

LEHRBUCH

Hans-Joachim Mittag
Katharina Schüller

Statistik

Eine Einführung mit interaktiven Elementen

6. Auflage

Freier
Zugang
zu interaktiven
Elementen

EBOOK INSIDE



Springer Spektrum

Statistik

Hans-Joachim Mittag · Katharina Schüller

Statistik

Eine Einführung mit interaktiven
Elementen

6., vollständig überarbeitete und ergänzte
Auflage



Springer Spektrum

Prof. Dr. Hans-Joachim Mittag
ehemals Fakultät für Kultur-
und Sozialwissenschaften
FernUniversität in Hagen
Hagen, Deutschland

Katharina Schüller
STAT-UP Statistical Consulting &
Data Science GmbH
München, Deutschland

ISBN 978-3-662-61911-7 ISBN 978-3-662-61912-4 (eBook)
<https://doi.org/10.1007/978-3-662-61912-4>

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

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert durch Springer-Verlag GmbH, DE, ein Teil von Springer Nature 2011, 2012, 2014, 2016, 2017, 2020

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung: Iris Ruhmann
Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Inhaltsverzeichnis

Vorwort	ix
I Beschreibende Statistik	1
Lernziele zu Teil I	2
1 Statistik, Daten und statistische Methoden	3
1.1 Statistik im Alltag, in Politik und Gesellschaft	3
1.2 Aufgaben und Teilbereiche der Statistik	7
1.3 Daten- und Methodenkompetenz	12
1.4 Veränderungen in der Methodenausbildung	18
2 Grundbegriffe der Statistik	21
2.1 Statistische Einheit, Merkmal und Grundgesamtheit	21
2.2 Merkmalsklassifikationen und Skalen	24
2.3 Operationalisierung von Merkmalen	28
3 Datengewinnung und Auswahlverfahren	31
3.1 Erhebungsarten und Studiendesigns	31
3.2 Stichprobenauswahl	41
3.3 Fehlschlüsse aus Daten in der Praxis	45
3.4 Träger amtlicher und nicht-amtlicher Statistik	49
4 Univariate Häufigkeitsverteilungen	51
4.1 Absolute und relative Häufigkeiten	51
4.2 Häufigkeitsverteilungen für klassierte Daten	61
4.3 Die empirische Verteilungsfunktion	69
5 Kenngrößen empirischer Verteilungen	73
5.1 Lagemaße	73
5.2 Streuungsmaße	80
5.3 Quantile und Boxplots	85
6 Analyse von Ereignisdaten	91
6.1 Anwendungsfelder und Grundbegriffe	91
6.2 Das Kaplan-Meier-Verfahren	94
7 Konzentration von Merkmalswerten	99
7.1 Die Lorenzkurve	99
7.2 Konzentrationsmaße	103

8 Indikatoren	109
8.1 Verhältniszahlen	109
8.2 Zusammengesetzte Indexzahlen	114
9 Bivariate Häufigkeitsverteilungen	123
9.1 Empirische Verteilungen diskreter Merkmale	123
9.2 Empirische Unabhängigkeit diskreter Merkmale	130
9.3 Empirische Verteilungen stetiger Merkmale	138
10 Zusammenhangsmaße	141
10.1 Nominalskalierte Merkmale	141
10.2 Metrische Merkmale	146
10.3 Ordinalskalierte Merkmale	155
II Wahrscheinlichkeitsrechnung und schließende Statistik	157
Lernziele zu Teil II	158
11 Zufall und Wahrscheinlichkeit	159
11.1 Grundbegriffe der Wahrscheinlichkeitsrechnung	159
11.2 Zufallsstichproben und Kombinatorik	166
11.3 Bedingte Wahrscheinlichkeiten	170
11.4 Sensitivität und Spezifität	175
11.5 Wahrscheinlichkeitsverteilungen	178
12 Diskrete Zufallsvariablen	183
12.1 Wahrscheinlichkeits- und Verteilungsfunktion	183
12.2 Kenngrößen diskreter Verteilungen	190
12.3 Die Binomialverteilung	193
12.4 Die hypergeometrische Verteilung	198
13 Stetige Zufallsvariablen	207
13.1 Dichtefunktion und Verteilungsfunktion	207
13.2 Kenngrößen stetiger Verteilungen	211
13.3 Normalverteilung und Standardnormalverteilung	213
13.4 χ^2 -, t- und F-Verteilung	221
14 Bivariate Verteilungen	229
14.1 Unabhängigkeit von Zufallsvariablen	229
14.2 Kovarianz und Korrelation	234
15 Schätzung von Parametern	239
15.1 Punktschätzungen und ihre Eigenschaften	241
15.2 Punktschätzung von Erwartungswerten und Varianzen	243

15.3 Punktschätzung von Anteilswerten	246
15.4 Intervallschätzung für Erwartungswerte	248
15.5 Intervallschätzung für Anteilswerte	251
16 Statistische Testverfahren	257
16.1 Arten statistischer Tests	258
16.2 Grundbegriffe und Gauß-Test für Erwartungswerte	260
16.3 Gütfunktion des Gauß-Tests	267
16.4 Signifikanzniveau und p -Wert	270
16.5 t -Test für Erwartungswerte	273
16.6 χ^2 -Test für Varianzen	276
16.7 Zweistichproben-Tests für Erwartungswerte	277
16.8 Unabhängigkeitstests	281
17 Das lineare Regressionsmodell	283
17.1 Das einfache lineare Regressionsmodell	285
17.2 KQ-Schätzung im einfachen Regressionsmodell	287
17.3 Das Bestimmtheitsmaß	291
17.4 Das multiple lineare Regressionsmodell	294
17.5 KQ-Schätzung im multiplen Regressionsmodell	300
17.6 Ausblick auf verallgemeinerte Regressionsmodelle	302
18 Grundzüge der Varianzanalyse	307
18.1 Das Modell der einfaktoriellen Varianzanalyse	309
18.2 Durchführung einer einfaktoriellen Varianzanalyse	311
18.3 Zweifaktorielle Varianzanalyse	317
III Anhänge	319
Lernziele zu Teil III	320
19 Tabellenanhang	321
19.1 Verteilungsfunktion der Binomialverteilung	321
19.2 Verteilungsfunktion der Standardnormalverteilung	328
19.3 Quantile der Standardnormalverteilung	330
19.4 Quantile der χ^2 -Verteilung	331
19.5 Quantile der t -Verteilung	332
19.6 Quantile der F-Verteilung	334
20 Übungsaufgaben und Lösungen	339
20.1 Übungsaufgaben zu Teil I	339
20.2 Übungsaufgaben zu Teil II	348
20.3 Lösungen zu Teil I	355
20.4 Lösungen zu Teil II	367

21 Verzeichnisse und Internet-Ressourcen	381
21.1 Literaturverzeichnis	381
21.2 Statistik-Software und Online-Ressourcen	384
21.3 Symbolverzeichnis	388
21.4 Autorenregister	390
21.5 Sachregister	392

Vorwort

Dieses Lehrbuch ist aus einem Kurs der FernUniversität Hagen hervorgegangen, der dort in mehreren Studiengängen zum Einsatz kam. Es ist betont interdisziplinär angelegt und deckt alle Inhalte einer traditionellen Einführung in die Statistik ab, insbesondere also die beschreibende Statistik sowie Grundlagen der Wahrscheinlichkeitsrechnung und der schließenden Statistik. Die Bearbeitung des Lehrtextes soll dazu befähigen, Daten und statistische Informationen nutzen, sachadäquat interpretieren und verständlich kommunizieren zu können. Diese als *statistische Daten- und Methodenkompetenz* bezeichnete Qualifikation ist für unsere digitale Wissensgesellschaft unverzichtbar.

Ursprung dieses Lehrtextes

Charakteristisch für das vorliegende Werk, das im März 2011 in 1. Auflage erschien, ist die Verknüpfung mit interessanten Web-Adressen, mit Lehrvideos und interaktiven Lernobjekten. Letztere ermöglichen es, statistische Verfahren „auszuprobieren“ und ausgewählte Datensätze unter Verwendung unterschiedlicher grafischer Instrumente zu explorieren.



Lernobjekte im Einsatz auf mobilen Endgeräten

Die interaktiven Lernobjekte sind auf der Basis von HTML5, Javascript und CSS programmiert. Bis Ende 2018 wurden in Kooperation mit der Hamburger Fern-Hochschule Lernobjekte realisiert und in einer virtuellen Bibliothek mit voneinander unabhängigen Elementen zusammengefasst. Diese Bibliothek repräsentiert eine frei zugängliche **Statistik-Web-App**. Die App wurde mit dem Comenius-EduMedia-Siegel 2015 der Gesellschaft für Pädagogik, Information und Medien, dem Innovationspreis Bildung 2015 des Bundesverbands für Bildung, Wissenschaft und Forschung e. V. (BBWF) sowie dem BBWF-Gütesiegel 2016 ausgezeichnet. Außerdem wurde sie für den eAward 2019 nominiert, einem der größten österreichischen IT-Wirtschaftspreise.



Ab 2019 wurden in Zusammenarbeit mit dem Institut für Kooperative Systeme, einem Aninstitut der FernUniversität Hagen, neue Lernobjekte zur interaktiven Visualisierung statistischer Methoden und interessanter Datensätze der amtlichen Statistik entwickelt. Diese wurden in einer weiteren, ebenfalls frei zugänglichen [neuen Statistik-Web-App](#) mit zusätzlichen Funktionalitäten und modernisiertem Design zusammengefasst.

Bedeutung der Icons
am Seitenrand

Wo interaktive Lernobjekte, Audios, Videos oder besondere Web-Links zu finden sind, wird am Rand durch runde Icons sichtbar gemacht:



Icons mit Links: Interaktives Lernobjekt, Audio, Video, externer Web-Link



Beispiel eines
interaktiven
Lernobjekts



Bei den interaktiven Lernobjekten und auch bei den meisten Videos sind zusätzlich *QR-Codes* (engl.: *Quick Response*) platziert, damit man sie direkt von der Printfassung des Lehrbuchs mit einem mobilen Endgerät aufrufen kann. Beim e-Buch kann man entweder auf den QR-Code oder auf das darunter eingestellte runde Icon klicken.

Innerhalb dieses Manuskripts findet man noch quadratische Icons, die nur der Orientierung dienen und nicht mit Links verknüpft sind. Diese Icons verweisen auf die *Kapitelvorschau* am Anfang eines Kapitels, auf die *Aufgaben* und die zugehörigen *Lösungen*, auf ergänzende *Literatur* sowie auf Textpassagen, die einen kritischen Blick auf Verfahren oder aktuelle Entwicklungen der statistischen Praxis unterstützen und damit *Denkanstöße* vermitteln:



Weitere Icons: Vorschau, Aufgabe, Lösung, Literatur, Denkanstöße

Externe und
interne Links

Von den Web-Links, die aus dem Lehrbuch hinausführen (externe Links), sind die interessantesten durch ein rundes Icon am Marginalienrand betont. Die meisten externen Links sind nur im Text anhand magentafarbener Schrift hervorgehoben. Verlinkungen innerhalb des Manuskripts (Sprünge zu Aufgaben oder zu Nummern von Gleichungen, Abbildungen oder Tabellen) erscheinen in blauer Schrift.

Ab der jetzigen 6. Auflage wird dieses Lehrbuch von einem Autorenteam verantwortet. Die neu hinzutretene Koautorin bringt langjährige Expertise aus der Statistikberatung für Unternehmen ein. Das Buch enthält nun zahlreiche neue Anwendungsbeispiele der Statistik aus weiteren Bereichen, z. B. der Medizin, der Qualitätssicherung, dem Kapitalmarktsektor und der Rechtsprechung. Für die vorliegende Auflage wurden u. a.

Was ist neu bei
der 6. Auflage?

- die Ausführungen zum Thema „Daten- und Methodenkompetenz“ deutlich erweitert, häufige Fehlschlüsse aus Daten in der Praxis beschrieben und ein Kapitel zur Analyse von Ereignisdaten aufgenommen;
- viele interaktive Lernobjekte sowie Lehrvideos neu entwickelt oder aktualisiert und erstmals auch Podcasts eingebunden;
- neue Grafiken, Fallbeispiele und Exkurse aufgenommen und Passagen mit dem Titel „Kritisch nachgefragt“ integriert, die Fragwürdiges mit Bezug zur Statistik anreißen und Denkanstöße geben;
- zu allen Aufgaben, die auch mit R lösbar sind, ergänzend die R-Codes unter einer externen Web-Adresse zugänglich gemacht;
- die Web-Links, die überblicksartige Zusammenstellung etablierter Statistik-Software sowie die Literaturliste auf den neuesten Stand gebracht und weitere Fotos aus der Praxis integriert.

Wie schon bei den vorausgegangenen Auflagen werden das e-Buch und die Printfassung des Buches als Paket angeboten („**eBook Inside**“). In jeder Printausgabe ist am Ende ein individualisierter Code enthalten, der den Zugang zum e-Buch vermittelt. Die Printausgabe ist durchgängig mehrfarbig. Beim e-Buch sind die interaktiven oder dynamischen Elemente direkt durch Anklicken der runden Icons am Seitenrand erreichbar. Auch die in den Text eingestreuten Web-Links zu Online-Ausgaben von Zeitschriften müssen hier nur angeklickt werden.

Buch und e-Buch
als Paket

Das e-Buch ist ein Hypertext, bei dem man Querverweise zu Gleichungen, Abbildungen oder Aufgaben (blau markiert) per Mausklick nachgehen oder vom Stichwortregister zu den dort aufgeführten Schlüsselbegriffen springen kann.

Vorzüge des e-Buchs

Printfassung mehrfarbig; Link zu interaktiven Experimenten, Audios, Videos über QR-Codes	 e-Buch mehrfarbiger Hypertext, interaktive Experimente, Audios und Videos integriert
--	---

Unterschiede zwischen Buch und e-Buch

Struktur des Buchs	Der erste Teil dieses Lehrbuchs widmet sich der beschreibenden Statistik, der zweite Teil der Wahrscheinlichkeitsrechnung und der schließenden Statistik. Der dritte Teil umfasst Aufgaben, Lösungen, statistische Tabelle und diverse Verzeichnisse. Die verwendeten Beispiele und Exkurse sind manchmal mit Hintergrundinformationen aus Online-Ausgaben über regionaler Zeitschriften verknüpft.
Danksagungen	Dank für die Programmierung der interaktiven Experimente gebührt Herrn Dr. T. <i>Augustin</i> , und Herrn A. <i>Michel</i> , beide Hagen. Frau A. <i>Dirks</i> , Hamburger Fern-Hochschule, konzipierte die Icons am Seitenrand. Herr Prof. Dr. R. <i>Männich</i> , Trier, und Herr Dr. Th. <i>Rahlf</i> , Bonn, stellten Grafiken zur Verfügung. Für inhaltliche Beiträge und Grafiken ist vom Team der Fa. STAT-UP, München, Herrn S. <i>Büsches</i> , Frau P. <i>Busch</i> , Herrn M. <i>Dumke</i> , Frau C. <i>Haas</i> , Herrn H. F. von <i>Koeller</i> , Herrn G. <i>Konecny</i> , Herrn P. <i>Kurthen</i> , Frau M. <i>Pröbstl</i> , Frau S. <i>Sieber</i> , Herrn Dr. R. <i>Sachse</i> sowie vor allem Herrn P. <i>Moreno Um</i> sehr zu danken. Verpflichtet sind wir ferner den nachstehenden Firmen und Institutionen, die kostenfrei Fotos oder andere Materialien zur Verfügung stellten:

- Fa. Bechem, Wetter (Herr C. *Hundertmark* und Herr K.-U. *Vieweg*),
- Fa. Böhme und Weihs GmbH, Sprockhövel (Herr Dr. N. *Böhme*),
- CeDiS, Berlin (Herr Dr. A. *Geukes* und Herr A. *Schulz*),
- Statistisches Bundesamt, Wiesbaden (Herr Dr. G. *Brückner*),
- Evonik Industries AG, Standort Essen (Herr Dr. W. *Wolfes*),
- GfK, Nürnberg (Herr Dr. S. *Maier* und Herr R. *Nicklas*),
- Hessischer Rundfunk, Frankfurt (Herr C. *Bender*),
- JMP / SAS, Köln (Herr Dr. V. *Kraft*),
- Kraftfahrtbundesamt, Flensburg (Herr U. *Siebert*),
- Q-DAS GmbH, Weinheim (Herr Dr. E. *Dietrich*),
- TNS Infratest, München (Herr M. *Kögel*),
- Volkswagen AG, Wolfsburg (Herr T. *Cramm*).

Zu danken ist schließlich Frau I. *Ruhmann* vom Springer Verlag für ihre sehr engagierte Unterstützung bei der Vorbereitung dieser Neuauflage.

Wetter / Ruhr und München, im Juni 2020

Hans-Joachim Mittag
mail@mittag-statistik.de

Katharina Schüller
katharina.schueler@stat-up.com

Teil I

Beschreibende Statistik



Lernziele zu Teil I

Nach der Bearbeitung des ersten Teils dieses Manuskripts sollten Sie

- wissen, warum Daten- und Methodenkompetenz heutzutage als Schlüsselqualifikation gilt;
- zentrale Aufgaben und Anwendungsfelder der Statistik kennen;
- mit wichtigen Grundbegriffen der Statistik vertraut sein (z. B. Merkmale und Merkmalstypen);
- alternative Ansätze zur Gewinnung von Daten und zur Entnahme von Stichproben kennen;
- einige Fehlschlüsse kennen, die bei der Erhebung und Analyse von Daten häufig auftreten;
- Datensätze unter Verwendung geeigneter Grafiken visualisieren können;
- in der Lage sein, Lage- und Streuungsparameter empirischer Verteilungen zu berechnen;
- wissen, wie sich Merkmalskonzentration messen und visualisieren lässt;
- den Einsatzzweck von Indikatoren sowie einige Beispiele für Indikatoren benennen können;
- in der Lage sein, Datensätze für zwei Merkmale anhand von Kontingenztafeln oder, bei stetigen Merkmalen, anhand von Streudiagrammen darzustellen;
- Maße kennen, mit denen sich ein Zusammenhang zwischen zwei Merkmalen quantifizieren lässt.



1 Statistik, Daten und statistische Methoden



Vorschau auf das Kapitel

Anhand von Beispielen aus verschiedenen Lebensbereichen und Anwendungsfeldern wird illustriert, welche Bedeutung der Statistik heute zukommt. Statistik wird als eine Wissenschaft definiert, die Methoden zur Gewinnung von Daten und zum Lernen aus Daten bereit stellt. Es wird sichtbar gemacht, welches breite Aufgabenspektrum die Statistik umfasst und welche Teilbereiche sich unterscheiden lassen. Dabei wird deutlich, dass Daten- und statistische Methodenkompetenz eine in immer mehr Arbeitsfeldern benötigte Schlüsselqualifikation darstellt, die auch im privaten Bereich nützlich ist. Dieser Kompetenz ist daher ein eigener Abschnitt gewidmet.

Am Ende des Kapitels werden aktuelle Entwicklungen in der Statistikausbildung skizziert.

1.1 Statistik im Alltag, in Politik und Gesellschaft

Die **Statistik** ist eine Wissenschaft, die alle Lebensbereiche durchdringt. Jeder von uns ist heute im Alltag mit einer Fülle von Daten und Visualisierungen von Daten konfrontiert, die uns über verschiedene Kanäle erreichen. Wenn wir am Morgen das Radio einschalten oder die Tageszeitung aufschlagen, erfahren wir etwas über die Entwicklung von Aktienkursen, über Trends auf dem Arbeitsmarkt oder über Ergebnisse der von der OECD getragenen PISA-Studien. Abends können wir im Fernsehen die Ziehung der Lottozahlen verfolgen oder uns über den Stand des aktuellen ZDF-Politbarometers informieren. Im Internet kann man gezielt nach Daten aller Art suchen, z. B. nach statistischen Informationen zu Migrationsströmen in Europa oder zur Entwicklung der Erwerbstätigkeit in Deutschland. Zugleich wird die Online-Präsentation von Daten immer benutzerfreundlicher. Dies gilt insbesondere für Daten der amtlichen Statistik – man studiere etwa die unter dem Etikett „Statistik anschaulich“ zusammengefassten interaktiven Anwendungen des **Statistischen Bundesamts**.

Statistische Daten im Alltag

Die große gesellschaftliche Relevanz der Statistik spiegelt sich darin wider, dass der 20. Oktober alljährlich als Weltstatistiktag und seit 2016 auch als Europäischer Statistiktag gilt. Am 20. Oktober gibt es daher jeweils zahlreiche Veranstaltungen von Universitäten, Statistikämtern, Unternehmen oder internationalen Institutionen.



Video von Eurostat zum Europäischen Statistiktag 2016

Beispiel 1.1: Analyse der Wählerstimmung und Trendidentifikation

Seit 1977 wird regelmäßig im Auftrag des ZDF eine Stichprobe von Wählern in Deutschland nach ihrer aktuellen Parteipräferenz, nach der Bewertung der bekanntesten Politiker und nach ihrer Haltung gegenüber aktuellen Entwicklungen in Politik und Gesellschaft befragt. Die Ergebnisse der als ZDF-Politbarometer bezeichneten Erhebung werden jeweils über Fernsehen und Internet verbreitet. Da die Personen in der Stichprobe so ausgewählt werden, dass sie als repräsentativ für die gesamte Bevölkerung anzusehen sind, können aus den Befragungsergebnissen Aussagen für alle Wähler in Deutschland abgeleitet werden. Aufgrund der Regelmäßigkeit der Befragungen gewinnt man nicht nur Aussagen für einen bestimmten Zeitpunkt, sondern Informationen zu langfristigen Trends und Veränderungen der politischen Stimmung.

Eine Frage des ZDF-Politbarometers, die sog. „Sonntagsfrage“, projiziert die aktuelle Parteipräferenz der befragten Wähler auf die nächste Bundestagswahl. Die Frage lautet: „Welche Partei würden Sie wählen, wenn am kommenden Sonntag Bundestagswahl wäre?“. Die Antworten zur „Sonntagsfrage“ vom 8. Dezember 2017 werden in diesem Manuskript mehrfach zur Illustration der Anwendung statistischer Konzepte herangezogen, u. a. bei der Analyse von Merkmalszusammenhängen.

Statistische Verfahren
im Wirtschaftsleben

Die Statistik spielt auch für Unternehmen eine wichtige Rolle. Bei industriellen Fertigungsprozessen und im Dienstleistungsbereich werden statistische Verfahren schon in der Designphase eines Produkts oder einer Serviceleistung eingesetzt, um Fehler zu vermeiden und Kundenzufriedenheit zu sichern. Mängel können zu Gewährleistungsprozessen und imageschädigenden Rückrufaktionen führen und die Existenz selbst größerer Unternehmen bedrohen. Statistische Instrumente sind auch in der Markt- und Werbeforschung nicht mehr wegzudenken. Marktforschungsinstitute ermitteln Marktanteile und Marktpotenziale, etwa über computergestützte Telefoninterviews. Die Einschaltquoten für Radio- und Fernsehsender werden auf Stichprobenbasis geschätzt und determinieren dann die Preise von Werbespots. Banken setzen statistische Modelle bei Entscheidungen über die Vergabe von Krediten und bei der Analyse von Kapitalmarktdaten ein. Große Lebensmittelkonzerne werten die an den Kassen gesammelten Scannerdaten aus und können damit die aktuellen „Renner“ und Ladenhüter identifizieren. Pharmahersteller benötigen statistische Testverfahren, um die bei der Zulassung neuer Medikamente geforderten Wirksamkeits- und Unbedenklichkeitsnachweise zu erbringen. Statistische Testverfahren werden auch eingesetzt, um die Wirksamkeit psychotherapeutischer Maßnahmen zu evaluieren.

Statistik ist
fachübergreifend

Die Statistik erfüllt für viele Wissenschaften eine wichtige Servicefunktion. In der *Soziologie*, der *Psychologie* oder auch der *Medizin* stützen sich Fachpublikationen maßgeblich auf Daten und deren statistischer Analyse. Die

Versuchsplanung, bei der es u. a. um die planmäßige Variation von Einflussfaktoren geht, ist ein weiteres Beispiel für den fächerübergreifenden Einsatz statistischer Methoden. Sie ist ein wichtiges Feld der experimentellen Psychologie und zugleich auch der *Ingenieurwissenschaften* – man denke an Experimente in der *Sozialpsychologie* zur Untersuchung von Motivationsstrukturen bei ehrenamtlich tätigen Personen oder an Belastungstests bei der Erforschung neuer Verbundwerkstoffe für Kraftfahrzeuge. Statistische Instrumente des Qualitätsmanagements werden in der *Bildungspädagogik* sowie in der *Gesundheitsökonomie* bei der Steuerung von Schulentwicklungen und Krankenhausbelegungen verwendet. Weitere Anwendungsfelder der Statistik sind die Beschreibung von Zufallsprozessen in der *Physik* (u. a. Brownsche Bewegung), die Berechnung von Lebensversicherungsprämien in der *Versicherungsmathematik*, die Verwendung von Zeitreihenmodellen in der *Kapitalmarktforschung*, die Analyse von Querschnitts- und Paneldaten in den *Wirtschaftswissenschaften*, die Modellierung von Wachstumsprozessen in der *Biologie* sowie die Gewinnung empirisch fundierter Aussagen zum Zustand von Wäldern und Gewässern in den *Umweltwissenschaften*.



Abb. 1.1: Qualitätskontrolle bei der Tensideherstellung (Säurezahlbestimmung und Eingabe für die statistische Auswertung); Quelle: Evonik Industries AG, Essen

Die Statistik spielt auch bei der *Politikplanung* und bei der Erfolgsbewertung von Politik eine gewichtige Rolle. Harmonisierte, d. h. über Ländergrenzen vergleichbare Daten, die *Eurostat*, das Statistische Amt Europas in Luxemburg, zusammenstellt und frei zugänglich macht, werden für nationale und europäische Politiken genutzt. So sind verlässliche Bevölkerungszahlen die Basis für Entscheidungen in der Gesundheits- und Bildungspolitik und werden für Abstimmungen des EU-Ministerrats nach dem Grundlagenvertrag von Lissabon benötigt (Erfordernis der „doppelten Mehrheit“ mit 55% der Staaten, die 65% der EU-Bevölkerung repräsentieren).

Statistik in der Politik

Beispiel 1.2: Monitoring strategischer Ziele der Politik



Interaktives Objekt
„Erwerbstätigkeit“



Interaktives Objekt
„Emission von
Treibhausgasen“

Im Jahr 2010 verständigten sich die Staats- und Regierungschefs der Länder der EU auf eine mit *Europa 2020* etikettierte Strategie, die wirtschaftliche und soziale Kernziele für Europa bis 2020 festlegte und anhand von acht Leitindikatoren operationalisierte. Ein Ziel ist z. B. die Erhöhung der Beschäftigungsquote der als erwerbsfähig geltenden EU-Bevölkerung auf 75 %. Der Erreichungsgrad dieses Ziels wird von Eurostat mit Hilfe des Indikators „Erwerbstätigenquote (Altersgruppe 20 - 64 Jahre)“ gemessen.

Ein weiteres Ziel, das mit der Unterzeichnung des Kyoto-Protokolls (Zusatzvereinbarung zur Klima-Rahmenvereinbarung der Vereinten Nationen) in Zusammenhang steht, beinhaltet die Senkung der Emissionen von Treibhausgasen für jedes Land um 20% gegenüber dem Stand von 1990. Bis zum Jahr 2030 sollen die Treibhausgasemissionen in der EU sogar 30% unterhalb des jeweiligen nationalen Stands von 1990 liegen.

Auch die *Vereinten Nationen* (UN) verfolgen globale Strategien und verknüpfen diese mit Indikatoren. So gilt seit Anfang 2016 die *UN Millennium Agenda 2030 für nachhaltige Entwicklung*. Dies ist ein Aktionsplan mit 17 Kernzielen, die nicht nur auf Entwicklungsländer und auf die Verminderung von Hunger und extremer Armut abgestellt sind. Vielmehr bieten sie einen für alle Länder gültigen Rahmen, der auf ein Umsteuern in Richtung nachhaltigen Wirtschaftens abzielt. Daten spielen auch hier eine Schlüsselrolle für das Politikmonitoring.

Aggregate aus verschiedenen Indikatoren, sog. zusammengesetzte Indikatoren, werden von internationalen Organisationen zur Beschreibung komplexer Entwicklungen eingesetzt, etwa zur Messung von Wohlfahrt oder Innovation.



**Kritisch
nachgefragt**



Interaktives Objekt
„Treibhausgasemissionen
pro Kopf“

Das von der EU-Kommission verfolgte Ziel, die Treibhausgasemissionen in jedem EU-Staat bis 2020 um 20% bzw. bis 2030 um 30% gegenüber dem im jeweiligen Land beobachteten Stand von 1990 zu senken, ist ein pragmatischer Ansatz, der den Vorteil hat, kein Land zu überfordern. Die Ausgangsniveaus im Jahr 1990 sind aber hierbei ausgeblendet – sie werden für alle Länder mit dem Wert 100 verbunden. Der Ansatz erlaubt es folglich nur, Erfolge bei der Verringerung von Treibhausgasen für jedes Land einzeln zu bewerten.

Will man die Emissionsniveaus zwischen Ländern vergleichen, kann man Pro-Kopf-Emissionswerte heranziehen. Dabei erhält man ein deutlich anderes Ranking. Schweden schneidet hier z. B. deutlich besser, Deutschland etwas schlechter und Luxemburg wesentlich schlechter ab. Es genügt demnach nicht, die Klimapolitik von Ländern danach zu beurteilen, wie stark die Emissionswerte eines Referenzjahrs unterschritten wurden. In den Medien und auch in der Politik wird aber fast ausschließlich die Erreichung oder Verfehlung der 20%- bzw. der 30%-Marke thematisiert und der Erfolg umweltpolitischer Maßnahmen allein hieran gemessen.

1.2 Aufgaben und Teilbereiche der Statistik

Die Statistik ist also eine Disziplin mit vielfältigen Aufgaben und Anwendungsbereichen. Das Spektrum reicht von der Planung der *Erhebung von Daten* über die *Beschreibung und Visualisierung* der erhobenen Befunde über die *Identifikation von Auffälligkeiten* in den Daten bis zur *Ableitung von Schlüssen*, die über die vorliegenden Daten deutlich hinausgehen. Die Statistik ist demnach eine Wissenschaft, die Methoden zur Gewinnung und Analyse von Daten sowie zum Lernen aus Daten bereit stellt.

Aufgaben der Statistik

Umgangssprachlich wird Statistik oft anders verstanden, nämlich als eine schwer zugängliche, spröde Disziplin, die sich der Sammlung und Auswertung von Zahlenfriedhöfen verschrieben hat. Dieses Fehlverständnis reduziert die Statistik auf Tätigkeitsfelder, die für die heutige Statistik keinesfalls repräsentativ sind. Statistik ist eine faszinierende Wissenschaft mit vielfältigen Bezügen zur Praxis und interdisziplinärem Charakter.

Öffentliche Wahrnehmung des Fachs

Für Statistiker ist der Begriff „Statistik“ nicht eindeutig belegt. Sie verstehen hierunter einerseits ihre *Wissenschaft* als Ganzes. Sie verwenden den Begriff aber auch für *Kenngrößen*, die sich aus statistischen Daten ableiten (z. B. den Mittelwert). Im allgemeinen Sprachgebrauch wird auch häufig ein *Datensatz* als eine Statistik angesprochen, etwa ein Datensatz mit der Medaillenverteilung bei den Olympischen Sommerspielen. In diesem Manuskript wird „Statistik“ im Sinne von „Wissenschaft“ verwendet.

Mehrdeutigkeit des Begriffs „Statistik“

Innerhalb der Statistik lassen sich die beschreibende und die schließende Statistik unterscheiden. Die **beschreibende Statistik** oder **deskriptive Statistik** (engl.: *descriptive statistics*) umfasst numerische und grafische Verfahren zur Charakterisierung und Präsentation von Daten. Ziel ist die Reduktion der in den Daten enthaltenen statistischen Informationen durch Verdichtung zu wenigen Kenngrößen, möglichst ohne größeren Informationsverlust. Das Europäische Amt für Statistik sammelt z. B. Daten zu Bruttoverdiensten für Millionen von Arbeitnehmern, die nur in aggregierter Form für die Politikplanung brauchbar sind. Techniken der Datenerhebung werden meist der beschreibenden Statistik zugerechnet.¹ Jede empirisch arbeitende Wissenschaft argumentiert mit Daten und bedient sich der Instrumente der beschreibenden Statistik.

Teilbereiche der Statistik:

Aus der beschreibenden Statistik ging mit den Fortschritten in der Informationstechnologie die **explorative Datenanalyse** hervor (engl.: *exploratory data analysis*). Diese geht über die beschreibende Statistik hinaus, weil hier – noch ohne Einsatz von Modellen – mit rechenintensiven

Beschreibende Statistik

¹In Anwendungsfeldern der Statistik, in denen die Datenerhebung im Rahmen umfassender Forschungsprozesse zu planen ist – wie etwa bei der Datengewinnung über Fragebögen in den Sozialwissenschaften oder über Experimente mit Versuchspersonen in der Psychologie – hat sie einen höheren Stellenwert und wird dort oft als eigenständiger Bereich angesehen.

Explorative Datenanalyse und „Big Data“

Verfahren nach auffälligen Mustern und Strukturen in Datenbeständen gesucht wird. So werden etwa die Scannerdaten eines Lebensmittelkonzerns von einem Verkaufstag routinemäßig nach Auffälligkeiten durchleuchtet, ohne dass sofort eine Hypothese im Spiel ist. So entdeckt man Trends im Käuferverhalten und kann zudem rechtzeitig Nachbestellungen organisieren. Man spricht hier von **Data Mining**. Die explorative Datenanalyse wird meist ebenfalls der beschreibenden Statistik zugeordnet. Unter dem heutzutage vielbenutzten Schlagwort „**Big Data**“ versteht man extrem große und mit hoher Geschwindigkeit anfallende digitale Datenmengen, die aus ganz unterschiedlichen Quellen stammen können (etwa Daten von Biosensoren und Überwachungskameras, Daten von Bezahlvorgängen oder von Aktivitäten in sozialen Netzwerken). Diese Daten werden computergestützt in Echtzeit mit mathematisch-statistischen Verfahren, u. a. mit Methoden des Data Mining, auf Zusammenhänge untersucht.

Wahrscheinlichkeitsrechnung und schließende Statistik

Die **schließende Statistik** oder **induktive Statistik** (engl.: *inferential statistics*) leitet aus Stichprobendaten Aussagen ab, die über die jeweilige Stichprobe hinausgehen und sich auf eine umfassendere Grundgesamtheit beziehen. Die Stichprobendaten werden als Ausprägungen von Zufallsvariablen interpretiert (Modellvorstellung) und das Verhalten solcher Zufallsvariablen durch sog. Verteilungsmodelle (engl: *probability distributions*) beschrieben.² Typische Aufgaben der schließenden Statistik sind das *Schätzen* von Modellparametern und das *Testen* von Hypothesen. Die aus den Daten abgeleiteten Folgerungen sind mit Unsicherheiten verknüpft (Schätzfehler beim Schätzen, Fehlentscheidungen beim Testen). Die **Wahrscheinlichkeitsrechnung** liefert die Grundlagen für die Berechnung von Wahrscheinlichkeiten auf der Basis von Verteilungsmodellen. Sie ist daher eng mit der schließenden Statistik verknüpft.

Exkurs 1.1: Künstliche Intelligenz und maschinelles Lernen



Audio „Künstliche Intelligenz“ (Zeit Online – Digital)

Künstliche Intelligenz (KI) bezeichnet ein Teilgebiet der Informatik, dessen Ziel es ist, Verhalten und Entscheidungen in unterschiedlichen Kontexten automatisiert von Computern und Maschinen ausführen zu lassen. Als KI-Schlüsseltechnologie kann **maschinelles Lernen** angesehen werden. Dieses zielt auf die Generierung von Wissen aus Erfahrung. Mit Hilfe von Lernalgorithmen wird hier aus Beispielen ein komplexes Modell entwickelt, das Muster in den Trainingsdaten erkennt und das erworbene Wissen auf unbekannte Daten überträgt. Auf Basis dieses Modells werden dann Entscheidungen abgeleitet.

In diesem Sinne ähneln sich Statistik und maschinelles Lernen, jedoch fußen die meisten Modelle der klassischen Statistik auf Verteilungsannahmen für Zufallsvariablen (sog. parametrische Modelle), während dies für die Modelle des maschinellen Lernens i. d. R. nicht zutrifft. Maschinelles Lernen reizt vor

²Die hier nur angerissenen Begriffe werden in Abschnitt 11.5 ausführlicher erklärt.

allem aufgrund der Fähigkeit, mit großen Datenmengen umgehen und komplexe System präzise darstellen zu können. Eine Gefahr liegt darin, dass mit dem Lernen aus Trainingsdaten eine Überanpassung verbunden sein kann, wodurch der Erkenntnistransfer auf unbekannte Daten verzerrt ist (vgl. Abschnitt 3.3).

Vorhersagen mit Verfahren der KI sind, anders als meist in der Statistik, nicht unmittelbar über eine Formel nachvollziehbar. Aus diesem Grund findet maschinelles Lernen bisher vor allem dort Anwendung, wo stetig Daten erzeugt werden und die schnelle Verarbeitung dieser zu kontinuierlich besseren Vorhersagen führen soll. Auf maschinellem Lernen basierende Systeme sind inzwischen in der Lage, radiologische Bilder mindestens so gut wie Mediziner zu analysieren und in komplexen Spielen, etwa Go und Poker, gegen Menschen zu gewinnen. Weitere Anwendungsfelder sind u. a. die Aufdeckung von Kreditkartenmissbrauch sowie die Sprach- oder Gesichtserkennung. Statistiker sollten sich jedenfalls heute der Möglichkeiten des maschinellen Lernens bewusst sein, aber auch der Vorteile ihrer Methoden gegenüber reinen KI-Verfahren.

KI-Verfahren bieten nicht nur Chancen, sondern auch Risiken für unsere Gesellschaft. Die Verknüpfung von Daten aus Überwachungskameras zur Personenidentifikation mit einem Sozialkreditsystem in China liefert ein aktuelles Anschauungsbeispiel (vgl. Abschnitt 8.1). Bedenklich ist auch, dass sich Videomaterial mit KI-Techniken leicht verfälschen lässt. Menschen begehen scheinbar Straftaten, die sich in Wirklichkeit nie ereignet haben. Es wird in Zukunft aufgrund des technischen Fortschritts immer schwieriger, bearbeitetes Material vom Original zu unterscheiden. Künstliche Intelligenz wird jedenfalls erhebliche Auswirkungen auf den Arbeitsmarkt haben. Es werden viele Routinetätigkeiten verschwinden und dafür neue Berufe hinzukommen (s. Exkurs 1.3).



Audio
„Gesichtserkennung in Deutschland“ (Zeit
Online – Digital)

Die Statistik ist heute längst, je nach Anwendungsschwerpunkt, in Spezialgebiete aufgefächert, die z. T. als eigenständige Disziplinen gelten. So hat die **Biometrie** ihren Schwerpunkt in statistischen Anwendungen im Bereich der Medizin und Biologie einschließlich der Identifikation von Mustern beim Menschen (Fingerabdruck-, Iris- und Gesichtserkennung). Die **Ökonometrie** bedient sich statistischer Methoden, um volks- und betriebswirtschaftliche Theorien zu quantifizieren und entsprechende Modelle anhand von Beobachtungsdaten zu überprüfen. Erwähnt sei auch die **Psychometrie**, bei der es um Theorien und Instrumente für das Messen in der Psychologie geht. Eine typische Aufgabe der Psychometrie ist z. B. die Entwicklung von Persönlichkeitstests.

Spezialgebiete der
Statistik und
verwandte
Disziplinen

In Stellenausschreibungen überregionaler Zeitungen findet man zunehmend Angebote für Datenwissenschaftler (engl.: *data scientists*). Die Datenwissenschaft ist von der Statistik nicht zu trennen. Es besteht aber Konsens, dass „Datenwissenschaft“ und „Statistik“ nicht dasselbe bezeichnen. Die **Datenwissenschaft** ist eine im Vergleich zur Statistik stärker anwendungsorientierte Disziplin, die sich der Analyse und dem Management von Daten widmet, insbesondere auch sehr großer Datenbestände

Sind
„Datenwissenschaft“
und „Statistik“
Synonyme?

(„**Big Data**“). Neben Verfahren der Statistik sind hier Kenntnisse der Informatik, der Datensicherheit sowie der effizienten Programmierung von Algorithmen aus den Bereichen Mathematik und Operations Research gefragt. Eine für die Datenwissenschaft typische Aufgabe besteht darin, aus unübersichtlichen Datenmengen (Rohstoff „Daten“) Informationen zu generieren, die für Managemententscheidungen benötigt werden.

Exkurs 1.2: Was ändert sich im Zeitalter von „Big Data“?

Es wurde schon erwähnt, dass wir mit einer stetig wachsenden Flut von Daten aus unterschiedlichen Quellen konfrontiert sind – z. B. Telefonie- und Internetnutzungsdaten, Daten von Finanzmarkttransaktionen, Messwerte von Sensoren in der Industrie oder Daten aus bildgebenden Untersuchungen in der Medizin. Die Analyse solcher komplexen Datenbestände ist eine Herausforderung für die Informationstechnologie. Die angewendeten Verfahren zielen darauf ab, das in den Daten steckende Informationspotenzial zu erschließen und nutzbar zu machen. Dabei werden in Echtzeit unter Einsatz von Hochleistungsrechnern u. a. Muster und eine Vielzahl von Korrelationen zwischen Variablen identifiziert. Aus den Korrelationen können sich dann Hypothesen für die Forschung ergeben.

MAYER-SCHÖNBERGER / CUKIER (2017) sprechen in diesem Kontext von einer Revolution für Wissenschaft und Gesellschaft. Revolutionär ist nach Überzeugung der Autoren, dass Hypothesen im Zeitalter von „Big Data“ nicht mehr aus Theorien abgeleitet und anhand von Daten geprüft werden müssen (theoriegetriebene Forschung), sondern das Ergebnis der automatisierten Anwendung von Analysealgorithmen sein können (datengetriebene Forschung).

Daten als Basis für datengestützte Entscheidungsfindung

In unserer heutigen Wissens- und Informationsgesellschaft werden überall Entscheidungen wesentlich durch Daten gestützt und empirisch abgesichert. **Datengestützte Entscheidungsfindung**, meist unter dem Etikett **Evidence Based Decision Making** firmierend, ist z. B. in der *Medizin* allgegenwärtig. Bei *komunalen Planungen* stützt man Entscheidungen über Investitionen auf Bevölkerungsdaten, etwa bei der Planung von Schulen. In der *Markt- und Meinungsforschung* werden Umsatzdaten als Basis für Entscheidungen über Sortimentsveränderungen und Produktinnovationen genutzt. Bei der Europäischen Kommission werden Entscheidungen zur Förderung strukturschwacher Regionen mit Mitteln des EU-Strukturfonds von der Datenlage bestimmt, d. h. statistische Informationen beeinflussen direkt die *Politikplanung*.

Die zunehmende gesellschaftliche Relevanz der Statistik spiegelt sich auch darin wider, dass namhafte Zeitungen und Zeitschriften in ihren Häusern Ressorts eröffnet haben, die sich einer neuen Form des datengestützten Online-Journalismus widmen. Bei der als **Datenjournalismus** (engl: *data journalism*) bezeichneten Entwicklung steht die Verbindung

interessanter Datensätze mit interaktiven Grafiken, Landkarten, Animationen und erläuterndem Text (Analysen, Kommentierungen) sowie sozialen Netzwerken im Vordergrund. Bei der englischen Tageszeitung *The Guardian* werden die verwendeten Daten zudem in frei zugänglichen Datenarchiven dem Leserpublikum zur Verfügung gestellt und in Datenblogs diskutiert. Es entsteht ein neues interaktives Erzählformat, bei dem die im Brennpunkt stehenden Daten mit Datenbanken verknüpft sind.

Exkurs 1.3: Neue Berufsfelder im Kontext der Digitalisierung

Neben dem Beruf des Datenjournalisten tun sich im Zuge der Digitalisierung weitere neuartige Berufsfelder auf. In allen Bereichen spielen Daten eine immer größere Rolle. Beispielsweise nimmt im medizinischen Kontext die Bedeutung der Bildverarbeitung stark zu. Diese ermöglicht es, automatisiert aus Bildern Diagnosen und Handlungsempfehlungen zu generieren, die ein Mensch nicht mit derselben Treffsicherheit oder Geschwindigkeit herleiten könnte.

Abbildung 1.2 vermittelt eine Übersicht über einige neue Berufe, bei denen der Umgang mit Daten zentral ist. Die Grafik enthält auch Bezeichnungen, die keine neuen Berufe sind, sondern symbolisieren, welche bisher schon existierenden Rollen (Entscheider, Fachexperte, Bürger) heute an welchen Stellen Datenkompetenzen benötigen. Die grauen vertikalen Pfeile deuten an, über welche Kompetenzbereiche sich die Berufe und Rollen erstrecken. Man sieht, dass sich die Kompetenzbereiche z. T. überlappen. Folglich gilt es in Stellenbeschreibungen genau auf die Aufgabenprofile zu achten. Berufsbezeichnungen für datenbezogene Tätigkeiten, die in Stellenanzeigen häufig in englischer Sprache wiedergegeben sind, wurden auch in der Grafik übernommen.

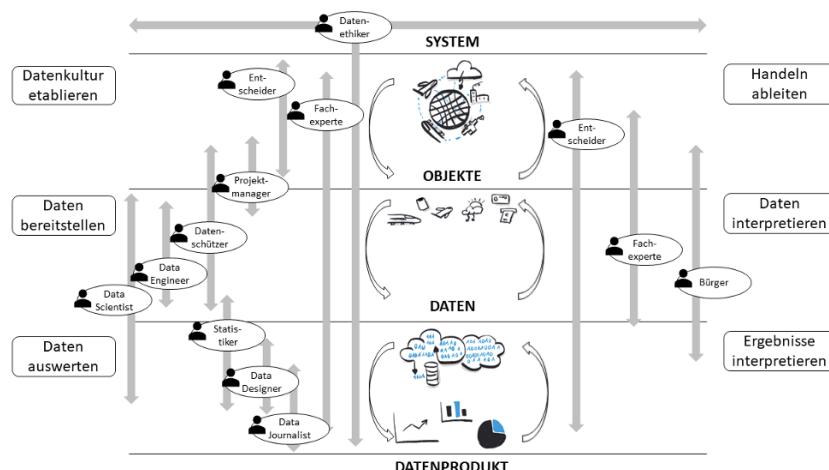


Abb. 1.2: Neue Berufe mit starkem Bezug zu Daten

Der *Data Designer* entwirft Konzeptstudien und sammelt Daten und visualisiert Auswertungen. Er ist entweder auf die Gestaltung von datenbezogenen

Prozessen und Architekturen (*Technical Data Designer*) oder auf das Design von Visualisierungen (*Data Artist*) spezialisiert. Die Begriffe *Data Scientist*, *Data Engineer* und *Data/Business Analyst* werden oft unscharf verwendet und nicht klar voneinander abgegrenzt. Man findet die Bezeichnung *Data Scientist* oft im Zusammenhang mit Tätigkeiten, die sich auf die Entwicklung / Anwendung von automatisierten Lernverfahren beziehen (besonders bei sehr großen und unstrukturierten Datensätzen), wohingegen ein *Data Engineer* sich mehr mit technischen Aspekten befasst, etwa mit der Verwaltung von Datenbanken und IT-Infrastrukturen. Da gerade im wirtschaftlichen Kontext, aber auch allgemein immer mehr Prozesse datengetrieben sind, werden in Anzeigen immer mehr *Data Analysts* gesucht, die neue Erkenntnisse aus Daten ableiten sollen, insbesondere aus „Big Data“.

Datenethik wird eine immer größere Rolle in der Berufswelt spielen. Im Gegensatz zum Datenschutz, der sich mit der Frage „Was darf ich?“ beschäftigt, lautet hier die Fragestellung „Was soll ich?“. Der *Datenethiker* muss „vorausdenken“, welcher Umgang mit Daten problematisch werden könnte, auch wenn eine aktuelle Praxis nicht gegen geltendes Recht verstößt. Dabei steht er mit dem Blick auf visionären Fragestellungen sowohl über der produzierenden (links in Abbildung 1.2) als auch über der rezipierenden Seite (rechts in Abbildung 1.2) im Prozess der Wertschöpfung aus Daten. Datenschutz kann ein Teilaспект von vielen sein, doch konkret befasst sich der Datenethiker mit dem gesellschaftlichen und normativen Wandel durch die Digitalisierung und den Konsequenzen auf die Wirtschaft und auf uns als einzelne Bürger. Dem Datenethiker geht es darum, wohin sich ein Gesellschaftssystem entwickelt, was für die Menschheit gut und schlecht ist und was die Ziele sind. Aktuell ist Datenethiker eher eine Rolle als ein gelernter Beruf – eine Rolle, die im Betrieb und in der Politik immer gefragter ist.

1.3 Daten- und Methodenkompetenz

Seit Jahren wird über Schlüsselqualifikationen und Kompetenzen diskutiert, die Menschen dazu befähigen, den sich wandelnden Anforderungen des Berufs und des gesellschaftlichen Lebens gerecht zu werden.

Schlüsselqualifikationen beziehen sich auf Fähigkeiten zur sachadäquaten *Anwendung von Wissen* und auf Strategien zur *Erschließung neuen Wissens*, gehen also über die Aneignung von Wissensinhalten hinaus. Es gibt unterschiedliche Arten von Schlüsselqualifikationen, etwa soziale Kompetenz (umfasst Kommunikationsfähigkeit im zwischenmenschlichen Bereich), Informationskompetenz (Fähigkeit zur effizienten Erschließung und Nutzung der kaum noch überschaubaren Informationsfülle) sowie Methodenkompetenz (Fähigkeit zur Nutzung unterschiedlicher Werkzeuge, Arbeitstechniken und Theorien zur Lösung von Problemen).

Da die Digitalisierung und die damit einhergehende Datafizierung das Leben und Arbeiten im 21. Jahrhundert nachhaltig verändern, werden Kompetenzen immer wichtiger, die sich auf den sachadäquaten Umgang mit Daten beziehen. Künstliche Intelligenz, vernetzte Produktion, kommunizierende Maschinen und selbstfahrende Autos werden von Daten gesteuert und produzieren selbst Daten am laufenden Band. Daten sind eine werthaltige Ressource („Öl des 21. Jahrhunderts“), Ausgangsbasis für Wissensschöpfung und Grundlage für bessere Entscheidungen.

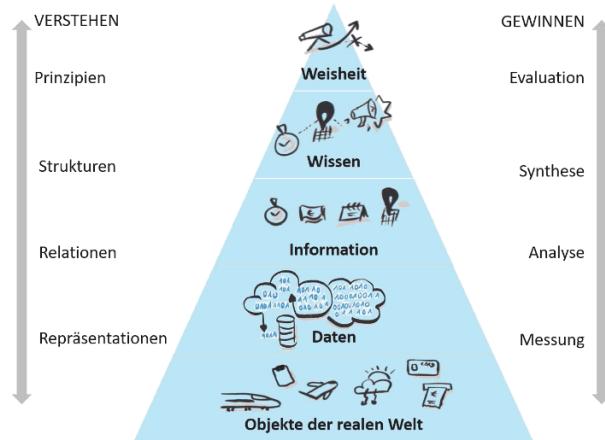


Abb. 1.3: Pyramidenmodell des Prozesses der Wertschöpfung aus Daten;
Quelle: SCHÜLLER / BUSCH / HINDINGER (2019)

Abbildung 1.3 zeigt ein Pyramidenmodell, das schematisch darstellt, wie durch einen zunehmenden Grad an Organisation Rohdaten im menschlichen Gehirn zu Informationen, Wissen und „Weisheit“ verarbeitet werden. Nach der Bereinigung und Verknüpfung einzelner Datenelemente zu bedeutsamer Information suchen wir nach Mustern, wenden Analyseprinzipien an und strukturieren die Informationen, beispielsweise durch Klassifikation oder Kategorisierung. Vorgelagert ist die Messung als Abbildungsprozess von Objekten der realen Welt in Daten.

Um systematisch Wissen bzw. Information aus Daten zu schöpfen, ist in allen Sektoren und Disziplinen die Fähigkeit von entscheidender Bedeutung, planvoll mit Daten umzugehen und sie im jeweiligen Kontext kritisch hinterfragen zu können. Diese als **Datenkompetenz** (engl: *data literacy*) bezeichnete Qualifikation beinhaltet die Fähigkeit, Daten zu sammeln, zu managen, zu bewerten und reflektiert anzuwenden. Sie umfasst außer einem breiten und tiefen Detailwissen über sich laufend verändernde statistische Methoden und Technologien auch ethische Grundhaltungen.

Abgrenzung von Kompetenzbegriffen

Eine weitere Kompetenz, die eng mit der sachadäquaten Nutzung von Daten verbunden ist, ist die **statistische Methodenkompetenz** (engl:

statistical literacy), ein Spezialfall der schon erwähnten generellen Methodenkompetenz. Sie bezieht sich auf die Fähigkeit, bei der Analyse und Präsentation von Daten geeignete statistische Verfahren auszuwählen, aus den Ergebnissen nachvollziehbare Schlüsse zu ziehen, diese verständlich zu kommunizieren und datenbasierte Entscheidungen zu treffen.

Die Begriffe „Data Literacy“ und „Statistical Literacy“ werden häufig synonym verwendet. Die aktuelle wissenschaftliche Diskussion tendiert aber dazu, Datenkompetenz als eine Fähigkeit zu verstehen, die statistische Methodenkompetenz einschließt und sowohl produzierende als auch rezipierende Schritte des Prozesses der Wertschöpfung aus Daten abdeckt. Die **Informationskompetenz** (engl: *information literacy*) ist noch umfassender. Sie ist eine rezeptive Fähigkeit, die sich auf den Umgang mit Informationen allgemein bezieht, nicht nur auf Daten und statistische Methoden.

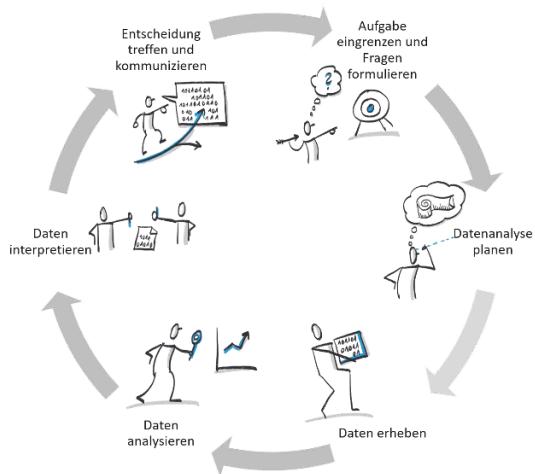


Abb. 1.4: Zyklisches Modell des Prozesses der Wertschöpfung aus Daten;
Quelle: SCHÜLLER / BUSCH / HINDINGER (2019)

Abbildung 1.4 zeigt ein zyklisches Prozessmodell, das noch besser als das Pyramidenmodell geeignet ist, sowohl die Erstellung von Datenprodukten durch den Fachexperten als auch den kompetenten Umgang mit Daten durch den Endanwender abzubilden. Insbesondere werden die Fähigkeit und Bereitschaft, eine Datenkultur zu etablieren (d.h. Nutzungsmöglichkeiten von Daten zur Entscheidungsfindung innerhalb eines gegebenen Systems zu identifizieren, zu spezifizieren und zu koordinieren) sowie die Fähigkeit und Bereitschaft, Handeln abzuleiten (d.h. Handlungsmöglichkeiten zu identifizieren, datengetrieben zu handeln und die erzielte Wirkung im System zu evaluieren) als Teilkompetenzen von Datenkompetenz verstanden.

Durch die Digitalisierung entstehen Daten häufig nicht mehr als geplantes Produkt zur Beantwortung einer Forschungsfrage („Science starts with a question“), sondern „zweckfrei“ in großer Menge und Heterogenität. Dabei eröffnen sich neue Fragen („Data Science starts with the data“). Datengetriebenes bzw. exploratives Vorgehen verbreitet sich in der Praxis zunehmend, so dass neue Kompetenzen hinsichtlich des Umgangs mit neuen Formen von Daten – z. B. Text, Ton, Bild – jenseits der bekannten Skalenniveau-Einteilungen und Speicherungsformen benötigt werden. Dabei ist das Risiko von Fehlschlüssen hoch, wenn ein grundlegendes Verständnis für statistische Fehler – etwa das Verwechseln von Korrelation und Kausalität – fehlt oder wenn unerkannte Mängel in der Qualität der Daten vorliegen, z. B. eine Verzerrung von Stichproben aufgrund fehlender Repräsentativität. **Ethische Kompetenz** (engl: *ethical literacy*) ist unbedingt erforderlich, wenn Daten frei kombiniert und für andere Zwecke als ihren ursprünglichen Erhebungszweck analysiert werden. Sie beschreibt die Bereitschaft und Fähigkeit, die Bedeutung von Daten zur Entscheidungsfindung vollständig zu erfassen. Dies bedingt, mögliche Interpretationen von Daten in unterschiedlicher Kontextualisierung zu reflektieren und kritisch zu bewerten. Datenkompetenz berührt jedenfalls auch ethische Fragen. Der Aufbau eines auf Daten gestützten Systems in China zur flächendeckenden Überwachung aller Einwohner des Landes macht dies deutlich.

Selbstverständlich dürfen Statistiken nicht bewusst manipuliert werden, um deren Nutzer zu bestimmten Entscheidungen zu verleiten. Doch selbst eine objektiv korrekte Analyse von Daten kann ethisch fragwürdig sein, wenn sie zu Diskriminierungen führt. Ein Beispiel hierfür ist die früher praktizierte unterschiedliche Preisgestaltung von Krankenversicherungstarifen in Abhängigkeit vom Geschlecht. Obwohl die durchschnittlichen Krankheitskosten von Frauen höher sind, darf das Geschlecht nach einem Urteil des Europäischen Gerichtshofs nicht mehr als Risikofaktor in die Tarifgestaltung einfließen. Zur Aufgabe, ein solches Risiko-Scoring handwerklich korrekt durchzuführen, gesellt sich also unmittelbar die Herausforderung, die Zulässigkeit des Scorings zu beurteilen, d. h. seinen Gebrauch und potenziellen Missbrauch zu beurteilen. Die Erhebung, Nutzung, Verarbeitung und Analyse von Daten kann nicht kontextunabhängig erfolgen, d. h. getrennt von deren Interpretation und Anwendung.

Die ethische Dimension von Datenkompetenz und statistischer Methodenkompetenz manifestiert sich u. a. in der Bereitschaft

- bei der Datenanalyse eine skeptische Grundhaltung zu bewahren;
- Informationsverluste infolge der Reduktion von Daten auf statistische Kennzahlen zu berücksichtigen;

Datenethik als Kompetenzdimension

Wünschenswerte Werthaltungen

- „good analytics standards“ einzuhalten, selbst wenn diese nicht explizit definiert sind (z. B. Gewährleistung von Daten-Nachhaltigkeit durch saubere Dokumentation und Archivierung von Daten);
- abzuwägen, ob ein gewünschtes Ergebnis mit den vorhandenen Daten überhaupt erreicht werden kann und ob das Ergebnis den benötigten Aufwand rechtfertigt (vgl. das folgende Beispiel 1.3);
- Präzision bei der Ergebniskommunikation auch bei knappen Resourcen und gegen Widerstände durchzusetzen, um Trugschlüsse zu verhindern und das Potenzial der Daten auszuschöpfen;
- „Analytical Fairness“ zu gewährleisten und z. B. Analysen zu unterlassen, wenn das Risiko eines Ergebnismissbrauchs hoch ist (vgl. das erwähnte Urteil zum Risiko-Scoring);
- bei Datenanalysen Objektivität als Grundhaltung zu bewahren, insbesondere in Situationen, in denen Datenlage und Fragestellung größeren Spielraum für die Analyse lassen;
- explizit oder implizit kommunizierte Interpretationen von Daten und Analyseergebnissen kritisch zu hinterfragen (auch eigenes Kontextwissen und dessen möglichen Einfluss auf die Interpretation);
- offen zu bleiben gegenüber neuen Erkenntnissen, auch wenn diese bisherigen Überzeugungen widersprechen.



Aufgaben 1.1-2

Beispiel 1.3: Ein fragwürdiges US-Projekt

Die DARPA (*Defense Advanced Research Projects Agency*), eine Forschungsbehörde des US-amerikanischen Verteidigungsministeriums, hat von 2014 - 2017 rund 67 Mio. US-Dollar in die automatisierte Erfassung und Auswertung der Webseiten von Escort-Services gesteckt, um Hinweise auf Zwangsprostitution zu finden (siehe [Artikel](#) von HUNDMAN / GOWDA / KEJRIWAL / BOECKING (2018)). In nur drei Fällen hat dieses große Anti-Trafficking-Projekt, an dem führende Forscher aus den Bereichen Datenwissenschaft und Künstliche Intelligenz mitgewirkt haben, zu einer Strafverfolgung geführt (vgl. [Magazinbeitrag](#) von BROWN (2019)). Dies zeigt eindrücklich, wie Ressourcen in hohem Maße verschwendet wurden, weil die Datenqualität und die Leistungsfähigkeit von Künstlicher Intelligenz massiv überschätzt wurden. Die aufgewendete Summe hätte z. B. besser für die Betreuung der Opfer von Zwangsprostitution eingesetzt werden können. Als Nebeneffekt führte die Kampagne dazu, dass alleinreisende Frauen in einigen Hotels unter Verdacht gerieten, Prostituierte zu sein.

Der Stellenwert, den das Thema „Data Literacy“ bzw. „Statistical Literacy“ inzwischen weltweit erlangt hat, spiegelt sich auch an Veränderungen der Lehrpläne von Schulen wider. Im Mathematikunterricht der Mittel- und Oberstufe weiterführender Schulen haben statistische Inhalte längst

Eingang in die Curricula gefunden. Einige Statistische Ämter haben E-Learning-Angebote konzipiert und implementiert, die statistische Basiskonzepte anhand amtlicher Daten illustrieren.

Erwähnenswert sind einige Projekte, die zur Verbesserung von Daten- und statistischer Methodenkompetenz beitragen. Das **Internationale Statistische Institut (ISI)**, eine nicht-kommerzielle Organisation zur Förderung internationaler Zusammenarbeit auf dem Feld der Statistik, hat das *International Statistical Literacy Project* initiiert, das auf die weltweite Vermittlung statistischer Grundkompetenzen bei Schülern abzielt. Das US-amerikanische *Consortium for the Advancement of Undergraduate Statistics Education (CAUSE)* stellt virtuelle Bibliotheken mit unterschiedlichen Ressourcen für die statistische Aus- und Weiterbildung bereit. Gleiches gilt für die ebenfalls als Open-Source-Sammlung angelegte *Statistics Online Computational Resource (SOCR)* der University of California in Los Angeles.

Projekte zur Förderung von Daten- und Methodenkompetenz

Exkurs 1.4: Kritik an Ergebnissen der PISA-Studien

Die seit 2000 in 3-jährigem Turnus weltweit laufenden **PISA-Studien** (PISA = Programme for International Student Assessment) zielen darauf ab, anhand großer ($n \geq 5000$ pro Land als Richtschnur) und möglichst repräsentativer Stichproben zu bewerten, inwieweit Schüler am Ende ihrer Pflichtschulzeit (Altersstufe 15 Jahre) für die Anforderungen unserer heutigen Wissensgesellschaft gerüstet sind. Operationalisiert wird diese Fähigkeit zur Anwendung von Wissen in realistischen Alltagssituationen durch eine standardisierte Messung der Leistungsfähigkeit in den Bereichen Lesen, Mathematik und Naturwissenschaften. Bei jeder Erhebung bildet reihum eines der drei genannten Kompetenzfelder den Schwerpunkt – in den Jahren 2000 und 2009 stand die Lesekompetenz im Vordergrund, 2006 und 2015 lag der Fokus auf den Naturwissenschaften, 2003 und 2012 sowie demnächst 2021 auf der Mathematik. Um den Einfluss sozioökonomischer Variablen auf den Bildungserfolg zu untersuchen, werden ergänzend auch Fragen zum familiären Hintergrund und zum schulischen Umfeld gestellt und ausgewertet.

Die letzte PISA-Studie fand 2018 statt. Erneut lag der Schwerpunkt auf der Messung von Lesekompetenzen. Deutschland schnitt in den drei Bereichen Lesefähigkeit, Mathematik und Naturwissenschaften vergleichsweise gut ab. Allerdings hatte sich an nicht-gymnasialen Sekundarschulen der Anteil der 15-jährigen Schüler mit Defiziten im Lesen und in Mathematik deutlich vergrößert. Die von deutschen Schülern erzielten Leistungen sind eng verknüpft mit dem sozialen Status und dem Migrationshintergrund.

Da die internationalen PISA-Studien wiederholt durchgeführt werden, werden ihre Ergebnisse – d. h. die in den drei Bereichen gemessenen und auf einer geeigneten Skala abgebildeten Schülerleistungen – zur Bewertung des Stands und der Entwicklung von Bildungssystemen herangezogen (kontinuierliches Bildungsmonitoring) und lösen Debatten zur Qualitätsverbesserung aus. Die

PISA-Resultate finden in Deutschland nicht zuletzt deswegen ein starkes Echo, weil sie einen Ergebnisvergleich für die einzelnen Bundesländer einschließen.

Es gibt aber auch kritische Kommentare zu den PISA-Studien, die entweder die technische Realisierung der Leistungsmessung betreffen oder aber den Grundgedanken der Steuerung von Bildungspolitik anhand standardisierter Tests. Erwähnt sei etwa ein Beitrag des Mathematikdidaktikers Th. JAHNKE in der *Neuen Zürcher Zeitung* vom 29. Januar 2012, in dem grundsätzliche Bedenken zur Methodik, zur Transparenz, zur Genauigkeit der Ergebnisse und vor allem zu den Zielen der PISA-Studien angemeldet werden. Erstaunlich ist, dass sich die breite öffentliche Diskussion bislang wesentlich auf die Rangplätze der Länder oder – auf nationaler Ebene – Regionen konzentrierte. Nach den Meta-Informationen, die das Zustandekommen der Rankings erst verständlich machen, wurde dabei kaum gefragt (vgl. auch die Beispiele 8.1 und 8.2).

Auch bei der öffentlichen Diskussion von Ergebnissen der PISA-Studie 2012 wurde Kritik daran laut, dass wesentliche Meta-Informationen ausgeblendet wurden – gemeint war hier die gegenüber 2003 veränderte Struktur der Schülerstichprobe. Der bessere Rangplatz Deutschlands sei, so der Tenor der kritischen Stimmen, möglicherweise allein oder überwiegend auf die veränderte Zusammensetzung der Stichprobe zurückzuführen, nicht aber notwendigerweise auf eine Verbesserung des Mathematikunterrichts.

1.4 Veränderungen in der Methodenausbildung

In der statistischen Aus- und Weiterbildung gab es in den letzten Jahren bemerkenswerte Veränderungen und neue Entwicklungen, die durch Fortschritte in der Informationstechnologie induziert wurden. Das klassische gedruckte Lehrbuch ist längst durch das „e-Buch“ ergänzt, wobei letzteres nicht immer einen erkennbaren Mehrwert gegenüber der gedruckten Version aufweist und das Internet oft nur als Transportmedium nutzt.

- | | |
|-----------------------------|--|
| Kostenfreie
Online-Kurse | Die Medien „Unterricht“ bzw. „Vorlesung“ haben in Online-Formaten eine Ergänzung gefunden. Inhalte einführender Statistikvorlesungen werden auf Online-Plattformen angeboten, z. B. bei <i>Coursera</i> , <i>EdX</i> , <i>Udacity</i> und der europäischen Plattform <i>iversity</i> . Die Kurse sind zumindest in der Basisversion überwiegend kostenfrei; Serviceleistungen – etwa die Ausstellung individualisierter Zertifikate – sind i. d. R. mit Gebühren verbunden. Die Geschäftsmodelle der Anbieter sind aber in Bewegung. |
|-----------------------------|--|

Die Euphorie, mit der diese unter dem Namen **MOOCs** (engl.: *Massive Open Online Courses*) bekannten Bildungsangebote anfangs aufgenommen wurden, ist inzwischen einer sachlichen Bestandsaufnahme gewichen. Zwei Standpunkte, die konträre Positionen in der Diskussion über MOOCs widerspiegeln, wurden in der Wochenzeitschrift *Die Zeit* in einem Artikel

vom 9. Januar 2014 von R. LANKAU und einem Beitrag von J. DRÄGER vom 5. Dezember 2013 veröffentlicht. Einen Vergleich konkurrierender MOOC-Plattformen einschließlich einer allgemeinen Bestandsaufnahme der Effekte von MOOCs findet man in einem Beitrag von J. POPE vom 15. Dezember 2014 in der *MIT Technology Review*. Es ist unbestritten, dass MOOCs sich inzwischen fest etabliert haben und mehr sind als nur ein flüchtiger Hype.

Außer umfassenden Online-Kursen in Form von MOOCs gibt es für die Statistikausbildung granulare Online-Lernmaterialien – auch als „Lern-Nuggets“ bezeichnete „Mini-Lernwelten“ – in der Gestalt von Animationen, interaktiven statistischen Experimenten oder Umgebungen zur dynamischen Datenvisualisierung – sowie Online-Sammlungen solcher Ressourcen. Ein Beispiel für letztere ist die schon erwähnte virtuelle Bibliothek **CAUSE** (*Consortium for the Advancement of Undergraduate Statistics Education*).

Granulare
Online-Ressourcen

Neben den MOOCs haben auch sog. **Webinare** in der Methodenausbildung ihren Platz gefunden. Der Name leitet sich aus *Web* und *Seminar* ab. Webinare sind Online-Seminare, die zu festgelegten Zeiten stattfinden und Interaktivität zwischen den Teilnehmern ermöglichen. Die Lehrkraft kann hier in einem Fenster am Bildschirm, für alle Teilnehmer sichtbar, mündlichen Vortrag mit Skizzen und Grafiken verknüpfen. Die Teilnehmer beteiligen sich unter Verwendung von Mikrofon und Web-Kamera. Ein Nachteil von Webinaren ist darin zu sehen, dass sie – zumindest bei erstmaliger Durchführung – mit größerem technischen Aufwand verbunden sind und bei unterschiedlicher Hardwareausstattung der Teilnehmer technische Probleme auftreten können.

Online-Seminare

Online-Kurse und kleinteilige Online-Ressourcen sind wie Webinare Formen computergestützten Lernens („e-Learning“). Als technische Plattform für die Präsentation und Distribution der Inhalte wird hier typischerweise ein Desktop-Computer verwendet. In neuerer Zeit werden aber auch zunehmend mobile Endgeräte eingesetzt („m-Learning“). Für mobile Endgeräte, vor allem für Smartphones mit kleinen Bildschirmen, müssen Lerninhalte möglichst kleinteilig und wenig textlastig sein. Für die Statistikausbildung eignen sich z. B. interaktive Experimente zu einzelnen statistischen Verfahren oder Modellen.

Einbezug mobiler
Endgeräte



Statistik-Web-App

Für das vorliegende Buch wurden zahlreiche interaktive Lernobjekte entwickelt und über QR-Codes eingebunden. Die einzelnen Lernobjekte sind voneinander unabhängig und sowohl für den Einsatz auf Desktops als auch auf mobilen Endgeräten konzipiert. Die Lernobjekte sind in zwei virtuellen Bibliotheken (Web-Apps) zusammengefasst, die sich auch für den Mathematikunterricht der Sekundarstufe II gut eignen (s. MITTAG 2017). Beide Apps sind im Vorwort dieses Buches näher beschrieben.



Abb. 1.5: *Statistisches Experiment für mobile Endgeräte*

„Blended Learning“ Heute werden bei der Vermittlung von statistischer Methodenkompetenz traditionelle Lehr- und Lernszenarien – Präsenzlehre, Einsatz von Printmaterialien – mit den beschriebenen Formen von e- und m-Learning zu integrierten Konzepten verknüpft („Blended Learning“). Die in der Praxis zu beobachtenden Blended-Learning-Konzepte können sich hinsichtlich der Gewichtung der einzelnen Komponenten des realisierten Medienmixes deutlich unterscheiden. Welcher Medienmix optimal ist, hängt von der Zielsetzung und der Zielgruppe ab.

Zunehmende Verwendung von R In der Statistikausbildung an Hochschulen, in der Forschung und in Unternehmen gibt es auch Veränderungen hinsichtlich der Software, die standardmäßig bei der Datenanalyse eingesetzt wird. Hier hat die Open-Source-Software R deutlich an Boden gewonnen und sich als Konkurrenz zu kommerzieller Statistiksoftware etabliert. Argumente für R sind neben der freien Verfügbarkeit der große Funktionsumfang und die Flexibilität. Für Anfänger ist allerdings das Fehlen einer grafischen Benutzeroberfläche ungewohnt.



2 Grundbegriffe der Statistik



Vorschau auf
das Kapitel

Es werden statistische Grundbegriffe vorgestellt, die bei der Datenerhebung wichtig sind, u. a. die Begriffe „Grundgesamtheit“ oder „Merkmal“. Außerdem erfolgt eine Klassifikation von Merkmalen nach der Anzahl der möglichen Ausprägungen (diskret vs. stetig), nach der Art der bei der Datenerfassung verwendeten Messskala (nominal-, ordinal- und metrisch skalierte Daten) sowie nach dem Typ der Merkmalsausprägungen (qualitativ vs. quantitativ).

Als Kriterien zur Beurteilung der Qualität von Messverfahren werden Objektivität, Reliabilität und Validität genannt und erläutert. Eingegangen wird auch auf die als Operationalisierung bezeichnete „Messbarmachung“ nicht direkt beobachtbarer Merkmale. Letztere werden auch als latente Variablen oder hypothetische Konstrukte bezeichnet.

2.1 Statistische Einheit, Merkmal und Grundgesamtheit

Wie jede Wissenschaft hat auch die Statistik ihre eigene Terminologie. Klare Begriffsbildungen sind notwendig, um den Rahmen, das Ziel und die Ergebnisse einer statistischen Untersuchung unmissverständlich zu beschreiben. Ausgangspunkt einer Untersuchung ist ein aus der Praxis oder der Forschung kommendes Problem. Die Problemlösung bedingt eine Konkretisierung des geplanten Untersuchungsablaufs. Erst nach sorgfältiger *Planung* kann die *Erhebung*, *Aufbereitung* und *Auswertung* von Daten erfolgen. In der Planungsphase gilt es festzulegen, welche Objekte Gegenstand einer Untersuchung sein sollen und welche Eigenschaften der Objekte von Interesse sind.

Wozu braucht man
eine statistische
Terminologie?

Manche Fragestellungen lassen sich bereits durch Auswertung vorhandenen Datenmaterials beantworten. Will man z. B. die Altersstruktur der Psychologen in Deutschland, deren Einsatzfelder und Träger der Beschäftigung (in eigener Praxis, in medizinischen Einrichtungen oder bei einer Behörde) sowie die geografische Verteilung untersuchen, so könnte man einfach die Mitgliederdateien von Berufsverbänden heranziehen, sofern diese frei zugänglich sind. Dennoch wären auch hier in der Planungsphase der Untersuchung noch Festlegungen zu treffen. So müsste entschieden werden, wie weit die Differenzierung bei den einzelnen Kategorien gehen sollte, z. B. bei der Untersuchung der räumlichen Verteilung nur Herunterbrechen auf Bundesländer oder auch tiefer.

Beispiel 2.1: Statistische Untersuchungen

Die Interdisziplinarität des Fachs „Statistik“ spiegelt sich auch in der Breite der Fragestellungen aktueller wissenschaftlicher Untersuchungen wider. Hier eine kleine Auswahl:

- In der Wirtschafts- und Sozialpolitik will man in einem Feldversuch neue Instrumente zur Bekämpfung von Jugenderwerbslosigkeit einsetzen und deren Effekt messen. Hier muss u. a. geklärt sein, welche Altersgruppe gemeint ist, wer als erwerbslos gilt und wie man Jugendliche in Ausbildung oder Umschulung behandelt.
- In der Sozialpsychologie wird untersucht, welche Determinanten die Bereitschaft zu ehrenamtlichem Engagement beeinflussen. Es ist hier festzulegen, welche Personengruppen man in die Untersuchung einbezieht, was an diesen Personen beobachtet wird und welche Untergruppen miteinander verglichen werden sollen.
- In der Fernsehforschung will man die Sehbeteiligung in Abhängigkeit von Alter und Tageszeit messen und auch das Ausbildungsniveau erwachsener Zuschauer berücksichtigen. Hier muss geklärt werden, welche Haushalte einbezogen werden, wie man den Ausbildungsstand erwachsener Haushaltsmitglieder misst und wie man die zunehmende Fernsehnutzung über mobile Endgeräte und Streamingdienste erfasst.
- In vielen Regionen Deutschlands mit starker Verkehrsbelastung wurden Umweltzonen eingerichtet. Ob ein Fahrzeug in diesen Zonen zugelassen ist, hängt von dessen Schadstoffausstoß ab. Bei der Einrichtung von Umweltzonen und der Schadstoffbewertung für Fahrzeuge ist zu klären, welche Schadstoffe zu messen sind und wo und wie man die Emissionen erfasst. Zu klären ist auch, ab welchen Schwellenwerten welche Nutzungseinschränkungen in einer Umweltzone verhängt werden.

Grundbegriffe

In der Statistik nennt man die Objekte, auf die sich eine statistische Untersuchung bezieht, **statistische Einheiten** oder **Merkmalsträger**. Daten werden also an statistischen Einheiten bzw. Merkmalsträgern erhoben. Die Menge aller für eine Fragestellung interessierenden statistischen Einheiten bildet eine **Grundgesamtheit**. Sie wird auch als **Population** bezeichnet. Wichtig ist, dass eine Grundgesamtheit klar abgegrenzt ist. Oft werden Teilmengen von Grundgesamtheiten (**Teilpopulationen**) betrachtet, etwa Differenzierung nach Geschlecht bei Untersuchungen zu delinquenter Verhalten bei Jugendlichen oder nach Fahrzeugtyp bei Untersuchungen zu Schadstoffemissionen im Straßenverkehr.



Video „Grundgesamtheit und statistische Einheiten“

Die Eigenschaften statistischer Einheiten werden **Merkmale** oder **Variablen** genannt. Die möglichen Werte, die ein Merkmal annehmen kann, heißen **Merkmalsausprägungen**. In der Statistik werden Merkmale üblicherweise mit Großbuchstaben gekennzeichnet, Merkmalsausprägungen mit Kleinbuchstaben.

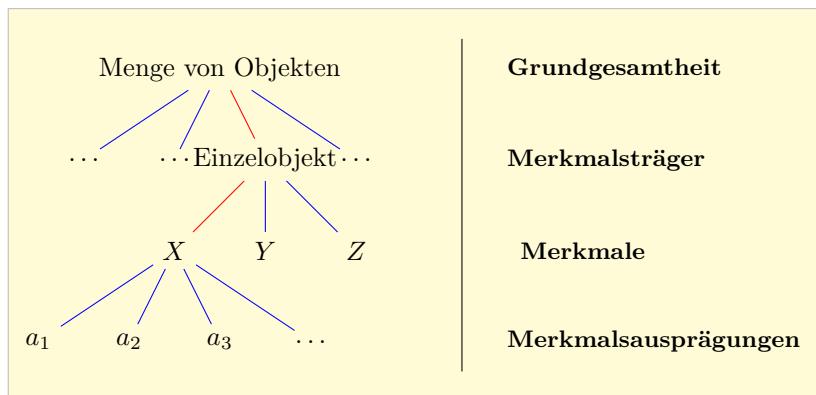


Abb. 2.1: Begriffshierarchien für statistische Grundbegriffe

Wählt man aus einer Grundgesamtheit nach einem Auswahlverfahren eine Teilmenge aus, spricht man von einer **Stichprobe**. Die in einer Grundgesamtheit oder einer Teilmenge einer Population beobachteten Werte für ein Merkmal nennt man **Daten!Roh-, Primärdaten** oder **Rohdaten**. Fasst man alle Urwerte in einer Liste zusammen, entsteht eine **Urliste**. In einer Urliste können Merkmalswerte mehrfach auftreten.

Wenn man, wie in der schließenden Statistik üblich, die Ausprägungen eines Merkmals als das Ergebnis eines Zufallsvorgangs interpretiert (Modellvorstellung), nennt man ein solches Merkmal **Zufallsvariable** und deren Ausprägungen auch **Realisierungen**. Beispiele für Realisierungen von Zufallsvariablen sind etwa die bei einer Serie von Roulettespielen beobachteten Ergebnisse oder die im März 2020 in Deutschland registrierten Fälle von Corona-Infektionen. In der schließenden Statistik werden auch die Elemente einer Stichprobe als Zufallsvariablen interpretiert und als **StichprobenvARIABLEN** bezeichnet.

Beispiel 2.2: Statistische Grundbegriffe

Eine Grundgesamtheit ist z. B. definiert durch

- alle Personen, die am 1. Mai 2018 in München ihren Erstwohnsitz hatten;
- Studierende einer Hochschule zu Beginn des Sommersemesters 2019, über die man via Telefonbefragung Informationen gewinnen will;
- die von einem stahlverarbeitenden Unternehmen im März 2017 produzierten Serienteile eines bestimmten Typs;
- die in Deutschland am 1. Januar 2020 gemeldeten PKWs.

Die statistischen Einheiten werden hier repräsentiert durch

- jede Person mit Erstwohnsitz in München am 1. Mai 2018;
- alle zum Sommersemester 2019 eingeschriebenen Studierenden;

- die im März 2017 gefertigten Serienteile;
- jeder am 1. Januar 2020 in Deutschland gemeldete PKW.

Interessierende Merkmale und Merkmalsausprägungen können hier sein

- der Familienstand der Person, etwa mit der Ausprägung „verheiratet“;
- das Alter der Studierenden, erfasst z. B. in Form von Altersklassen;
- der Durchmesser des Serienteils, etwa mit der Ausprägung 112 mm;
- die von einem PKW emittierte Menge an CO_2 in g/km.

Die Merkmalsausprägungen „verheiratet“ oder „Altersgruppe 20 – 24 Jahre“ werden bei der hier betrachteten Grundgesamtheit zweifellos mehrfach auftreten. Die Urliste, in der die Werte für das Merkmal „Familienstand“ zusammengefasst werden, enthält viele Elemente, aber nur wenige unterschiedliche Ausprägungen. Ob sich auch bei einer Urliste für ein Längenmaß Wiederholungen ergeben, hängt davon ab, mit welcher Präzision gemessen wird. Misst man z. B. nicht in Millimetern, sondern in Mikrometern, erhält man seltener gleiche Messwerte.



Aufgabe 2.1

2.2 Merkmalklassifikationen und Skalen

Einteilung von
Merkmälern nach der
Anzahl der
Ausprägungen



Video „Diskrete
Zufallsvariablen“

Merkmale lassen sich nach verschiedenen Kriterien klassifizieren. Ein erstes Einteilungskriterium ist die *Anzahl der möglichen Ausprägungen*. Man unterscheidet hier zwischen diskreten und stetigen Merkmälern.

Ein **diskretes Merkmal** ist ein Merkmal, das nur endlich viele Ausprägungen oder höchstens abzählbar unendlich viele Ausprägungen annehmen kann.¹ Zählvariablen sind stets diskret. Ein **stetiges Merkmal** ist hingegen dadurch gekennzeichnet, dass die Ausprägungen ein Intervall bilden. Für je zwei Merkmalsausprägungen eines stetigen Merkmals gilt, dass auch alle Zwischenwerte angenommen werden können. Die Unterscheidung von diskreten und stetigen Merkmälern lässt sich insbesondere auch auf Zufallsvariablen beziehen.

Ob ein Merkmal diskret oder stetig ist, hängt nicht davon ab, wie das Merkmal in der Praxis tatsächlich angegeben wird. Die Körpergröße ist z. B. stetig, obwohl man sie in der Praxis kaum genauer als auf volle Zentimeter gerundet ausweist. Ähnliches gilt für die Größe einer Wohnung, die meist in vollen Quadratmetern angegeben wird.² Generell kann man jedes stetige Merkmal durch Rundung oder Gruppierung in diskrete Variablen überführen, wobei damit immer ein Informationsverlust

¹Der Fall „abzählbar unendlich“ ist für die Praxis von geringerer Relevanz. Eine Menge heißt *abzählbar unendlich*, wenn sich ihre Elemente umkehrbar eindeutig auf die Menge der natürlichen Zahlen abbilden lassen. Die Elemente einer abzählbar unendlichen Menge lassen sich fortlaufend nummerieren. Beispiele sind die Menge der Primzahlen oder die der geraden ganzen Zahlen.

²Solche Merkmale werden gelegentlich auch als *quasi-stetig* bezeichnet. Diese Bezeichnung wird aber im vorliegenden Manuskript nicht weiter verwendet.

einhergeht. So wird man das Bruttojahreseinkommen von Arbeitnehmern eines größeren Landes der Eurozone anhand von Einkommensklassen erfassen, also auf die Angabe der exakten Merkmalswerte (Rohdaten oder Urwerte) in Euro und Cent verzichten. Die Klassenmitten werden dann bei der Datenanalyse als Repräsentanten für die jeweilige Klasse verwendet. Mit der Bildung von Klassen erreicht man vor allem bei größeren Datensätzen für stetige Merkmale mehr Übersichtlichkeit, kann dann aber innerhalb der Klassen nicht mehr differenzieren.

Beispiel 2.3: Diskrete und stetige Merkmale

Diskret sind z. B. die Anzahl der Fachsemester von Studierenden, Güteklassen bei Lebensmitteln oder Hotels, der Familienstand einer Person oder die Anzahl der zu einem Haushalt gehörenden Personen.

Stetig sind Zeitangaben, Längen, Gewichte oder Merkmale zur Quantifizierung der Schadstoffbelastung von Luft und Wasser. Monetäre Größen, etwa Bruttoeinkommen oder Mietpreise in Euro und Cent, sind ebenfalls stetige Merkmale. Auch hypothetische Konstrukte bzw. deren Operationalisierungen, in der Psychologie etwa das Merkmal „Intelligenzquotient einer Person“, werden häufig als stetige Variablen interpretiert.

Eine zweite Merkmalsklassifikation basiert auf der *Art der verwendeten Messskala*. Man unterscheidet drei Skalenniveaus, nämlich Nominalskalen, Ordinalskalen und metrische Skalen.

Einteilung von
Merkmälern nach der
Skalierung

Eine **Nominalskala** ist eine Messskala, bei der die Ausprägungen eines Merkmals lediglich Namen oder Kategorien darstellen, etwa Branchenzugehörigkeit von Arbeitnehmern, das Studienfach von Studierenden oder das Transportmedium von Pendlern. Nominalskalierte Daten sind Daten, die anhand einer Nominalskala erfasst werden. Typisch für sie ist, dass es keine natürliche Rangordnung gibt. Die Bildung von Differenzen oder Quotienten ist bei nominalskalierten Daten nicht sinnvoll.



Bei einer **Ordinalskala** oder **Rangskala** gibt es hingegen eine natürliche Rangordnung, aber die Differenzen- und Quotientenbildung ist ebenfalls nicht sinnvoll erklärt. Beispiele für ordinalskalierte Daten sind Schulnoten oder Bonitätsbewertungen von Sparkassenkunden auf einer mehrstufigen Skala. Es gibt hier zwar eine Rangordnung zwischen den Stufen, Abstände zwischen zwei Stufen sind aber nicht direkt vergleichbar.



Video „Skalentypen“

Eine **metrische Skala** oder **Kardinalskala** ist dadurch gekennzeichnet, dass hier auch Abstände (Differenzen) zwischen den Merkmalsausprägungen interpretierbar sind. Eine metrische Skala heißt **Verhältnisskala** oder **Ratioskala**, wenn ein natürlicher Nullpunkt existiert; ansonsten spricht man auch von einer **Intervallskala**. Temperaturmessungen in

°Celsius erfolgen z. B. auf einer Intervallskala. Letzteres impliziert, dass die Bildung von Quotienten aus zwei Merkmalsausprägungen nicht sinnvoll ist. Das Merkmal „Geschwindigkeit“ ist hingegen ein Merkmal mit natürlichem Nullpunkt. Aussagen des Typs 100 km/h ist doppelt so schnell wie 50 km/h sind hier zulässig, d. h. auch die Division ist erklärt. Ein Spezialfall der Verhältnisskala ist die **Absolutskala**. Bei dieser gibt es außer einem natürlichen Nullpunkt zusätzlich eine natürliche Einheit. Das Merkmal „Anzahl der Fachsemester“ ist ein solches Merkmal.

		sinnvolle Operationen			
	Skala	auszählen	ordnen	Differenz bilden	Quotienten bilden
Nominalskala		ja	nein	nein	nein
Ordinalskala		ja	ja	nein	nein
Metrische Skala	Intervall-skala	ja	ja	ja	nein
	Verhältnis-skala	ja	ja	ja	ja
	Absolut-skala	ja	ja	ja	ja

Tab. 2.1: Sinnvoll interpretierbare Operationen bei verschiedenen Skalen

Die genannten Skalenniveaus stellen eine Hierarchie dar, bei der die Nominalskala das niedrigste Niveau und die Verhältnisskala – bzw. die Absolutskala als Sonderfall der Verhältnisskala – das höchste Niveau repräsentiert. Man kann ein Merkmal, das ordinalskaliert ist, auch auf einer Nominalskala messen und ein metrisch skaliertes Merkmal stets auch auf einer Ordinalskala oder Nominalskala – allerdings in beiden Fällen unter Informationsverlust. Genannt sei als Beispiel das Merkmal „Bruttojahreseinkommen“, das man in Euro und Cent erfassen kann (metrische Skala) oder über wenige Einkommensklassen. Wenn bei der Erfassung des Merkmals „Einkommen“ nur die Zugehörigkeit zu Einkommensbereichen abgefragt wird, kann man das Merkmal nur noch als ordinalskaliert behandeln und Einkommensunterschiede zwischen zwei Personen nicht mehr in Euro und Cent beziffern.

Likert-Skala:
Datenerhebungs-instrument der empirischen Sozialforschung In den Sozialwissenschaften, in der Markt- und Meinungsforschung sowie in der Psychologie misst man häufig persönliche Einstellungen oder Empfindungen, also nicht direkt beobachtbare Variablen, die den Charakter hypothetischer Konstrukte haben (**latente Variablen**) – etwa die individuellen Ausprägungen der Merkmale „Leistungsmotivation“, „Lebenszufriedenheit“ oder „Umweltbewusstsein“. Dazu legt man den Personen Aussagen vor (sog. „Items“) und erfasst den Grad der Zustim-

mung oder Ablehnung dieser Aussagen anhand einer mehrstufigen, von „trifft zu“ bis „trifft nicht zu“ reichenden Skala. Die Anzahl der Stufen bei einer solchen Skala, die nach dem amerikanischen Sozialforscher Rensis LIKERT (1903 - 1981) auch **Likert-Skala** genannt wird, kann ungerade oder gerade sein. Die Stufen lassen sich anhand von Zahlen codieren. Bei einer ungeraden Anzahl von Stufen steht eine neutrale Bewertung in der Mitte der Skala, während bei gerader Stufenzahl eine neutrale Position ausgeschlossen wird. Die Antworten auf einer Likert-Skala sind ordinal-skaliert, weil man nicht voraussetzen kann, dass die Abstände zwischen den einzelnen Stufen gleich sind. Likert-Skalen werden auch in der Medizin verwendet, etwa zur Einschätzung von Schmerzintensitäten.

Eine Likert-Skala ist jedenfalls keine grundsätzlich andere Skala, die die in Tabelle 2.1 wiedergegebene Klassifikation erweitert. Sie ist vielmehr ein in der empirischen Sozialforschung und der Psychologie sowie der Medizin für Befragungen häufig anzutreffendes Instrument zur Erhebung ordinalskalierter Daten. In der Praxis wird allerdings oft – ähnlich wie bei Schulnoten – Äquidistanz der Stufen unterstellt. Mit dieser Annahme wird dann bei Daten, die anhand einer Likert-Skala gewonnen wurden, die Anwendung von Operationen gerechtfertigt, welche eigentlich nur für metrisch skalierte Daten zulässig sind, z. B. die Mittelwertbildung.

Da die Werte einer Likert-Skala auf Einschätzungen beruhen (engl.: *rating*), spricht man auch von einer **Ratingskala**. Der Begriff der Ratingskala wird allerdings nicht nur im Zusammenhang mit der Messung von persönlichen Einstellungen und Empfindungen verwendet, sondern bekanntermaßen auch bei der Bewertung der Bonität von Staaten, Unternehmen oder individuellen Kreditnehmern.

Beispiel 2.4: Skalenniveaus für Merkmale

Weitere Beispiele für Merkmale mit unterschiedlicher Skalierung:

- Nominalskalierte Merkmale sind „Parteipräferenz von Wählern“, „Konfessionszugehörigkeit“, „Geschlecht“.
- Ordinal- oder rangskaliert sind „Militärischer Rang“ oder „Höchster erreichter Bildungsabschluss“. Auch das Merkmal „Temperatur“ kann als rangskaliert behandelt werden, wenn man nur zwischen „kalt, normal, warm, heiß“ unterscheidet. Ebenfalls ordinalskaliert sind die Antworten zu Aussagen, die anhand einer fünfstufigen Likert-Skala mit den Stufen „trifft zu (1)“ – „trifft eher zu (2)“ – „weder noch (3)“ – „trifft eher nicht zu (4)“ – „trifft nicht zu (5)“ gewonnen werden.
- Metrisch sind „Geburtsjahr“ (Intervallskala) und „Lebensalter“ oder „CO₂-Emissionen von PKWs“ (Verhältnisskala).



Aufgabe 2.2

Einteilung von
Merkmälern
nach dem Typ der
Ausprägungen

Eine weitere Klassifikation für Merkmale bezieht sich auf den *Typ der Merkmalsausprägungen* (Kategorie oder Zahl). Wenn die Ausprägungen *Kategorien* sind, spricht man von einem **qualitativen Merkmal**. Die Merkmalsausprägungen spiegeln hier eine Qualität wider, keine Intensität. Ein qualitatives Merkmal kann nominal- oder ordinalskaliert sein – im ersten Falle sind die Kategorien ungeordnet (z. B. beim Merkmal „Konfessionszugehörigkeit“), im zweiten geordnet (z. B. „Hotelkategorie“). Auch wenn den Ausprägungen qualitativer Merkmale für die statistische Analyse oft Zahlencodes zugeordnet werden (etwa „2“ für „Familienstand = verheiratet“), sind die Zahlen nur Etiketten, mit denen man nicht im üblichen Sinne rechnen kann. Sind die Ausprägungen eines Merkmals hingegen „echte“ *Zahlen*, so liegt ein **quantitatives Merkmal** vor. Metrisch skalierte Merkmale sind stets quantitativ.

2.3 Operationalisierung von Merkmalen

Bevor eine Variable gemessen wird, ist ihre Messbarkeit zu sichern. Dies geschieht durch die als **Operationalisierung** bezeichnete Festlegung von Messanweisungen. Vor allem bei latenten Variablen ist die Operationalisierung nicht trivial – es gibt hier i. Allg. mehr als eine Möglichkeit. In jedem Falle geht es darum, ein Verfahren zu spezifizieren, mit dem sich ein Merkmal quantifizieren lässt.

Qualitätsbewertung
für Messverfahren

Die Beurteilung der Qualität von Messverfahren erfolgt anhand dreier Kriterien. Es sind dies die **Objektivität** (intersubjektive Nachvollziehbarkeit), die **Reliabilität** (Messgenauigkeit) sowie die **Validität** (Gültigkeit) des Verfahrens. Von letzterer spricht man, wenn wirklich das gemessen wird, was man messen will. Validität bezieht sich also auf den inhaltlichen Aspekt der Messung, während die Reliabilität auf die technische Ebene abstellt. Ein nicht-reliables Messverfahren ist i. Allg. auch nicht-valide und auch ein hoch-reliables Messverfahren kann wenig valide sein. Letzteres trifft zu, wenn ein Verfahren zwar etwas genau misst, aber inhaltlich etwa anderes erfasst werden sollte.

Eine detaillierte Behandlung der genannten Gütekriterien findet man z. B. bei SEDLMEIER / RENKEWITZ (2018).



Beispiel 2.5: Operationalisierung latenter Variablen

Die Notwendigkeit der Operationalisierung von Merkmalen zeigt sich z. B. bei der Formulierung und Überprüfung von Forschungshypothesen. Wenn man etwa postuliert, dass ein höherer Bildungsstand i. d. R. mit einem höheren Einkommen verknüpft ist, muss vor einer Überprüfung der Hypothese geklärt werden, wie man das nicht direkt beobachtbare Merkmal „Bildungsstand einer Person“ messen will. Dazu wird üblicherweise ein messbares Merkmal als Proxyvariable herangezogen, d. h. eine näherungsweise verwendbare beobachtbare Variable. Für das Merkmal „Bildungsstand“ kämen etwa der höchste erreichte Bildungsabschluss oder die Anzahl der erfolgreich an Bildungsinstitutionen verbrachten Jahre als Proxyvariablen in Betracht. Bei der Messung der Rechenfertigkeit von Schülern wird man auf geeignete Mathematikaufgaben zurückgreifen, von denen man annimmt, dass sie einzelne Aspekte der latenten Variablen treffen, etwa die Fähigkeit Rechenfähigkeiten in Alltagssituationen anwenden zu können.

Schwieriger ist die Operationalisierung latenter Variablen, wenn diese unterschiedlich interpretierbar sind oder Sinnfragen berühren. Im *Guardian* wurde im Juli 2012 ein Interview mit dem Ökonomen R. LAYARD und dem Philosophen J. BAGGINI wiedergegeben, bei dem die Frage der Messbarkeit von „Glück“ sehr kontrovers diskutiert wurde. Um Messbarkeit zu erreichen, muss das oft überstrapazierte Konstrukt „Glück“ von verwandten Konstrukten wie „Wohlbefinden“ oder „subjektive Lebenszufriedenheit“ abgegrenzt werden. Daten zur subjektiven Lebenszufriedenheit werden z. B. im Rahmen des (*Sozioökonomischen Panels (SOEP)*) gewonnen. Der seit 2011 alljährlich von der Deutschen Post veröffentlichte *Glücksatlas Deutschland* fasst die Ergebnisse für die einzelnen Bundesländer in Form eines Indexes für subjektive Zufriedenheit zusammen („Glücksindex“) und verknüpft die Zahlenwerte mit einer interaktiven Karte.

Aber selbst bei der Messung von Merkmalen, die direkt beobachtbar sind (**manifeste Variablen**) – z. B. das Bruttoeinkommen von Arbeitnehmern – ist es wichtig, genau zu spezifizieren, was gemessen werden soll. Es ist ein Verdienst von *Eurostat*, dem Europäischen Amt für Statistik, eine Harmonisierung der in Europa von Statistischen Ämtern erhobenen Daten zu sichern. Die Harmonisierung erfolgt über EU-Verordnungen, die in den Mitgliedstaaten Rechtskraft besitzen. Die Verordnungen regeln, welche Komponenten zu einer Variablen gehören und welche nicht. Dies sichert die Vergleichbarkeit von Daten über Ländergrenzen hinaus und macht die amtliche Statistik von aktuellen Politiken nationaler Regierungen unabhängiger. Welche Regierung sähe z. B. nicht gerne vor Wahlen positive Zahlen für den Arbeitsmarkt? Eurostat besitzt Vollmachten für das Monitoring von Daten, die für die Stabilität der Eurozone besonders relevant sind – etwa Daten zur Entwicklung von Staatsschulden. Die Europäisierung der amtlichen Statistik wirkt der möglichen Manipulation durch Veränderung der Operationalisierung von Merkmalen entgegen.



Beispiel 2.6: Operationalisierung in der amtlichen Statistik

Bei der Erfassung von Bruttoeinkommen in der EU gilt es zu klären, welche Einkommensanteile einzubeziehen, wann sie zu verbuchen sind und auf welche Branchen oder Branchenaggregate sich die Datenerfassung beziehen soll. Die einschlägige Kommissionsverordnung regelt z. B., dass staatliche Sozialtransferszahlungen, etwa das Kindergeld, nicht als Einkommenskomponente gelten, Sonderzahlungen wie Weihnachts- oder Urlaubsgeld jedoch zählen. Schwierig ist auch die Bewertung von Aktienoptionen als Einkommenskomponente. Um mittlere Stundenverdienste zu errechnen, muss man bei Lehrern regeln, wie die häusliche Vorbereitung von Unterricht zeitlich zu bewerten ist, und bei Fabrikarbeitern ist zu klären, ob Pausenzeiten als Arbeitszeit gelten.



Interaktives Objekt
„Erwerbstätigkeit“

Politisch brisanter ist die Operationalisierung von Erwerbs- oder Arbeitslosigkeit. Als *erwerbslos* gilt nach der z. Z. angewandten Definition der **International Labour Organization (ILO)** in Genf eine Person im erwerbsfähigen Alter, die weniger als eine Stunde wöchentlich gegen Entgelt (beliebiger Höhe) arbeitet und aktiv auf der Suche nach mehr Arbeit ist. Als *erwerbsfähig* werden Personen angesehen, die der Altersklasse von 15 - 64 Jahren angehören – häufig wird auch die Altersklasse 20 - 64 Jahre zugrunde gelegt. Die Quote der Erwerbstätigen bzw. der Erwerbslosen wird über Telefonumfragen im Rahmen des Mikrozensus erfasst. Die Erwerbslosenquote wird oft mit der Quote der registrierten Arbeitslosen verwechselt, die von der Nürnberger **Bundesagentur für Arbeit (BA)** erfasst wird. Eine Person gilt als *arbeitslos*, wenn sie vorübergehend in keinem Beschäftigungsverhältnis steht und sich als arbeitslos registrieren ließ. Die Registrierung erfolgt nur, wenn mindestens 15 Arbeitsstunden pro Woche angestrebt werden. Die Statistiken zur Arbeitslosigkeit sind umfassender als die zur Erwerbslosigkeit und ermöglichen auch aussagekräftige Vergleiche zwischen Regionen. Das **Statistische Bundesamt** weist sowohl die europaweit angewendete Erwerbslosenstatistik als auch die Arbeitslosenzahlen der BA aus.



Statistik-App
der BA

Damit das einem Datensatz der amtlichen Statistik zugrunde liegende Messverfahren nachvollziehbar ist, werden die Daten in der amtlichen Statistik durch Meta-Daten ergänzt, die den methodischen Hintergrund und eventuelle Besonderheiten der Datenerfassung offen legen. Wenn sich etwa die Bruttoverdienste für eine Branche in einem EU-Land auf alle in dem Wirtschaftszweig tätigen Arbeitnehmer beziehen, in einem anderen Land aber nur auf Arbeitnehmer, die in Unternehmen einer bestimmten Mindestgröße tätig sind, so wird dieser die Vergleichbarkeit der Ergebnisse einschränkende Unterschied als Meta-Information zusammen mit den Daten ausgewiesen.



3 Datengewinnung und Auswahlverfahren



Vorschau auf
das Kapitel

Im Zentrum dieses Kapitels stehen Klassifikationen für Datenerhebungen. Man kann z. B. danach differenzieren, ob sich eine Erhebung auf selbst gewonnene Daten stützt (Primärerhebung) oder bereits vorhandene Daten nutzt (Sekundärerhebung). Die Erhebung eigener Daten kann anhand einer Befragung, via Beobachtung oder per Experiment erfolgen. Beim Experiment werden Einflussgrößen planmäßig variiert und die damit verbundenen Effekte gemessen.

Weitere Klassifikationen für Erhebungen beziehen sich auf den zeitlichen Zusammenhang der Daten (Querschnitts- vs. Längsschnittdaten) oder auf den Umfang der erhobenen Daten (Teil- vs. Vollerhebung). Für den in der Praxis dominierenden Fall der Teilerhebung (Verwendung von Stichproben) werden zufallsgesteuerte und systematische Auswahlprozeduren vorgestellt und der Begriff der Inferenz erläutert. Einige praxisrelevante mehrstufige Auswahlverfahren werden ausführlicher präsentiert. Näher eingegangen wird auch auf Fehlschlüsse, die in der statistischen Praxis bei der Gewinnung, Interpretation und Verarbeitung von Daten häufig auftreten.

Am Ende des Kapitels findet der Leser eine Übersicht über wichtige Institutionen, die auf nationaler oder supranationaler Ebene amtliche oder nicht-amtliche Daten sammeln und der Öffentlichkeit zur Verfügung stellen.

3.1 Erhebungsarten und Studiendesigns

Für die empirische Überprüfung von Forschungsfragen werden **Daten** benötigt, d. h. Werte eines Merkmals oder mehrerer Merkmale in einer Grundgesamtheit von Merkmalsträgern. Die Qualität der Aussagen, die sich aus der Analyse statistischer Daten ableiten lassen, hängt wesentlich von der Datenqualität ab. Die Vorgehensweise bei der Datengewinnung ist daher bei einer statistischen Untersuchung sorgfältig zu planen. Die Gewinnung von Daten bezeichnet man als **Datenerhebung**, während die Planung der Datengewinnung **Erhebungsdesign** genannt wird.

Datenerhebungen lassen sich nach verschiedenen Kriterien klassifizieren. Nach der Art der Datenquelle unterscheidet man zwischen Primär- und Sekundärerhebungen. Bei **Primärerhebungen** werden die Daten eigens für das jeweilige Untersuchungsziel gewonnen. Dieser Verfahrensweise begegnet man z. B. in der Arzneimittelforschung oder der Psychologie. Bei **Sekundärerhebungen** wird hingegen auf Daten aus schon vorhandenen Quellen zurückgegriffen. Gelegentlich spricht man auch von **Tertiärerhebungen**, nämlich dann, wenn statistische Information aus Sekundärerhebungen noch transformiert oder aggregiert wird.

Klassifikation von
Erhebungen
hinsichtlich der
Datenquelle

Beispiel 3.1: Primär-, Sekundär- und Tertiärerhebungen

Die regelmäßig erscheinenden Berichte des Münchener IFO-Instituts zum aktuellen Geschäftsklima in Deutschland beziehen sich auf *Primärerhebungen*, denn sie basieren auf Daten, die direkt für die Erstellung der Berichte erhoben werden. Statistische Analysen, die sich z. B. auf Daten des Statistischen Bundesamts stützen, sind *Sekundärerhebungen*.

Die *Europäische Gehalts- und Lohnstrukturerhebung* basiert auf Individualdaten für Millionen von Arbeitnehmern in europäischen Staaten. Die Daten werden von nationalen Statistikämtern erhoben und von *Eurostat* zusammengeführt. Aufgrund der Vertraulichkeit der originären Mikrodaten werden die Daten anonymisiert und bearbeitet, also nur in Form *tertiärstatistischer Daten* freigegeben. Die Aussagekraft statistischer Auswertungen ist natürlich reduziert, wenn die ursprünglich vorhandene statistische Information verkürzt wird.

Arten der Datengewinnung Daten können auf unterschiedliche Weisen gewonnen werden, insbesondere per Befragung, durch Beobachtung von Personen, Objekten und Prozessen im Feld oder anhand von Experimenten.

Varianten der Befragung Die **Befragung** ist das dominierende Instrument sozialwissenschaftlicher Forschung. Sie lässt sich mündlich (persönlich oder per Telefon), schriftlich und auch internetgestützt durchführen. Eine *mündliche Befragung* kann unstrukturiert, teilstrukturiert oder strukturiert erfolgen. Eine *unstrukturierte Befragung* hat einen offenen Charakter und kann ohne Fragebogen realisiert werden. Bei *teilstrukturierten* und *strukturierten* Interviews ist die Befragung teilweise oder ganz standardisiert. Dies lässt sich durch die Verwendung von Fragebögen mit teilweise oder vollständig vorgegebenen Antwortalternativen erreichen. Im letztgenannten Fall spricht man auch von *geschlossenen Fragen*.

Mündliche Befragungen lassen sich mit modernen Kommunikationstechnologien verknüpfen. So kann etwa eine direkte oder telefonische Befragung per Interview mit softwaregesteuerter Interviewführung und automatisierter Ergebnisverarbeitung erfolgen. In der Literatur findet man in diesem Kontext die Abkürzungen **CAPI** (*computer assisted personal interviewing*) für das persönlich geführte Interview mit Notebook oder Tablet und **CATI** (*computer assisted telephone interviewing*) für das fernmündlich geführte Interview, bei dem der Interviewer mit Sprecheinrichtung vor dem Computer sitzt und die Antworten der befragten Person direkt eingibt. Die *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)* erfolgt z. B. auf CAPI-Basis. In beiden Fällen spricht man von einem *interviewer-administrierten Interview*, weil die Antworteingabe am Computer vom Interviewer vorgenommen wird.¹ Abbildung

¹Der *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM)* gibt auf seiner Internetseite an, dass 2018 in den Mitgliedsinstituten des ADM mehr als 3 800 CATI-Plätze eingerichtet und über 8 500 CAPI-Geräte im Einsatz waren.

3.1 zeigt ein Telefonstudio eines Marktforschungsinstituts, in dem fernmündliche Befragungen durchgeführt werden.



Abb. 3.1: Befragungen via Telefonstudio (CATI); Quelle: TNS Infratest

Bei der *schriftlichen* Befragung werden Fragebögen per Post oder per E-Mail an ausgewählte Adressaten verteilt oder auf einer Internetseite bereitgestellt. Netzbasierte schriftliche Befragungen können interaktive Programme sein, die den Befragten flexibel durch einen Fragenkatalog führen. Da der Befragte die Antworteingabe selbst vornimmt, spricht man auch von einer *selbst-administrierten* computergestützten Befragung.

Auch die **Beobachtung** ist ein verbreitetes Verfahren der Datenerhebung. Beobachtung kann sich auf ganz unterschiedliche Objekte beziehen, etwa auf Volkswirtschaften, auf technische Prozesse in Unternehmen, auf Umweltbelastungen oder auf das Verhalten von Personen. In den *Wirtschaftswissenschaften* werden z. B. Aktienindizes, Inflationsraten oder Beschäftigungsquoten fortlaufend verfolgt, wobei die Beobachtung mit Maßnahmen verbunden sein kann, etwa mit Interventionen durch die Europäische Zentralbank. Bei der industriellen *Qualitätssicherung* werden Fertigungsprozesse kontinuierlich beobachtet und dokumentiert, i. d. R. automatisiert unter Einsatz moderner Messtechniken, mit dem Ziel der Vermeidung nicht-spezifikationskonformer Produkte.

Wo werden Daten per Beobachtung gewonnen?

Abbildung 3.2 illustriert dies anhand eines Fotos aus der Produktion von Kraftfahrzeugen. Das Bild zeigt den Einsatz robotergesteuerter optischer Sensoren bei der Fertigung von Karosseriekomponenten. Mit den Sensoren erfolgt eine berührungslose Überprüfung von Abmessungen der Anbauteile. So werden Unterbrechungen der Fertigungslinie vermieden und Qualitätsprobleme sofort identifiziert.

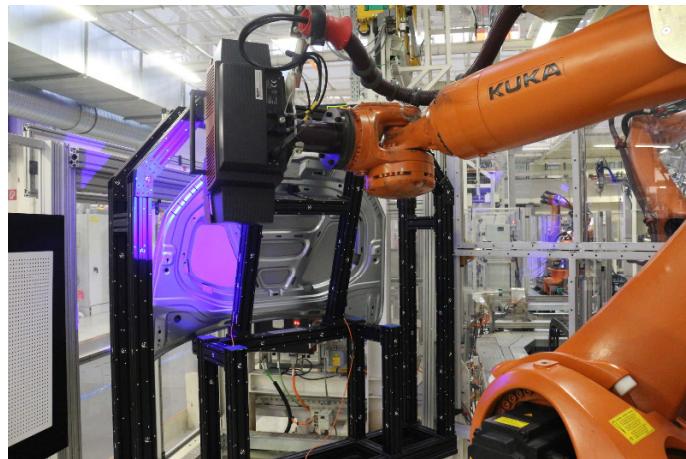


Abb. 3.2: Berührungslose Datenerfassung bei der Fertigung von Kraftfahrzeugen (Quelle: Volkswagen AG)

Beim *Umweltmonitoring* werden Schadstoffemissionen gemessen. Die Daten werden bei internationalen Klimakonferenzen als Diskussionsbasis herangezogen und auch auf nationaler Ebene für Politikentscheidungen verwendet, in Deutschland z. B. bei der Verhängung von Fahrverboten für ältere Diesel-Kraftfahrzeuge oder bei Verkehrsbeschränkungen in Zonen mit hoher Feinstaubbelastung. In den *Sozialwissenschaften* und der *Psychologie* geht es um die Beobachtung von Einzelpersonen oder Gruppen. Die Beobachtung kann offen oder verdeckt erfolgen. Typisch für Beobachtungen in der empirischen Sozialforschung ist, dass die Beobachtungen systematisch geplant und dokumentiert werden und einem spezifizierten Forschungszweck dienen. Für die Dokumentation der Beobachtungen bedient man sich eines Beobachtungsprotokolls.

Die verdeckte Beobachtung oder auch die Auswertung von Verhaltensspuren – z. B. die Durchführung von Logfile-Analysen zur Untersuchung des Verhaltens von Internetnutzern – sind **nicht-reaktive Erhebungsverfahren**. Hierunter versteht man Erhebungstechniken, die keine Veränderungen bei den zu untersuchenden Objekten hervorrufen. Bei der verdeckten Beobachtung von Personen nehmen diese i. d. R. gar nicht wahr, dass sie Gegenstand einer Beobachtung sind. Hier sind natürlich ethische und datenschutzrechtliche Richtlinien zu beachten.

Anwendungsfelder nicht-reaktiver Erhebungsverfahren

In der *Markt- und Konsumforschung* werden nicht-reaktive Methoden unter Nutzung moderner Datenverarbeitungstechnologie eingesetzt. *Google Analytics* ist z. B. ein Informationsdienst, der Verhaltensspuren im Internet auswertet. Als weiteres Beispiel genannt seien Frequenzzählungen oder Aufzeichnungen von Blickbewegungen von Kunden in den Gängen von Supermärkten, mit denen Unternehmen Informationen zur Optimierung des Warensortiments gewinnen. Auch Geoinformationssysteme

werden zur Identifikation raumbezogener Zusammenhänge genutzt, etwa bei der Messung von Pendlerströmen.

Biometrische Daten sind ebenfalls den Beobachtungsdaten zuzurechnen. Sie werden zur Identifikation von Personen verwendet, z. B. als Berechtigungs nachweis für den Zugang zu Gebäuden oder mobilen Endgeräten. Erwähnt sei der elektronische Reisepass, der in Deutschland ein digitalisiertes biometrisches Passbild und – optional – Fingerabdrücke speichert. Ebenso kann beim Bezahlen eines Einkaufs an der Kasse mit einer Bankkarte zur Verifikation der auf dem Magnetstreifen gespeicherten Daten ein Fingerabdruck verwendet werden.

Verwendung
biometrischer Daten



Abb. 3.3: Bezahlen im Supermarkt per Irisscan
(Quelle: José Giribás Marambio, Fotograf in Berlin)

Neuerdings kommen auch vermehrt Irisscanner zum Einsatz. Abbildung 3.3 zeigt ein solches Gerät im Supermarkt eines Camps des Flüchtlings hilfswerks UNCHR der Vereinten Nationen in Jordanien. Die individuellen Irisdaten der Campbewohner sind auf einer Datenbank gespeichert. Im Rahmen des Welternährungsprogramms tätigt das UNHCR monatliche Zahlungen, die ebenfalls über die Datenbank laufen. Beim Einkauf wird anstelle von Bargeld, Bankkarten oder Ausweisen nur der Irisscanner benötigt. Dieser verifiziert betrugssicher die Übereinstimmung mit den hinterlegten Irisdaten, prüft den Kontostand und löst eine Abbuchung des Zahlbetrags aus. Dabei wird zwecks Vermeidung von Bankgebühren die sog. Blockchain-Technologie verwendet, die von den Kryptowährungen her bekannt ist (vgl. hierzu Exkurs 5.2). Ähnliche Pilotprojekte laufen auch im Gesundheitsbereich.

Eine andere Möglichkeit der Datengewinnung beinhaltet den Einsatz von **Experimenten**. Dieser als **Versuchsplanung** (engl.: *design of experiments*, kurz *DoE*) bezeichnete Ansatz wurde zuerst in den *Agrar-* und

Anwendungsfelder
für Experimente

Naturwissenschaften und später in der *Technik* angewendet, ist aber auch in der *Medizin* und der *Psychologie* weitverbreitet. Bei einem Experiment geht es um die empirische Überprüfung von Hypothesen über kausale Zusammenhänge zwischen Merkmalen. Die Überprüfung erfolgt anhand einer geplanten Untersuchung, bei der die Ausprägungen eines Merkmals oder mehrerer Merkmale (**unabhängige Variablen, Einflussfaktoren**) unter Laborbedingungen systematisch variiert und der Effekt auf ein anderes Merkmal (**abhängige Variable, Zielgröße**) studiert wird. Dabei will man durch ein geeignetes Untersuchungsdesign den Einfluss weiterer Variablen möglichst ausschalten (Kontrolle von **Störvariablen**). Die Untersuchungsanordnung wird durch einen **Versuchsplan** festgelegt.

In den Wirtschafts- und Sozialwissenschaften sind Experimente kaum verbreitet, weil sich Forschung hier typischerweise auf Feldbeobachtung bezieht und selten auf Laborsituationen übertragen lässt. Stellt man hier einen empirischen Zusammenhang zwischen zwei Merkmalen fest, lässt sich daraus – anders als beim Experiment – nicht zwingend auf einen Kausalzusammenhang schließen.

Fachspezifische
Unterschiede beim
Design von
Experimenten

Zwischen Experimenten in den einzelnen Anwendungsbereichen, etwa in der Technik, Psychologie oder Medizin, gibt es Unterschiede, die durch die Natur der zu untersuchenden Merkmale bedingt sind. In der *Technik* geht es darum, Merkmale unbelebter Objekte zu untersuchen. Die planmäßige Veränderung von Formparametern eines Kraftfahrzeugs und die Untersuchung des Effekts auf den Luftwiderstand im Windkanal ist ein Beispiel für eine industrielle Anwendung von Versuchsplänen. Die Messung der Merkmalsausprägungen anhand moderner Messtechniken ist hier kein Problem und auch Messwiederholungen lassen sich leicht realisieren.



Abb. 3.4: Experiment zur Untersuchung der Reißfestigkeit von Elastomeren. Quelle: Fa. Bechem, Wetter / Ruhr

Ein weiteres, in Abbildung 3.4 wiedergegebenes Beispiel bezieht sich auf die Untersuchung des Einflusses verschiedener Schmierstoffe auf die Reißfähigkeit von verformbaren Kunststoffen (Elastomere), die häufig in technische Bauteile integriert sind. Bei dem Experiment werden genormte Elastomer-Prüfkörper mit einem zu bewertenden Schmierstoff bei einer festgelegten Temperatur bis zum Zerreißen gespannt, wobei die hierzu benötigte Kraft und Wegstrecke dokumentiert werden.

Experimente in der *Psychologie* beziehen sich hingegen auf individuelle Merkmale von Personen, etwa auf die Ausprägungen der latenten Variablen „Leistungsmotivation“, „Introvertiertheit“ oder „Lebenszufriedenheit“. Hier ist oft schon die Operationalisierung der Variablen schwierig. Ähnliches gilt für die Ausschaltung von Störeinflüssen oder die Wiederholung von Messungen. Typisch für Experimente in der Psychologie und auch in der *Medizin* ist die Ergänzung der Gruppe von Versuchspersonen um eine **Kontrollgruppe**. Nur in der **Versuchsgruppe** werden dann Einflussfaktoren variiert. Bei echten experimentellen Designs erfolgt die Zuordnung zu den beiden Gruppen durch Zufallsauswahl. Nicht immer ist eine zufallsgesteuerte Zuordnung von Personen zu einer Kontroll- und einer Versuchsgruppe realisierbar oder ethisch vertretbar. Man denke an eine Untersuchung von Effekten neuer Behandlungsmethoden in der *Medizin*, die aus ethischen Gründen so organisiert wird, dass sich die beteiligten Patienten für eine von zwei alternativen Behandlungsmethoden frei entscheiden können. Man spricht bei einem solchen Erhebungsdesign mit nicht-randomisierter Zuordnung von einem **Quasi-Experiment**.

Die vorgestellten Grundtypen „Befragung“, „Beobachtung“ und „Experiment“ sind bei SEDLMEIER / RENKEWITZ (2018, Kapitel 4 - 5) ausführlich beschrieben, die beiden erstgenannten Typen auch bei SCHNELL / HILL / ESSER (2018, Abschnitte 7.1 - 7.2).



Beispiel 3.2: Beobachtungen in verschiedenen Anwendungsfeldern

Die Ergebnisse des *Mikrozensus* sind eine für Planungen in *Politik und Wirtschaft* zentrale Informationsquelle, die sich aus *mündlichen Befragungen* speist. Es werden hier alljährlich 1% der Haushalte in Deutschland (ca. 370 000 Haushalte mit etwa 830 000 Personen) auf der Basis von Zufallsstichproben ausgewählt. Erfasst werden u. a. neben Geschlecht, Alter und Familienstand vor allem Daten über die Wohnung, Art und Umfang der Erwerbstätigkeit sowie das Nettoeinkommen. Dabei gehen Interviewer im Auftrag der Statistischen Landesämter in die Haushalte und geben die Befragungsergebnisse sofort in mitgebrachte Notebooks ein (Datenerhebung via CAPI). Die Interviewsteuerung einschließlich der Prüfung der Antwortkonsistenz wird von der auf dem Notebook vorinstallierten Software geleistet.

Beobachtung in der *Arbeits- und Organisationspsychologie* kann sich auf die Erfassung und Bewertung von menschlichem Verhalten in einem Vorstellungsgespräch beziehen (offene Beobachtung). Hier lassen sich mehrere für die künftige Tätigkeit relevante Merkmale anhand einer Ratingskala bewerten. Die Ergebnisse gehen dann in Entscheidungen zur Personalauswahl ein. Ein Beispiel für ein Experiment in der *Lernpsychologie* ist die Untersuchung des Lernerfolgs in der Statistikgrundausbildung mit und ohne Einsatz neuer Medien, etwa bei Vorlesungen mit und ohne Einbezug multimedialer Elemente und virtueller Lernumgebungen. Der Lernerfolg lässt sich über die Punktzahl bei der Abschlussklausur abbilden. Man bildet zwei Gruppen, wobei nur eine Gruppe die neuen Medien nutzt. Es wäre nicht sachadäquat, wenn die Beteiligten sich selbst eine Gruppe auswählen dürften, weil man dann mit unerwünschten Verzerrungen und Störeinflüssen rechnen müsste.

Klassifikation von
Erhebungen



Interaktives Objekt
„PKW-Neuzulassun-
gen, nach Marken“

Bei Beobachtungsstudien kann man zwischen **Querschnittsstudien** und **Längsschnittsstudien** unterscheiden. Wenn an verschiedenen Merkmalsträgern zu einem festen Zeitpunkt die Ausprägungen eines Merkmals erfasst werden, resultiert eine **Querschnittsreihe**. Verfolgt man hingegen ein Merkmal an einer statistischen Einheit im Zeitverlauf, erhält man eine **Zeitreihe**. Als Beispiel einer vielbeachteten Zeitreihe seien die Werte des *Deutschen Aktienindexes (DAX)* an einem Börsentag genannt. Auch die Anzahl der in Deutschland pro Quartal neu zugelassenen PKWs einer bestimmten Marke im Zeitraum 2006 - 2018 konstituiert eine Zeitreihe.



Abb. 3.5: Quartalsdaten zu PKW-Neuzulassungen in Deutschland für drei Marken (obere Kurve: Skoda, mittlere Kurve: Citroen, untere Kurve: Seat). Quelle: Kraftfahrt-Bundesamt, Flensburg

Ein **Panel** kombiniert Querschnitts- und Zeitreihendaten. Hier werden für dieselben Objekte wiederholt Merkmalsausprägungen ermittelt. Bei Panel-Untersuchungen, die sich auf Personen beziehen und sich über einen

längerem Zeitraum erstrecken, ist es kaum zu vermeiden, dass Teilnehmer ausscheiden, etwa durch Krankheit oder Umzug. Man spricht in diesem Zusammenhang von **Panelmortalität**. Diese kann mit unerwünschten Verzerrungen einhergehen.

Eine weitere Klassifikation für Erhebungen bezieht sich auf den Umfang der erhobenen Daten. Bei einer **Vollerhebung** bezieht man *alle Elemente* einer Grundgesamtheit in die Erhebung ein, während bei einer **Teilerhebung** oder **Stichprobenerhebung** nur Daten für eine *Teilmenge* der für die jeweilige Fragestellung relevanten Grundgesamtheit herangezogen werden. Die *Volkszählungen* des Jahres 1987 in der alten Bundesrepublik Deutschland und 1981 in der damaligen DDR waren Vollerhebungen, während der alljährlich durchgeführte *Mikrozensus* eine Stichprobenerhebung darstellt. Die letzte Volkszählung wurde für alle Länder der EU im Jahr 2011 durchgeführt (*Zensus 2011*), wobei man in Deutschland erstmals aus Kostengründen und wegen einer höheren Akzeptanz bei der Bevölkerung wesentlich auf Melderegister und Register der Bundesagentur für Arbeit zurückgriff (*registergestützter Zensus*). In Deutschland beruhten die amtlichen Bevölkerungszahlen bis 2011 noch auf Fortschreibungen der Volkszählung von 1987 bzw. 1981 anhand der Mikrozensusdaten. Mit dem Zensus von 2011 stand fest, dass die Einwohnerzahl Deutschlands Ende 2011 bei 80,3 Millionen lag und damit 1,5 Millionen niedriger als bisher angenommen. Zuverlässige Bevölkerungsdaten sind aber für viele Bereiche des öffentlichen Lebens unabdingbar, etwa für Planungen auf kommunaler Ebene, für den Länderfinanzausgleich, für den Zuschnitt von Wahlkreisen sowie auch für die Bemessung der Beiträge Deutschlands zum EU-Haushalt.

Stichprobenerhebungen sind vor allem bei sehr großen Grundgesamtheiten geboten oder oft auch der einzige gangbare Weg, weil Vollerhebungen teuer, aufwändig und nicht immer praktikabel sind. Dies gilt für die Gewinnung von sozioökonomischen Daten für große Regionen, etwa Daten zu Arbeitskosten in Deutschland. Stichprobenbasierte Erhebungen liefern auch u. U. zuverlässigere Ergebnisse, weil hier für die Datengewinnung für jeden Merkmalsträger mehr Zeit investiert werden kann. In der industriellen Qualitätssicherung ist die Merkmalserfassung manchmal – z. B. bei der Ermittlung der Lebensdauer von Leuchtmitteln – mit der Zerstörung des Merkmalsträgers verbunden. Hier gibt es zur Stichprobenprüfung keine Alternative. Bei der Prüfung sicherheitsrelevanter Produkte, etwa bei Airbags oder Reißleinen von Fallschirmen, sind hingegen Vollerhebungen geboten, weil Restrisiken nicht vertretbar sind.

Vorteile und Grenzen von Stichproben-erhebungen



Aufgabe 3.1

Beispiel 3.3: SOEP und ALLBUS

Das *Sozioökonomische Panel (SOEP)* ist eine seit 1984 durchgeführte stichprobensbasierte Befragung von über 15 000 Haushalten (gleichbleibende Haushalte), die auf die Identifikation politischer und gesellschaftlicher Veränderungen in Deutschland abzielt. Die Befragung bezieht sich auf alle erwachsenen Haushaltsglieder und erfasst u. a. Persönlichkeitsmerkmale, Lebensbedingungen, Erwerbssituation, berufliche Mobilität, Wertvorstellungen, Gesundheit und Lebenszufriedenheit. Die Befragung wird in Form persönlicher Interviews von einem Umfrageinstitut durchgeführt. Die Ergebnisse werden dann vom Deutschen Institut für Wirtschaft (DIW) in Form anonymisierter Mikrodaten an die interessierte Fachöffentlichkeit weitergegeben. Anders als beim Mikrozensus ist die Teilnahme am Sozioökonomischen Panel freiwillig.

Die *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)* ist eine seit 1980 im Zweijahresrhythmus durchgeführte Mehrthemenbefragung einer Stichprobe von ca. 3 500 Personen. Die Erhebung dient der Dauerbeobachtung gesellschaftlichen Wandels. Die Fragen beziehen sich u. a. auf Einstellungen, Erwerbstätigkeit, Umwelt und Politik. Anders als beim *SOEP* wird bei jeder Erhebung eine neue Stichprobe gezogen (Querschnittsdesign). Die Befragungen werden von wechselnden Marktforschungsinstituten im Auftrag der *Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen* durchgeführt, die seit 2008 den Namen *GESIS – Leibniz-Institut für Sozialwissenschaften* trägt. Auch hier werden die Ergebnisse der Fachöffentlichkeit zugänglich gemacht.



**Kritisch
nachgefragt**



Video des Welt-
biodiversitätsrates zur
abnehmenden
Artenvielfalt

Die Wirkung von Datenvisualisierungen kann stark von der Wahl des veranschaulichten Ausschnitts abhängen. Bei Zeitreihengraphen wird der Eindruck z. B. vom betrachteten Zeitraum beeinflusst. So veröffentlichte der *Deutsche Imkerbund* vor dem Volksbegehren „Artenvielfalt in Bayern – Rettet die Bienen“ vom Februar 2019 eine auf eigenen Daten basierende Abbildung, die für die Periode 1991 - 2018 insgesamt eine Abwärtsentwicklung der Anzahl der Bienenpopulationen in Deutschland zeigte und damit das Volksbegehren unterstützte.

Der *Bayerische Bauernverband* hielt dagegen und präsentierte auf seiner Internetseite einen auf den Zeitraum 2008 - 2018 beschränkten Ausschnitt dieser Grafik (in Abbildung 3.6 ockerfarben hinterlegt), bei dem sich der Abwärtstrend nicht widerspiegelte. Die *Süddeutsche Zeitung* visualisierte am 11. Februar 2019 unter Rückgriff auf Daten der FAO (Food and Agriculture Organization; UN-Organisation für Ernährung und Landwirtschaft) die Entwicklung des Bestands an Bienenpopulationen in Deutschland ab 1961.

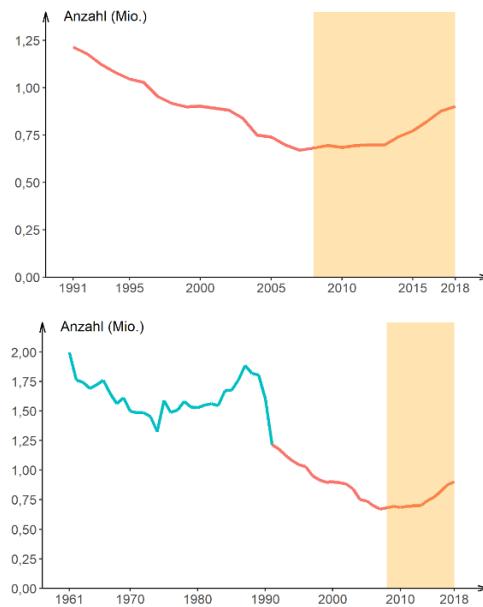


Abb. 3.6: Anzahl der Bienenvölker in Deutschland (oben: Periode 1991 - 2018; unten: Periode 1961 - 2018, Datenquelle: FAO bis 1990, türkis, Deutscher Imkerbund ab 1991, rot). Jeweils betont: vom Bayerischen Bauernverband verwendeteter Ausschnitt

Es gibt natürlich keine Ausschnittswahl, die als die einzige Richtige gelten kann. Man muss sich nur vergegenwärtigen, dass die Festlegung des Ausschnitts auch manipulativen Zwecken dienen kann.

3.2 Stichprobenauswahl

Bei Teilerhebungen ist die Verfahrensweise bei der Auswahl von Stichprobenelementen festzulegen sowie der Umfang der Stichprobe. Ziel ist es, aus einer Teilmenge einer Grundgesamtheit Aussagen abzuleiten, die sich auf die Grundgesamtheit übertragen lassen. Der Stichprobenentnahme vorgelagert ist eine eindeutige Festlegung der Grundgesamtheit. Wenn es etwa darum geht, aus einer Stichprobe von Bürgern einer Großstadt Aussagen für die gesamte Stadt zu gewinnen, muss u. a. durch räumliche Abgrenzung und inhaltliche Vorgaben (z. B. Einbezug nur der an einem Stichtag in der Stadt wohnhaften Personen) klargestellt sein, wer zur Grundgesamtheit gehört und wer nicht. In der Praxis kann es passieren, dass die Population, aus der eine Stichprobe gezogen wird, die sog. **Auswahlpopulation** oder **Auswahlgesamtheit**, Elemente enthält, die nicht zu der im Untersuchungsdesign definierten Grundgesamtheit gehören oder auch, dass einige Elemente der definierten Grundgesamtheit



bei der Stichprobenziehung gar nicht berücksichtigt werden. Im letztgenannten Fall spricht man von **Undercoverage**, im erstgenannten von **Overcoverage**. Bei der Erhebung von Bevölkerungsdaten für eine Großstadt könnten etwa Personen in der Stadt wohnen, ohne amtlich gemeldet zu sein oder aber gemeldet sein, obwohl schon längst verzogen.



Video „Over- und Undercoverage“

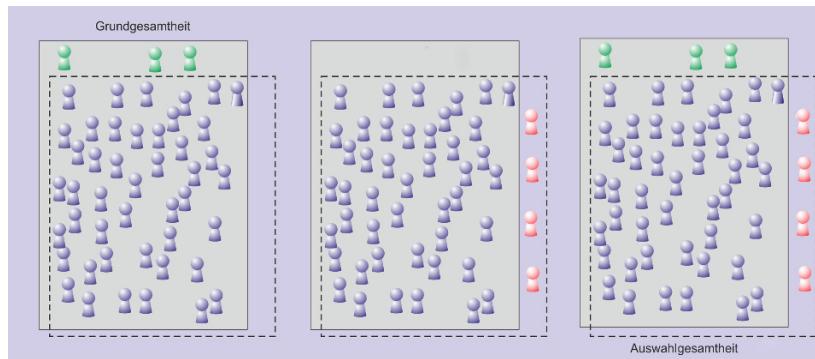


Abb. 3.7: Undercoverage (linke Teilgrafik), Overcoverage (Mitte), gleichzeitiges Auftreten von Under- und Overcoverage (rechts)



Video „Inferenzschluss“

Um mit der Stichprobe ein repräsentatives Abbild der Grundgesamtheit zu bekommen, zieht man eine **Zufallsstichprobe**. Bei einer Zufallsstichprobe hat jedes Element der Grundgesamtheit eine von Null verschiedene Wahrscheinlichkeit in die Stichprobe zu gelangen. Nur bei Realisierung einer Zufallsauswahl kann von einer Stichprobe mit einer kontrollierten kleinen Irrtumswahrscheinlichkeit auf die zugrunde liegende Grundgesamtheit zurückgeschlossen werden. Dieser auch als **Inferenzschluss** bezeichnete Rückschluss von Eigenschaften einer Stichprobe auf Eigenschaften einer Grundgesamtheit anhand von Schätz- und Testverfahren ist Gegenstand der **schließenden Statistik**. Ein Inferenzschluss ist stets mit Unsicherheit verknüpft, die sich daraus ergibt, dass nur die Teileinformation der Merkmalsträger der Stichprobe und nicht die volle Information aller Merkmalsträger der Grundgesamtheit zur Verfügung steht. Man spricht in diesem Zusammenhang von einem **Stichprobenfehler**.

Beispiel 3.4: Bestimmung der Einschaltquoten von Fernsehsendern

Die *GfK* in Nürnberg erfasst im Auftrag der *Arbeitsgemeinschaft Fernsehforschung (AGF)* die Nutzung von Bewegtbildinhalten auf stationären Fernsehgeräten und mobilen Endgeräten. Als Datenbasis dient eine für die ca. 37 Millionen Fernsehhaushalte in Deutschland repräsentative Stichprobe von ca. 5 000 Haushalten (AGF-Fernsehpanel).

Für die Gewinnung von Daten zur stationären TV-Nutzung werden die Audiosignale der Sender aufgezeichnet und gespeichert. Bei den Panelhaushalten werden solche „akustischen Fingerabdrücke“ mit einem Messgerät gewonnen, das die GfK den Haushalten zur Verfügung stellt. Abbildung 3.8 zeigt ein derartiges Messgerät im Einsatz.



Abb. 3.8: Gewinnung von Daten zur Nutzungsdauer von TV-Sendern.

Quelle: GfK Nürnberg



Einschaltquoten
vom Vortag

Durch einen Abgleich der in den Haushalten aufgezeichneten digitalen Signaturen mit denen in der Zentrale erfolgt eine Senderzuordnung und eine Erfassung der Nutzungsdauer. Um auch die Nutzung von Streaming-Inhalten, die über das Internet auf mobilen Endgeräten abgerufen werden, zu quantifizieren, werden die Panelhaushalte mit Netzwerk-Routern ausgestattet. Diese ermöglichen die Identifikation von Nutzungsart und Nutzungsdauer.

Für jede Sendung lässt sich so am Ende eines Fernsehtages eine Einschaltquote ermitteln, die auf die Grundgesamtheit aller Fernsehhaushalte bezogen wird. Die Einschaltquoten werden jeden Morgen veröffentlicht und determinieren die Preise für Fernsehwerbung. Zur Sicherung der Repräsentativität des Fernsehpanels werden Gewichtungen vorgenommen. Die Zusammensetzung des Panels und die verwendete Gewichtung werden regelmäßig überprüft.

Die Aussagekraft der Einschaltquoten wird immer wieder diskutiert, zumal das Einschalten einer Sendung noch wenig über die Konzentration beim Zuschauen aussagt. Strittig ist auch, ob sich die öffentlich-rechtlichen Sendeanstalten bei der Programmgestaltung zu sehr von Einschaltquoten und weniger von der inhaltlichen Qualität ihrer Sendungen leiten lassen.

Bei einer **einfachen Zufallsstichprobe** des Umfangs n ist die Stichprobenauswahl nicht nur zufällig, sondern auch so geplant, dass jede Teilmenge der Grundgesamtheit mit n Elementen dieselbe Auswahlwahrscheinlichkeit besitzt. Gedanklich kann man sich die Verfahrensweise anhand eines hypothetischen Gefäßes mit Kugeln oder Losen verdeutlichen (**Urnenmodell**), wobei aus dem Gefäß entweder in einem Zug oder nacheinander n Elemente gezogen werden. Die Ziehung der Lottozahlen ist z. B. so organisiert.

Zweistufige Verfahren



Aufgabe 3.2

Manchmal verfügt man auch über Vorinformation, die bei der Auswahl der Stichprobenelemente herangezogen werden kann und i. d. R. zu verlässlicheren Inferenzschlüssen führt. Dies gilt für die **geschichtete Zufallsauswahl**, ein in der Praxis verbreitetes Verfahren der Stichprobenziehung. Man zerlegt hier die Grundgesamtheit in sich nicht überlappende (= disjunkte) Teilgesamtheiten, sog. **Schichten**. Die Schichten sollen bezüglich des zu untersuchenden Merkmals in sich möglichst homogen und untereinander möglichst heterogen sein. Aus jeder Schicht wird eine Zufallsstichprobe gezogen. Die Vorinformation besteht aus der Kenntnis des auch als **Schichtungsvariable** bezeichneten Merkmals, nach dem die Grundgesamtheit in Schichten zerlegt wird. Bei einer Einkommenserhebung bei Hochschulabsolventen könnte nach Berufsgruppen geschichtet werden. Beim Sozioökonomischen Panel werden z. B. Haushalte von Deutschen und Ausländern in getrennten Schichten untersucht.

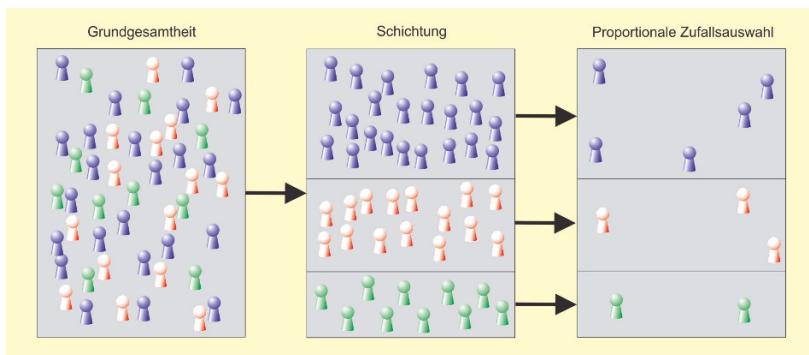


Abb. 3.9: Schichtung mit proportionaler Stichprobenauswahl

Eine geschichtete Stichprobenauswahl ist ein *zweistufiges Auswahlverfahren*, bei der eine Grundgesamtheit mit N Elementen zunächst anhand eines Hilfsmerkmals – der Schichtungsvariablen – in L disjunkte Teilgesamtheiten des Umfangs N_1, N_2, \dots, N_L zerlegt wird ($N_1 + N_2 + \dots + N_L = N$), aus denen im zweiten Schritt Zufallsstichproben des Umfangs n_1, n_2, \dots, n_L gezogen werden ($n_1 + n_2 + \dots + n_L = n$). Je nachdem, ob der Anteil $\frac{n_i}{N_i}$ ($i = 1, 2, \dots, L$) der einer Schicht entnommenen Stichprobenelemente fest ist oder nicht, liegt eine *proportional geschichtete Stichprobe* resp. eine *disproportional geschichtete Stichprobe* vor. Abbildung 3.9 zeigt eine

Grundgesamtheit von $N = 50$ Elementen, bei der zunächst eine Zerlegung in drei Schichten mit den Umfängen $N_1 = 25$, $N_2 = 15$, $N_3 = 10$ und dann in jeder Schicht eine zum Schichtumfang proportionale Zufallsstichprobe gezogen wird. Bei dem Illustrationsbeispiel beträgt der Auswahlwahrsatz 20% der Elemente einer Schicht.

Bei einer disproportional geschichteten Stichprobe ist die Auswahlwahrscheinlichkeit der Stichprobenelemente innerhalb einer Schicht konstant, nicht aber von Schicht zu Schicht. Man muss hier die Stichprobenelemente beim Rückschluss auf die Grundgesamtheit gewichten – die Gewichte sind dabei zu den Auswahlwahrscheinlichkeiten reziprok. Disproportionale Schichtung wird z. B. angewendet, wenn Schichten dünn besetzt sind. Bei geschichteten Zufallsstichproben wird eine Grundgesamtheit anhand eines Hilfsmerkmals (Schichtungsvariable) in disjunkte Teilmengen zerlegt. Manchmal zerfällt eine Grundgesamtheit auf „natürliche“ Weise in disjunkte Teilgesamtheiten, die hier **Klumpen** genannt werden. Bei einer Grundgesamtheit von Schülern könnten die Klumpen durch Klassenverbände und bei Tieren durch Herden gegeben sein. In solchen Fällen zieht man manchmal ein anderes zweistufiges Auswahlverfahren heran, die sog. **Klumpenstichprobe**. Hier wird im ersten Schritt eine Zufallsstichprobe aus der Menge aller Klumpen gezogen. Im zweiten Schritt werden dann alle Elemente der ausgewählten Klumpen untersucht.

In der Praxis, etwa in der Markt- und Meinungsforschung, werden Stichproben nicht immer zufällig, sondern auf der Basis einer Systematik ausgewählt. Ein Beispiel für ein **systematisches Stichprobenauswahlverfahren** ist die **Quotenauswahl**. Bei dieser versucht man, eine Stichprobe durch Vorgabe von Quoten bezüglich eines meist sozioökonomischen Merkmals, z. B. Geschlecht oder Alter, so zu erzeugen, dass die Stichprobe hinsichtlich dieses Merkmals – damit allerdings nicht zwingend auch hinsichtlich des eigentlich interessierenden Untersuchungsmerkmals – eine Art verkleinertes Abbild der Grundgesamtheit darstellt.

Bei BAMBERG / BAUR / KRAPP (2017, Kapitel 18) findet man eine knapp gehaltene Einführung in die Datengewinnung anhand von Stichproben. Eine umfassendere Darstellung von Stichprobenverfahren, auch ein- und mehrstufiger Zufallsauswahlverfahren, liefert eine Monografie von KAUERMANN / KÜCHENHOFF (2011).



Video
„Klumpenstichprobe“

Systematische
Auswahlprozeduren



3.3 Fehlschlüsse aus Daten in der Praxis

Bei der Arbeit mit Daten – gleich ob diese selbst erhoben sind oder aus vorhandenen Quellen stammen – ist die Prüfung der Datenqualität ein wichtiger Schritt. In der Praxis wird dieser oft vernachlässigt und das Augenmerk nur auf die verwendeten statistischen Methoden gelegt.

Mangelhafte Datenqualität kann aber zu falschen Schlussfolgerungen führen. Im Folgenden sind einige Punkte aufgeführt, die bei der Bewertung von Datenqualität und allgemein bei der Interpretation von Ergebnissen statistischer Analysen zu beachten sind.

Sicherung der Repräsentativität von Stichproben

Wenn aus Stichproben Aussagen über eine umfassendere Grundgesamtheit abgeleitet werden (Inferenzschluss), ist zu gewährleisten, dass die verwendete Stichprobe für die Gesamtpopulation repräsentativ ist. Stützt man einen Inferenzschluss auf eine nicht-repräsentative Stichprobe, kommt zu dem unvermeidlichen Stichprobenfehler noch eine systematische Verzerrung hinzu, der sog. **Auswahlbias**. Der Inferenzschluss kann dann zu gravierenden Fehlschlüssen führen. Oft ist die Vermeidung solcher Fehlschlüsse die größte Herausforderung, da man die wesentlichen Charakteristika der Grundgesamtheit kennen und darauf basierend hochrechnen muss, wie eine repräsentative Stichprobe auszusehen hat. Wichtige Kennzahlen, auf die man bei der Einschätzung der Repräsentativität einer Bevölkerungsstichprobe achten sollte, sind etwa das Geschlechterverhältnis oder die Altersstruktur.



Video „Verzerrte Stichprobe“

Würde man z. B. anhand eines Verzeichnisses stationärer Telefonanschlüsse eine Stichprobe auswählen, hieße dies, von vornehmesten einen erheblichen Teil der Bevölkerung auszuschließen. Vor allem jüngere Menschen haben oft nur noch ein Mobiltelefon. Nicht zulässig sind auch stichprobenbasierte Vergleiche der Kriminalitätsraten von Deutschen und von nach Deutschland Geflüchteten, wenn die Anzahl begangener Delikte einfach auf die beiden Gesamtpopulationen hochgerechnet wird. Ein direkter Vergleich der Kriminalitätsraten ist aus statistischer Sicht hier nicht seriös, weil sich die demographischen Charakteristika beider Grundgesamtheiten deutlich unterscheiden. Während die deutsche Bevölkerung tendenziell überaltert ist und ein weitgehend ausgeglichenes Geschlechterverhältnis aufweist, sind unter den Geflüchteten größtenteils Männer und zwar junge Männer (vgl. Abbildung 4.12). Männer sind aber allgemein strafauffälliger als Frauen und junge Männer strafauffälliger als Senioren. Die Kriminalitätsrate bei Geflüchteten wird folglich bei einem direkten Vergleich mit der deutschen Bevölkerung aufgrund des Auswahlbias höher als bei der deutschen Bevölkerung sein.

Fehlschlüsse bei zu kleinen Stichproben

Fehlende Repräsentativität von Stichproben kann auch auf zu kleinen Stichproben beruhen. An medizinischen Studien nehmen z. B. häufig nur wenige Personen teil. Es ist dann schwer, aus der Stichprobe allgemeinere Schlussfolgerungen zu ziehen. Auch politische Umfragen haben mit dem Problem zu geringer Stichprobenumfänge zu kämpfen. Schließt man etwa von nur ca. 1 000 Befragten auf die gesamte Bevölkerung Deutschlands, ist ein solcher Inferenzschluss ebenfalls sehr fehleranfällig. Dies kann eine Erklärung dafür sein, dass umfragebasierte Wahlprognosen in Zeiten mit großen Wechselwähleranteilen relativ oft unzuverlässig sind.

Beispiel 3.5: EU-Umfrage zur Abschaffung der Zeitumstellung

Zu kleinen Stichproben, die keinen belastbaren Rückschluss auf eine umfassendere Grundgesamtheit erlauben, können auch mit der Durchführung der Datenerhebung Zusammenhänge. Ein Beispiel liefert eine 2018 durchgeführte Online-Umfrage der EU zur Abschaffung der zweimal jährlich erfolgenden Zeitumstellungen. Die Teilnahme an dieser Umfrage war freiwillig. Die Umfrageergebnisse sprechen auf den ersten Blick klar für eine Abschaffung der regelmäßigen Zeitumstellungen – etwa 80% der Teilnehmer waren dafür. Eine genaue Betrachtung der Ergebnisse zeigt jedoch ein anderes Bild. Durch die freiwillige Teilnahme machten tendenziell eher Befürworter einer Abschaffung der Zeitumstellung mit. Zwar basiert Demokratie allgemein auf freiwilliger Mitwirkung, doch fehlten bei der Befragung zur Sommerzeit ausreichende Vorabinformation und Teilnehmermobilisierung – anders als etwa bei Wahlen. Ferner führt die Beschränkung auf Online-Rückmeldungen zu einer geringeren Beteiligung alter Menschen. Insgesamt stimmten EU-weit 4,6 Millionen Bürger der damaligen EU-28 mit 511,4 Millionen Einwohnern ab. In den meisten Ländern lag die Beteiligungsquote deutlich unter 1%, nur in Deutschland lag sie bei 3,8. Dies ist dadurch erklärbar, dass die Zeitumstellung in den einzelnen EU-Staaten unterschiedlich stark diskutiert wird, sehr intensiv aber in Deutschland. Das Ergebnis der EU-Umfrage als repräsentatives Meinungsbild für den EU-Raum zu deuten und hiermit eine Entscheidung über die Abschaffung oder Beibehaltung der Zeitumstellung in der EU zu begründen, wäre jedenfalls aus statistischer Sicht höchst fragwürdig.

Die Herkunft der verwendeten Daten kann eine wichtige Rolle spielen, wenn man mit neuronalen Netzen oder anderen komplexen Modellen arbeitet, die eigene Entscheidungsregeln entwickeln. Oft werden diese Modelle trainiert, um das menschliche Fehlverhalten zu neutralisieren. Dabei kann übersehen werden, dass die verwendeten Datensätze zum Training der Modelle eben auf genau diesem Fehlverhalten basieren. Diese Erfahrung musste vor kurzem der Online-Versandhändler *Amazon* machen. Um menschliche Vorurteile bei der Rekrutierung neuer Mitarbeiter auszuschließen, wurde ein intelligentes System trainiert, welches auf der Basis der Lebensläufe von Bewerbern automatisch Einstellungsempfehlungen liefert. Vorurteile etwa gegen die Einstellung von Frauen oder Personen mit Migrationshintergrund sollten so vermieden werden. Doch das System wurde mit den Einstellungsdaten der letzten 10 Jahre trainiert – Jahre, in denen Menschen anstelle von Algorithmen vorentschieden, welche Personen zu einem Auswahlgespräch eingeladen werden sollten. Die Maschine lernte nun mit den Daten und erkannte die oft unterbewusst angewandte Praxis, häufiger Männer anzustellen. Ein neuronales Netz erkennt diese unterschwellige Regel und gibt dieser ein Gewicht. Das Resultat bei *Amazon* war, dass Lebensläufe, die das Wort „weiblich“ oder „Frau“ enthielten, häufiger vom System aussortiert wurden.

Schlüsse aus Daten lernender Systeme

Vertrauenswürdigkeit von Daten	Bedeutsam ist auch die Frage, wie viel Vertrauen der Aussagekraft von Daten geschenkt werden kann. Oft werden Daten als gültige Wahrheit ausgegeben, die mehr auf Erinnerungen als auf Fakten basieren. Ein Beispiel hierzu liefert eine von Anthony Mawson, einem bekannten Impfgegner von der Jackson State University in Mississippi, geleitete Pilotstudie zur Untersuchung des Einflusses von Impfstoffen auf Krankheiten (MAWSON / RAY / BUIYAN / JACOB (2017)). Befragt wurden Mütter von geimpften und ungeimpften Kindern zum Gesundheitszustand ihrer Kinder in den letzten Jahren. Weder medizinische Daten noch ärztliche Bescheinigungen wurden berücksichtigt, lediglich das Erinnerungsvermögen der Mütter zählte bei der Datenerhebung. Dass das Erinnerungsvermögen einer Mutter in Bezug auf die Gesundheit ihres Kindes tendenziell fragwürdig ist, sollte jedem klar sein. Mütter tendieren häufiger dazu, den Gesundheitszustand ihrer Kinder dramatischer zu sehen. Auf erinnerten Daten eine statistische Analyse aufzusetzen, deren Ergebnisse von Impfgegnern weltweit zitiert und gefeiert wurden, ist mehr als bedenklich.
Verwechslung von Korrelation und Kausalität	In der statistischen Praxis gibt es auch Fehlschlüsse, die nicht auf unzureichende Datenqualität, sondern auf die unzulässige Gleichsetzung von Korrelation und Kausalität zurückzuführen sind. Bei zwei metrisch skalierten Merkmalen spricht man von Korrelation , wenn bei einem linearen Anstieg der Ausprägungen eines der beiden Merkmale auch eine lineare Veränderung beim anderen Merkmal beobachtet wird (vgl. hierzu die linke Hälfte der vierteiligen Abbildung 10.2). Die in der Technik und auch anderen Bereichen eingesetzten Versuchspläne zielen darauf ab, solche Zusammenhänge systematisch durch Veränderung von Einflussgrößen zu untersuchen. Wenn etwa die Temperatur bei der in Abbildung 3.4 veranschaulichten Reißfestigkeitsprüfung systematisch variiert wird, lässt sich ein kausaler und reproduzierbarer Zusammenhang zwischen Temperatur und Reißfestigkeit ableiten. Eine ganz andere Grundsituation liegt bei den in der wirtschafts- und sozialwissenschaftlichen Forschung verbreiteten Beobachtungsstudien vor. Hier lässt ein beobachteter Zusammenhang zwischen zwei Merkmalen nicht zwingend auf einen Kausalzusammenhang schließen. Es ist durchaus möglich, dass zwei Merkmale korreliert sind, ohne dass zwischen beiden ein sachlogischer Zusammenhang besteht. Fehlt bei zwei korrelierten Merkmalen ein kausaler Zusammenhang, spricht man in der Literatur von Scheinkorrelation (engl. <i>spurious correlation</i>). Weniger missverständlich wäre der Begriff „Scheinkausalität“.
Ethische Gesichtspunkte in der Medizin	In der Medizin ist es häufig ethisch nicht vertretbar Experimente durchzuführen. Man kann z. B. bei einem Kleinkind nicht bewusst und ohne Wissen der Eltern eine Standardimpfung auslassen, nur um zu untersuchen, ob dies einen negativen Effekt auf die Gesundheit hat. Zu hoch wäre das Risiko, dass das Kind eine Krankheit bekommt, gegen die normalerweise geimpft wird. Daher werden die meisten medizinischen

Studien als Beobachtungsstudien durchgeführt. Bei diesen kann aus einer beobachteten Korrelation zwischen zwei Variablen nicht ohne Weiteres auf einen kausalen Zusammenhang geschlossen werden.

3.4 Träger amtlicher und nicht-amtlicher Statistik

Entscheidungen in Wirtschaft und Politik in nationalem wie auch in supranationalem Kontext basieren wesentlich auf statistischen Informationen. Letztere werden nicht nur für die Entscheidungsvorbereitung, sondern auch für die Kommunikation mit dem Bürger sowie für das Monitoring und die Erfolgsbewertung von Politiken benötigt und von nationalen und internationalen Trägern amtlicher Statistik bereitgestellt. Daten stammen aber nicht nur von Statistischen Ämtern, sondern ebenfalls von nicht-amtlichen Trägern, die statistische Information auch auf Anforderung liefern, etwa für Werbezwecke. Im Folgenden werden einige Träger amtlicher und nicht-amtlicher Statistik vorgestellt.

In manchen Ländern, etwa in Japan, gehört die amtliche Statistik zum Aufgabenbereich eines Ministeriums. In Deutschland ist sie hingegen weitgehend losgelöst von Ministerien und wird von eigenständigen Behörden verantwortet (Prinzip der „fachlichen Konzentration“). Dies sichert größere Unabhängigkeit von der Tagespolitik. Für Datensammlungen, die ganz Deutschland betreffen, ist das **Statistische Bundesamt** zuständig, für regionale Daten die **Statistischen Landesämter**. Daneben gibt es auch einige **kommunale Statistikämter**. Nur wenige amtliche Statistiken werden unter direkter Kontrolle von Ministerien geführt, etwa die Arbeitsmarktstatistik der *Bundesagentur für Arbeit*, bei der das *Bundesministerium für Arbeit und Soziales* Mitverantwortung trägt.

Organisation der
amtlichen Statistik in
Deutschland

Während die Träger der amtlichen Statistik eine Informationspflicht gegenüber der Öffentlichkeit haben, gilt dies nicht für die Träger der nicht-amtlichen Statistik. Zu diesen zählen Institutionen und Firmen mit sehr unterschiedlichen Zielsetzungen, etwa Wirtschaftsforschungsinstitute, Interessen- und Wirtschaftsverbände (Gewerkschaften, Arbeitgeber, Kammern) sowie private Institute für Markt- und Meinungsforschung. Die oft an Universitäten angegliederten **Wirtschaftsforschungsinstitute** widmen sich vor allem der Analyse statistischer Daten, etwa im Rahmen der Politikberatung, und weniger der Datengewinnung. Die größten Wirtschaftsforschungsinstitute in Deutschland sind das *Institut für Wirtschaftsforschung (IFO)* in München, das *Deutsche Institut für Wirtschaftsforschung (DIW)* in Berlin, das *Rheinisch-Westfälische Institut für Wirtschaftsforschung (RWI)* in Essen, das *Institut für Weltwirtschaft (IfW)* in Kiel und das *Institut für Wirtschaftsforschung Halle (IWH)*.

Träger
nicht-amtlicher
Statistik

In die Markt- und Meinungsforschung, die im Auftrag von Unternehmen oder öffentlichen Einrichtungen erfolgt, werden erhebliche Summen investiert. Der *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM)* in Bonn bezifferte das Geschäftsvolumen für Marktforschung für das Jahr 2018 in Deutschland auf ca. 2,36 Milliarden Euro. Relativ bekannte Institute sind z. B. die *Gesellschaft für Konsumforschung (GfK)* in Nürnberg, die u. a. das Fernsehverhalten in Deutschland untersucht, oder das aus dem Zusammenschluss von EMNID und Infratest hervorgegangene Institut *TNS Infratest*, das u. a. für das *Eurobarometer* verantwortlich zeichnet. Zu nennen ist auch die *Forschungsgruppe Wahlen*, die vor allem mit dem *Politbarometer* und mit Berichten zu Bundestags- und Europawahlen in der Öffentlichkeit sichtbar wird. Die *Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS)* ist ein Zusammenschluss von Instituten, die Methodenberatung und umfangreiche Datenarchive für die empirische Sozialforschung anbieten.

Internationale Träger amtlicher Statistik

Als bedeutender Träger amtlicher supranationaler Statistik ist **Eurostat** zu nennen, das in Luxemburg ansässige **Europäische Amt für Statistik**. Dieses spielt für die europäische Politik eine wichtige Rolle. Eurostat führt nicht nur Datenbestände der Ämter von EU-Mitgliedstaaten und EU-Beitrittskandidaten zusammen, sondern ist vor allem federführend bei der Harmonisierung der Datengewinnung. Letzteres geschieht durch die Entwicklung und fortlaufende Aktualisierung von Verordnungen, in denen die Datenerhebung auf allen politikrelevanten Feldern auf europäischer Ebene verbindlich geregelt wird. Erst so wird vergleichbar, was die nationalen Ämter an statistischer Information bereitstellen. Eurostat bietet unter dem Label *Statistics Explained* ein Wissensportal an, das zu den Themenfeldern der amtlichen Statistik Texte und Grafiken für den interessierten Laien bereit stellt. Die Texte enthalten Verknüpfungen zur Datenbank von Eurostat und anderen internationalen Organisationen sowie zu Publikationen der EU-Kommission.

Internationale amtliche Daten werden auch von der **OECD**, der **Organisation für wirtschaftliche Zusammenarbeit und Entwicklung** (engl.: Organisation for Economic Co-operation and Development) bereitgestellt. Die OECD wird vor allem im Zusammenhang mit den PISA-Studien häufig in den Medien genannt. Zu erwähnen ist auch die **UN Statistics Division**, das Statistikreferat der Vereinten Nationen. Dieses veröffentlicht Daten zu gesellschaftsrelevanten Bereichen für alle Länder und unterstützt u. a. die für die Vergleichbarkeit internationaler Daten erforderliche Harmonisierung von statistischen Methoden und Definitionen. Bekannt sind der *Human Development Index* (vgl. hierzu Exkurs 8.1) und die *UN Millennium Agenda 2030 für nachhaltige Entwicklung*.

4 Univariate Häufigkeitsverteilungen



Vorschau auf
das Kapitel

In diesem Kapitel geht es um die Beschreibung und grafische Darstellung von Merkmalswerten. Wenn das betrachtete Merkmal *diskret* ist, kann man die Häufigkeit für die einzelnen Merkmalsausprägungen zählen (Feststellung der absoluten Häufigkeiten) und die Zählergebnisse durch den Umfang des Datensatzes dividieren (Berechnung der relativen Häufigkeiten). Die so definierten absoluten oder relativen Häufigkeitsverteilungen lassen sich anhand von Kreis-, Ring-, Säulen-, Balkendiagrammen visualisieren. Wenn zwei Merkmale im Spiel sind, lassen sich gestapelte Säulen- und Balkendiagramme verwenden oder auch Mosaikplots. Illustriert wird dies an Datensätzen zum Politbarometer.

Bei *stetigen* Merkmalen lassen sich die Merkmalsausprägungen zu Klassen zusammenfassen. Zur Visualisierung der absoluten oder relativen Besetzungshäufigkeiten der Klassen zieht man Histogramme heran. Diese sind durch Balken repräsentiert, deren Breite den Klassenbreiten und deren Fläche den Klassenbesetzungshäufigkeiten entspricht. Als Beispiel angeführt wird die in Form eines Doppelhistogramms darstellbare Altersstruktur der deutschen Bevölkerung (je ein Histogramm für Männer und Frauen). Für die Klassenbildung bieten sich hier die Jahrgänge an.

Wenn man die absoluten oder relativen Häufigkeiten eines diskreten Merkmals, für dessen Ausprägungen eine Rangordnung erklärt ist (mindestens Ordinalskala), bis zu einem Schwellenwert aufsummiert, erhält man eine kumulierte Häufigkeitsverteilung. Die grafische Darstellung liefert eine Treppenfunktion. Dies wird anhand von Daten zu Würfeexperimenten und Roulettespielen veranschaulicht. Eine kumulierte Häufigkeitsverteilung wird im Falle relativer Häufigkeiten auch empirische Verteilungsfunktion genannt.

4.1 Absolute und relative Häufigkeiten

Bei statistischen Erhebungen werden Ausprägungen von Merkmalen erfasst und ausgewertet. Da in der Regel die Ausprägungen vieler Einzelmmerkmale erhoben werden, fällt i. Allg. eine kaum überschaubare Fülle von Datensätzen an, die es zu charakterisieren und zu visualisieren gilt. Um auch bei großen Datenmengen eine Übersicht zu gewinnen, wird die in den Daten steckende Information unter Verwendung statistischer Kenngrößen (Lage- und Streuungsparameter) und einfacher grafischer Instrumente verdichtet. Je nachdem, ob man Daten für ein Merkmal oder für mehrere Merkmale auswertet, spricht man von **univariater** oder **multivariater Datenanalyse**. Bei letzterer steht die Analyse von Zusammenhängen zwischen Merkmalen im Vordergrund. Im Folgenden geht es erst einmal nur um die univariate Datenanalyse.

Klassenbildung bei stetigen Merkmalen

Betrachtet sei eine Erhebung, bei der für ein beliebig skaliertes Merkmal X an n Merkmalsträgern oder Untersuchungseinheiten jeweils die Merkmalsausprägung festgestellt wird. Die beobachteten oder gemessenen Merkmalswerte x_1, \dots, x_n konstituieren die Urliste. Da sich die Urliste hier auf ein einziges Merkmal bezieht, liegt eine **univariate Urliste** vor. In dieser können Werte mehrfach auftreten. Dieser Fall tritt bei diskreten Merkmalen zwangsläufig auf, wenn die Länge n der Urliste die Anzahl k der möglichen Merkmalsausprägungen überschreitet. Wenn man z. B. eine Münze mehr als zweimal wirft, wird mindestens einer der beiden möglichen Ausgänge „Kopf“ und „Zahl“ des Münzwurfexperiments mehr als einmal beobachtet. Bei stetigen Merkmalen ist das wiederholte Auftreten von Merkmalswerten um so seltener, je genauer gemessen wird. Bei hoher Messgenauigkeit kann es auch bei großer Anzahl n von Beobachtungswerten passieren, dass alle Merkmalswerte unterschiedlich ausfallen, d. h. die Anzahl der realisierten Ausprägungen mit n übereinstimmt. Wenn man z. B. in einer kleineren Kommune für alle Haushalte die jährlich anfallenden Rechnungsbeträge der Stadtwerke für Wasser und Strom ohne Rundung auf volle Eurobeträge auswiese, so würden kaum zwei Beträge exakt übereinstimmen. In solchen Fällen kann man die Daten zu Gruppen oder Klassen zusammenfassen. Dies geschieht dadurch, dass man den Gesamtbereich, in dem die Merkmalsausprägungen liegen, in eine überschaubare Anzahl von Teilintervallen zerlegt und die Daten den Teilintervallen zuordnet. Man spricht dann von **gruppierten Daten** oder von **klassierten Daten**. Bei einer Urliste mit Bruttostundenverdiensten für alle Arbeitnehmer eines Landes könnte man etwa wenige Einkommensklassen unterscheiden (z. B. Stundenverdienste von „0 bis unter 5 Euro“, „5 bis unter 10 Euro“, …, „45 bis unter 50“ und die nach oben offene Klasse „50 und mehr“).

Verteilung von absoluten und relativen Häufigkeiten

Urlisten werden mit wachsender Länge n und sich wiederholenden Merkmalswerten rasch unübersichtlich. Es empfiehlt sich dann, die in den Rohdaten enthaltene Information durch Angabe von Häufigkeiten für die Merkmalsausprägungen oder, bei gruppierten Daten, für Klassenbesetzungshäufigkeiten – zusammenzufassen. Hat man ein diskretes Merkmal mit Ausprägungen a_1, \dots, a_k , so ist die im Folgenden mit

$$h_i := h(a_i) \quad i = 1, 2, \dots, k \quad (4.1)$$

bezeichnete **absolute Häufigkeit** für die Ausprägung a_i die Anzahl der Elemente der Urliste, die mit dem Wert a_i übereinstimmen.

Absolute Häufigkeiten haben den Nachteil, dass sie von der Länge n der Urliste abhängen. Es gibt z. B. wenig Sinn, die Häufigkeit des Auftretens von Depressionen in Bremen und Niedersachsen anhand der Fallzahlen in den beiden genannten Bundesländern zu vergleichen. Um auch Häufig-

keiten vergleichbar zu machen, die sich auf Datensätze unterschiedlichen Umfangs beziehen, teilt man die absoluten Häufigkeiten durch den Umfang n der Beobachtungsreihe. Es resultieren **relative Häufigkeiten**

$$f_i := f(a_i) = \frac{h(a_i)}{n} \quad i = 1, 2, \dots, k. \quad (4.2)$$

Die durch (4.2) definierten Anteilswerte kann man auch in Form von Prozentwerten ausweisen (Multiplikation mit 100).¹

Häufigkeiten lassen sich in Tabellenform ausweisen. Dabei resultieren **Häufigkeitsverteilungen** für absolute oder relative Häufigkeiten. Die absoluten Häufigkeiten addieren sich zu n und die relativen Häufigkeiten zu 1. Eine Häufigkeitsverteilung für ein Merkmal X wird auch als **empirische Verteilung** für dieses Merkmal bezeichnet.

Häufigkeitstabellen lassen sich auch grafisch darstellen. Dabei kommen, wie anhand von Beispiel 4.1 illustriert, unterschiedliche Visualisierungsoptionen in Betracht. Bei einem **Kreisdiagramm** werden die absoluten oder relativen Häufigkeiten durch Kreissektoren repräsentiert. Der Mittelpunktwinkel α_i , der die Größe des Kreissektors definiert, ist sowohl bei absoluten Häufigkeiten h_i als auch bei relativen Häufigkeiten f_i durch $f_i \cdot 360^\circ$ gegeben. Eine Variante des Kreisdiagramms ist das **Ringdiagramm**, auch **Donut-Diagramm** genannt. Hier wird der Kreis durch einen Kreisring ersetzt.

Statt einen einzigen Kreis in Segmente aufzuteilen, kann man auch für jede Häufigkeit einen eigenen Kreis vorsehen. Die Kreisflächen sind dann proportional zum jeweiligen Häufigkeitswert zu wählen. Für die resultierende Grafik findet man die Bezeichnung **Blasendiagramm** (engl.: *bubble chart*). Zur Veranschaulichung von Regionaldaten kann man Blasendiagramme mit Landkarten verknüpfen. Ab März 2020 veröffentlichte die *Berliner Morgenpost* z. B. ein täglich aktualisiertes Blasendiagramm, das den Stand der bisher registrierten Corona-Fälle in den deutschen Bundesländern widerspiegelt. Die Mittelpunkte der Kreise waren hier in der Karte so platziert, dass sie in dem betreffenden Bundesland lagen. Innerhalb der Kreise wurden durch unterschiedliche Farbgebung nach noch akuten, schon ausgeheilten sowie tödlich verlaufenen Fällen differenziert.

Als Alternative zum Kreis- oder Ringdiagramm werden Stab- und Säulen- oder Balkendiagramme verwendet. Beim **Stabdiagramm** werden die Häufigkeiten durch vertikale dünne Stäbe (Striche), beim **Säulendiagramm** und beim **Balkendiagramm** durch vertikal resp. durch horizont-



Interaktives Objekt
„Depressionshäufigkeiten
in Deutschland 2015“

Visualisierung
von Häufigkeits-
verteilungen

¹Die Bezeichnungen für Häufigkeiten sind in der Literatur uneinheitlich. Die hier verwendete Notation h_i für absolute und f_i für relative Häufigkeiten ist allerdings sehr verbreitet – vgl. z. B. die Lehrbücher von FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 2.1.1) oder STELUND (2016, Abschnitt 1.5.2).

tal angeordnete dicke Stäbe dargestellt. Wenn die Merkmalsausprägungen Kategorien mit längeren Namen sind (etwa Namen von Staaten, Bundesländern oder Parteien), empfiehlt es sich, Codes zu verwenden.



Interaktives Objekt
„Militärausgaben“

Abbildung 4.1 vergleicht Militärausgaben von 24 Ländern für das Jahr 2019 anhand eines Balkendiagramms. Die Ausgaben sind als Anteil am Bruttoinlandsprodukt (BIP) ausgewiesen. Die Ländernamen werden bei der nebenstehenden interaktiven Visualisierung auch codiert ausgewiesen.² Die in Abbildung 4.1 eingehenden Daten stammen von *SIPRI* (Stockholm International Peace Research Institute), einem unabhängigen Institut für Friedensforschung und Rüstungskontrolle. Vergleicht man die Militärausgaben der Länder anhand der Anteile am jeweiligen Bruttoinlandsprodukt oder anhand von Pro-Kopf-Ausgaben, erhält man ein ganz anderes Bild (s. Tabelle 8.1).

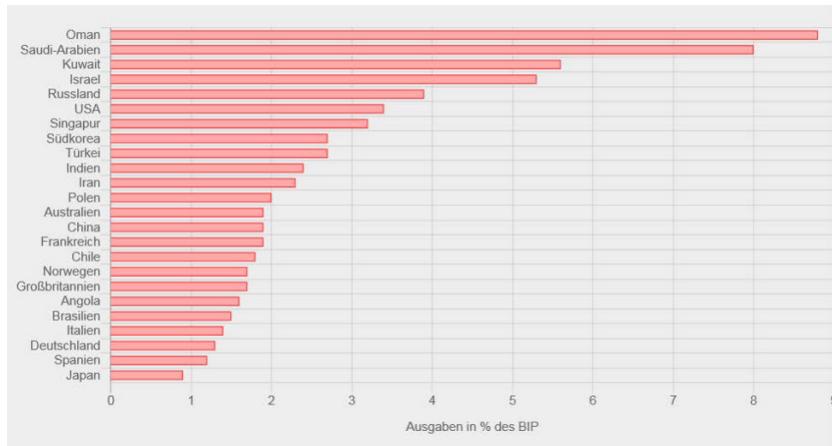


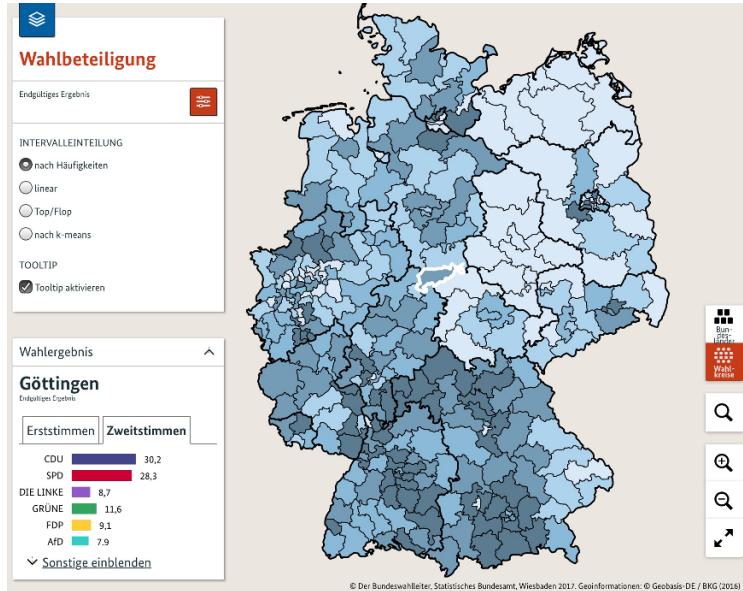
Abb. 4.1: Militärausgaben 2019 von 24 Ländern in % des BIP (Säulendiagramm); Quelle: SIPRI; Datenextraktion: April 2020

Nutzer-
freundlichkeit
der amtlichen
Statistik

Die amtliche Statistik bedient sich heute einer nutzerfreundlichen Datenkommunikation. Manchmal werden grafische Darstellungen von Häufigkeitsverteilungen dabei mit Landkarten verknüpft. Abbildung 4.2 zeigt eine interaktive Online-Präsentation von Ergebnissen der Bundestagswahl 2017. Man kann hier alternativ die Erst- oder Zweitstimmenanteile auf Ebene der Bundesländer oder der Wahlkreise in Form eines Balkendiagramms anzeigen lassen (in %). Die Blaustufen der Karte repräsentieren unterschiedliche Wahlbeteiligungen. Mit der Maus oder – bei mobilen

²Ländercodes: AO - Angola, AU - Australien, BR - Brasilien, CL - Chile, CN - China, DE - Deutschland, ES - Spanien, FR - Frankreich, UK - Großbritannien, IL - Israel, IN - Indien, IR - Iran, IT - Italien, JP - Japan, KR - Südkorea, KW - Kuwait, NO - Norwegen, OM - Oman, PL - Polen, RU - Russland, SA - Saudi-Arabien, SG - Singapur, TR - Türkei, US - USA. Die Militärausgaben umfassen auch Aufwendungen für Personal und Instandhaltung.

Endgeräten – mit dem Finger lässt sich auswählen, für welche Region man die Wahlergebnisse sehen möchte.



Bundestagswahl 2017

Abb. 4.2: Wahlbeteiligung bei der Bundestagswahl 2017 mit Balkendiagramm für die Zweitstimmenanteile der Parteien im Wahlkreis Göttingen. Quelle: Statistisches Bundesamt

Beispiel 4.1: Ergebnisse des ZDF-Politbarometers vom 8. 12. 2017

Bei der bekannten „Sonntagsfrage“ – einer im Auftrag des ZDF im Zweiten-Wochen-Turnus durchgeführten Telefonbefragung – wird die Wahlentscheidung für den fiktiven Fall erfragt, dass am nächsten Sonntag Bundestagswahlen stattfinden. Abbildung 4.3 zeigt die Ergebnisse vom 8. Dezember 2017. Die Stichprobe umfasste $n = 1451$ zufällig ausgewählte Wahlberechtigte:

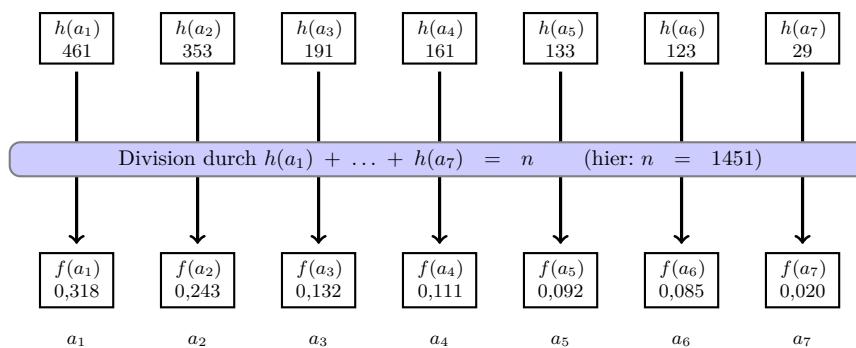


Abb. 4.3: Häufigkeiten beim ZDF-Politbarometer vom 8. Dezember 2017.
Quelle: Forschungsgruppe Wahlen

Die Ausprägungen a_1, \dots, a_7 des Merkmals „präferierte Partei“ stehen für die CDU/CSU (Union), die SPD, die Grünen, die Linke, die FDP, die AfD resp. für „Sonstige“. Angegeben sind die absoluten Häufigkeiten $h(a_i)$ und die auf drei Dezimalstellen gerundeten relativen Häufigkeiten $f(a_i)$ ($i = 1, 2, \dots, 7$).

Tabelle 4.1 gibt die Häufigkeiten für das Merkmal „Parteipräferenz“ wieder. In der letzten Tabellenspalte sind – zum Vergleich mit den relativen Häufigkeiten in der dritten Tabellenspalte – auch die Zweitstimmenanteile bei der Bundestagswahl vom 24. September 2017 ausgewiesen (in Klammern: Zweitstimmenanteile der Bundestagswahl vom 27. September 2009).

Merkmalsaus- prägung a_i	„Sonntagsfrage“ vom 8. 12. 2017		Bundes- tagswahl
	Absolute Häufigkeit $h(a_i)$	Relative Häufigkeit $f(a_i)$	Zweitstimmen- anteil 2017 (2009)
	a_1	461	0,318
	a_2	353	0,243
	a_3	191	0,132
	a_4	161	0,111
	a_5	133	0,092
	a_6	123	0,085
Sonstige	a_7	29	0,020
	Summe	$n = 1451$	1
			1 (1)

Tab. 4.1: Politbarometer vom 8. Dezember 2017 und Bundestagswahlergebnisse vom September 2017 (in Klammern: September 2009)

Abbildung 4.4 veranschaulicht die Ergebnisse der „Sonntagsfrage“ vom 8. Dezember 2017 in Form je eines Kreis-, Ring-, Säulen- und Stabdiagramms. Beim Kreis- und Ringdiagramm lassen sich Anteile ähnlicher Größe – hier z. B. die relativen Häufigkeiten $f(a_5)$ und $f(a_6)$ für FDP und AfD oder die Anteile $f(a_3)$ und $f(a_4)$ der Grünen und der Linken – nicht so gut wie beim Säulen- oder Stabdiagramm vergleichen.

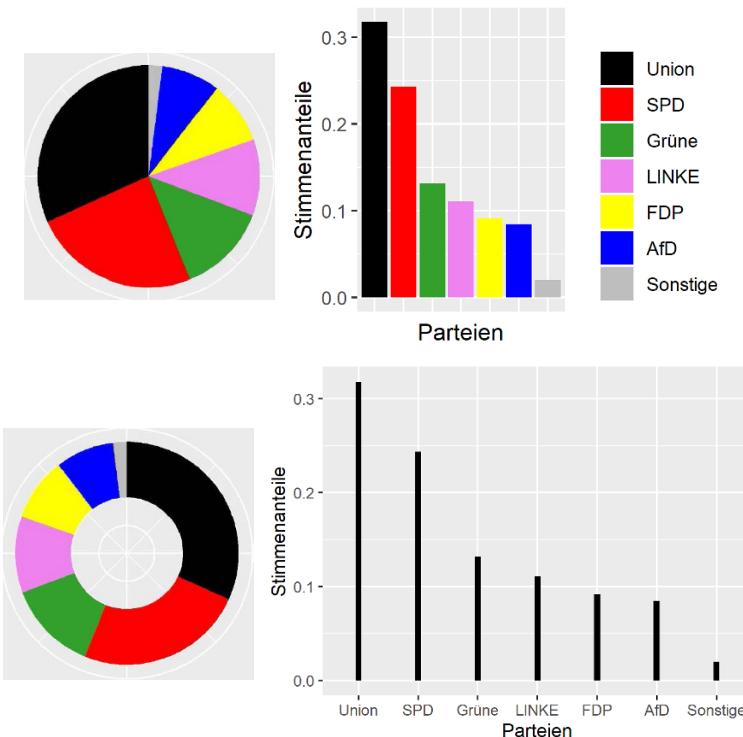


Abb. 4.4: Kreis-, Säulen-, Ring- und Stabdiagramm (Politbarometer)

Zur Visualisierung von Häufigkeitsverteilungen für zwei oder mehr qualitative Merkmale kann man **gestapelte Säulendiagramme** bzw. **gestapelte Balkendiagramme** verwenden. Hier sind die Säulen resp. Balken in zwei oder mehr Teile zerlegt. Die Komponenten können durch unterschiedliche Schraffierung oder Färbung unterschieden werden. Bei Zerlegung der Säulen oder Balken in mehr als zwei Komponenten wird die Darstellung schnell unübersichtlich. Durch Einblendung der numerischen Angaben in die Komponenten wird dieser Nachteil gemildert.

Weitere grafische Instrumente

3D-Säulendiagramme werden manchmal als Alternative zu gestapelten Säulendiagrammen verwendet. Hier sind die einzelnen Säulenabschnitte nicht übereinander, sondern hintereinander gestellt und dreidimensional ausgestaltet. Dies geht aber mit perspektivischen Verzerrungen einher (s. Abbildung 4.5). Ebenfalls nicht zu empfehlen sind **3D-Kreisdiagramme**, die z. B. in Excel angeboten werden. Bei diesen ist die dritte Dimension inhaltsleer und man hat auch hier perspektivische Verzerrungen.

Abbildung 4.4 zeigte Möglichkeiten der Veranschaulichung der Häufigkeitsverteilung für das qualitative Merkmal „präferierte Partei“. Differenziert man nach Geschlecht, betrachtet also die beiden qualitativen Merkmale „von Frauen präferierte Partei“ und „von Männern präferierte Partei“,

kann man außer gestapelten Säulen- oder Balkendiagrammen auch sog. **Mosaik-Plots** (engl.: *mosaic plot*) heranziehen, die auch **Marimekko-Diagramme** genannt werden. Ausgangspunkt ist ein Rechteck, das im Falle zweier qualitativer Merkmale vertikal oder horizontal in zwei farblich unterschiedene Streifen zerlegt wird, deren Breite zur Anzahl der Beobachtungswerte für jedes Einzelmerkmal proportional ist. Jeder Streifen wird in Teilabschnitte zerlegt, deren Länge die Häufigkeit der einzelnen Kategorien widerspiegelt.



Kacheldiagramm zum
Bundshaushalt 2017

Zur Veranschaulichung der Größenordnung quantitativer Daten sind auch **Kacheldiagramme** (engl.: *treemap*) geeignet. Bei diesen wird ein Rechteck in kleinere, ineinander verschachtelte Rechtecke aufgeteilt, deren Flächeninhalte zur Größe der darzustellenden Daten proportional sind. Während beim Mosaik-Plot die einzelnen Streifen in gleich viele Segmente zerlegt sind, genau wie auch beim gestapelten Säulendiagramm die einzelnen Säulen, kann bei einem Kacheldiagramm die Unterteilung der Äste der Baumstruktur unterschiedlich differenziert ausfallen.

Eine umfassende Sammlung von Beispielen für die Visualisierung von Daten mit den vorstehend präsentierten und weiteren Grafiktypen einschließlich deren Programmierung in *R* findet man bei RAHLF (2018).

Beispiel 4.2: Geschlecht und Parteipräferenz

Tabelle 4.2 zeigt nochmals die Ergebnisse des Politbarometers vom 8. Dezember 2017, nun mit Differenzierung nach Geschlecht. Von den 1 451 Personen, die nach ihrer Wahlpräferenz befragt wurden, waren 824 männlich und 627 weiblich.

Merkmalsaus- prägung a_i	Anzahl $h(a_i)$	Anzahl $h(a_i)$	Anteil $f(a_i)$	Anteil $f(a_i)$
	σ	φ	σ	φ
	235	226	0,162	0,156
	219	134	0,151	0,092
	86	105	0,059	0,072
	86	75	0,059	0,052
	88	45	0,061	0,031
	93	30	0,063	0,021
Sonstige	17	12	0,012	0,008
Summe	824	627	0,568	0,432

Tab. 4.2: Politbarometerergebnisse mit Differenzierung nach Geschlecht
(σ : Werte für die befragten Männer; φ : Werte für die Frauen)

Die folgende Grafik zeigt im linken Teil ein gestapeltes Säulendiagramm für die nach Geschlecht differenzierten relativen Häufigkeiten der Tabelle. Die Farben für die sieben Säulen wurden wie in Abbildung 4.4 gewählt, wobei die Wahlpräferenzen der Frauen repräsentierenden Komponenten etwas heller dargestellt sind. Der rechte Abbildungsteil veranschaulicht dieselbe Information in Form eines 3D-Säulendiagramms.

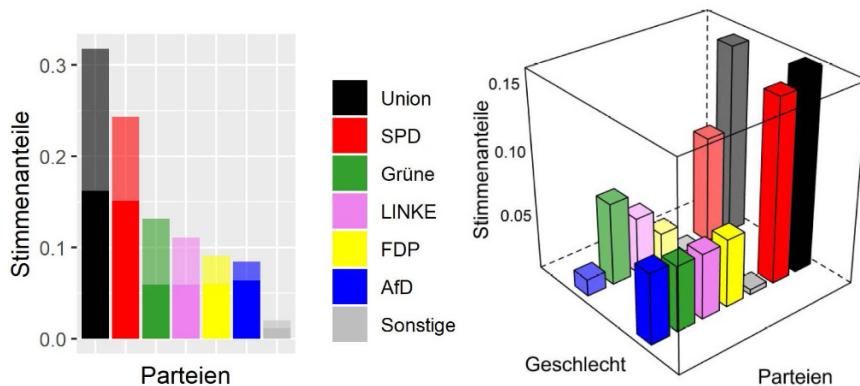


Abb. 4.5: Gestapeltes Säulendiagramm und 3D-Säulendiagramm für die geschlechtsspezifischen Parteipräferenzen

Die Häufigkeitsverteilungen aus Tabelle 4.2 lassen sich auch als Mosaik-Plot darstellen. Die Breiten der beiden Streifen, in die das Basisrechteck zerlegt ist, spiegeln das Verhältnis der Geschlechter in der Stichprobe wider. Jeder Streifen ist in sieben farbige Abschnitte gegliedert, wobei die Farben für die einzelnen Parteien aus Abbildung 4.4 übernommen wurden. Die Längen der sich entsprechenden Abschnitte in den beiden Streifen unterscheiden sich z. T. erheblich. Man sieht dies z. B. gut an den gelben resp. blauen Abschnitten, die die Präferenzen für „FDP“ und „AfD“ repräsentieren. Die Parteipräferenzen von Männern und Frauen differieren offenbar wesentlich.

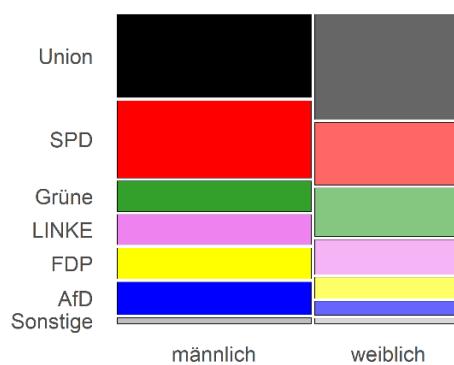


Abb. 4.6: Mosaik-Plot für geschlechtsspezifische Parteipräferenzen



**Kritisch
nachgefragt**



Corona-Dashboard

SUCCe-D

Nachdem im März 2020 in Deutschland die Corona-Pandemie (COVID-19; engl: *coronavirus disease 2019*) einsetzte, wurden wir mit absoluten und relativen Häufigkeiten für positive Tests und mit Todesfallzahlen geradezu überschüttet. Statistiken fanden wohl nie zuvor in der Öffentlichkeit so große Beachtung. Daten zur Entwicklung der Pandemie wurden täglich von öffentlichen Einrichtungen publiziert, in Deutschland u. a. durch das **Robert-Koch-Institut (RKI)**, international durch die **Weltgesundheitsorganisation (WHO)** sowie die **Johns-Hopkins-Universität**. In den Medien fand man interaktive und z. T. auch mit Landkarten verknüpfte Visualisierungen zum aktuellen Verlauf. Die sachadäquate Interpretation der vielen und z. T. widersprüchlichen Kennzahlen bereitete dem Publikum aber zunehmend Probleme. Eine sachgerechte Bewertung ist in der Tat nicht trivial und erfordert eine genauere Betrachtung.

Wir illustrieren dies anhand einer Datenreihe aus dem **Situationsbericht** des Robert-Koch-Instituts vom 20. Mai 2020. Hier wurden die von Anfang März – Mitte Mai 2020 von über 200 Laboren durchgeföhrten Tests auf COVID-19 beziffert und dabei ausgewiesen, wie hoch jeweils die Anzahl der positiven Testungen war. Die nachstehende Abbildung zeigt im oberen Teil ein gestapeltes Säulendiagramm, bei der die Säulenlängen die Gesamtzahl der wöchentlichen Tests repräsentieren. Die Tests mit negativem bzw. positivem Ergebnis sind durch unterschiedliche Farben der Säulenabschnitte unterschieden. Der untere Abbildungsteil zeigt den jeweiligen Anteil der positiven Tests.

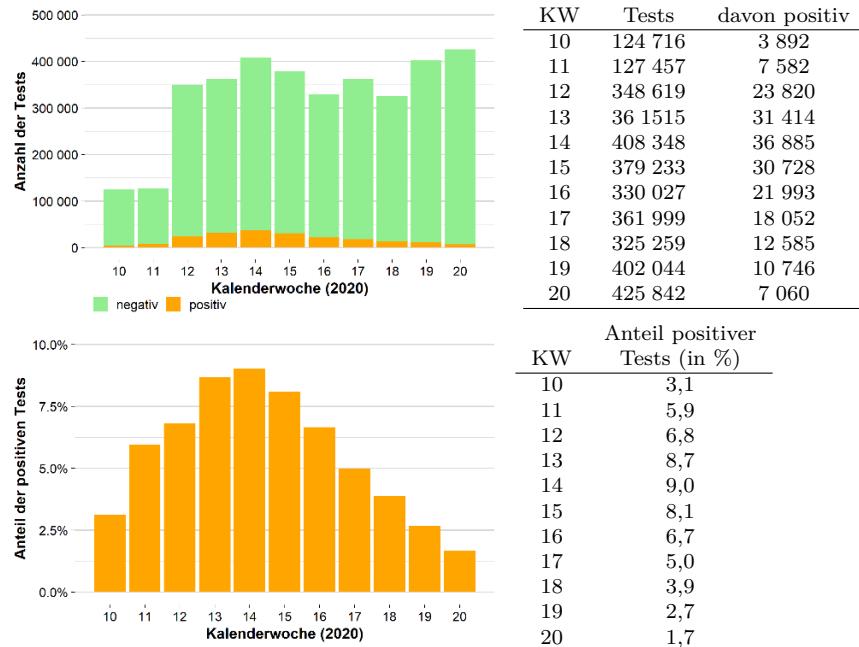


Abb. 4.7: *COVID-19-Tests in Deutschland in den Kalenderwochen 10 - 20 des Jahres 2020. Oben: Anzahl aller durchgeföhrten Tests. Unten: Anteil der Tests mit positivem Befund (in %). Quelle: Robert-Koch-Institut*

Man erkennt, dass die Anzahl der Tests bis Mitte März hochgefahren wurde – bis zum 19. Mai 2020 waren in Deutschland etwa 3,6 Millionen Tests auf COVID-19 durchgeführt – und dann auf hohem Niveau blieb. Der Anteil der Tests mit positivem Ausgang ging hingegen ab der 14. Kalenderwoche zurück. Man könnte also vermuten, dass die Durchseuchungsrate der Bevölkerung bis etwa Anfang April (Kalenderwoche 14) zunahm, um dann wieder zurückzugehen.

Es ist allerdings größte Vorsicht bei der Interpretation der Testbefunde geboten. Der Anteil positiver Tests kann nur bedingt zur Abschätzung des Anteils unerkannt infizierter Personen (Dunkelziffer) herangezogen werden. Zum einen hängt die Aussagekraft der Werte für neue Infektionsfälle von der Erhebungsmethodik ab, vor allem davon, ob die COVID-19-Tests auf der Basis repräsentativer Stichproben durchgeführt wurden. Dies ist aber hier nicht der Fall. Die Tests wurden in Deutschland in dem betrachteten Zeitraum praktisch nur an Personen mit spezifischen Symptomen und vorherigem Kontakt zu nachgewiesenermaßen Infizierten durchgeführt. Deshalb lieferten die Testergebnisse hochgradig unvollständige, nicht repräsentative Bilder der tatsächlichen Infektionslage. In anderen Ländern wurde gänzlich anders getestet, so dass auch verlässliche internationale Vergleiche unmöglich sind.

Nicht wenige Personen wurden zudem mehrfach getestet, weil die verwendeten PCR-Tests³ nur während einer zeitlich begrenzten Phase der Infektion anschlagen und schätzungsweise bis zu etwa 20 % der Getesteten fälschlich als nicht-infiziert eingestuft wurden (vgl. hierzu auch Abschnitt 11.4).

Wenn in einem Land die Fallzahlen ansteigen, weil mehr getestet wird (z. B. systematisch in Kliniken oder in Sammelunterkünften) und damit mehr Infizierte entdeckt werden, kann dennoch die Pandemie zunehmend unter Kontrolle sein. Umgekehrt können abnehmende Fallzahlen möglicherweise mit einer abnehmenden Zahl an Tests zusammenhängen. Die Fallzahlen für positiv Getestete lassen jedenfalls nur bedingt Rückschlüsse auf den Gesundheitsstatus aller Personen in der Region zu, so lange die Getesteten keine Zufallsstichprobe der betrachteten Grundgesamtheit konstituieren.

4.2 Häufigkeitsverteilungen für klassierte Daten

Bei klassierten Daten bezieht sich eine Häufigkeitsverteilung auf Klassenbesetzungshäufigkeiten. Auch hier kann man die Häufigkeiten anhand von Säulen darstellen, wobei sich die Breite der Säulen an der Breite der Klassen orientiert, d. h. die durch Rechtecke repräsentierten Besetzungshäufigkeiten schließen direkt aneinander an. Die resultierende Grafik nennt man **Histogramm**. Die Klassenbesetzungshäufigkeiten sind zu den Flächeninhalten der einzelnen Rechtecke proportional. Bei Wahl gleicher Klassenbreiten lassen sich die Klassenbesetzungshäufigkeiten direkt anhand der Länge der Säulen miteinander vergleichen.

³PCR steht abkürzend für *polymerase chain reaction* und bezeichnet ein Verfahren zur Vervielfältigung und Identifikation einer DNA-Sequenz des COVID-19-Virus.

Visualisierung von
Einkommens-
verteilungen



Aufgabe 4.2

Abbildung 4.8 zeigt Bruttojahresverdienste von Arbeitnehmern in Spanien und Portugal im Bereich „Industrie und Dienstleistungen“ für 2002. Die Daten stammen aus einer im 4-Jahres-Turnus durchgeführten Europäischen Verdienststrukturerhebung (engl: *Structure of Earnings Survey*).⁴ Die Jahreseinkommen umfassen auch Sonderzahlungen, etwa Boni, Weihnachtsgeld und Urlaubsgeld. Die Daten für die beiden Länder werden hier anhand von Histogrammen visualisiert (Klassierung der Individualdaten).⁵ Die Klassen in der Abbildung sind 5 000-Euro-Intervalle, die jeweils die rechte Intervallgrenze nicht einschließen. Die letzte der 15 Klassen, zu der im Vergleich zur vorletzten Klasse ein etwas höheres Rechteck gehört, ist nach oben offen. Würde man die Anzahl der Klassen deutlich erhöhen, entfiele die leichte Auffälligkeit der Höhe des letzten Rechtecks.

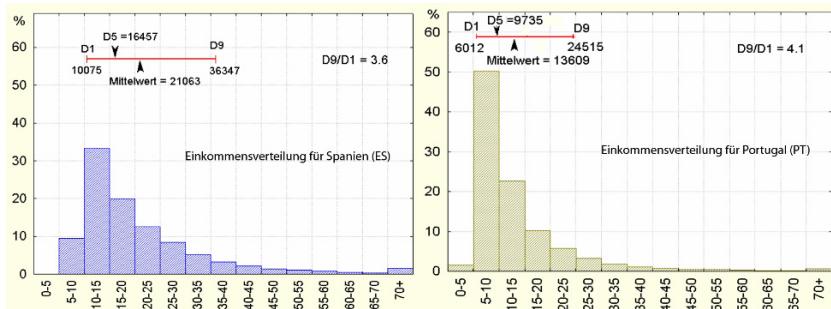


Abb. 4.8: Bruttojahresverdienste 2002 in Spanien und Portugal in Tausend Euro (Histogramme der nationalen Einkommensverteilungen)

Visualisierung von
Alterstrukturen

Auch bei Bevölkerungsdaten für größere Grundgesamtheiten bietet sich eine Klassenbildung an, z. B. nach Jahrgängen oder nach mehrere Jahre umfassenden Altersklassen. Das *Statistische Bundesamt* präsentiert z. B. eine ansprechende interaktive Visualisierung der Bevölkerungsentwicklung in Deutschland für den Zeitraum 1950 bis 2060. Gezeigt werden zwei vertikal und spiegelbildlich zueinander angeordnete Histogramme, die die Anzahl von Männern und Frauen für 100 Jahrgänge (0 bis 100 Jahre) ausweisen. Im Juli 2019 wurde die 14. Bevölkerungsvorausberechnung für Deutschland der Öffentlichkeit vorgestellt. Die Berechnungen basieren auf einem Set von drei alternativ herangezogenen Annahmen zur Geburtenhäufigkeit pro Frau, zur Lebenserwartung Neugeborener im Jahr 2060 und zur Stärke von Zuwanderungsbewegungen:

⁴Die Ergebnisse zum bisher letzten *Structure of Earnings Survey*, der sich auf das Referenzjahr 2018 bezog, sind ab Anfang 2021 verfügbar.

⁵Die Zusatzinformationen oberhalb der beiden in Abbildung 4.8 wiedergegebenen Histogramme werden in Kapitel 5 erläutert.

Geburtenhäufigkeit pro Frau 2060	Lebenserwartung Neugeborener 2060	positiver Wanderungssaldo pro Jahr
G1: Rückgang auf 1,4	L1: 82,5 für Jungen / 86,4 für Mädchen	W1: 147 000 Personen
G2: stabil bei ca. 1,55	L2: 84,4 für Jungen / 88,1 für Mädchen	W2: 221 000 Personen
G3: Anstieg auf 1,7	L3: 86,2 für Jungen / 89,6 für Mädchen	W3: 311 000 Personen

Tab. 4.3: Annahmen des Statistischen Bundesamtes bei der 14. Bevölkerungsvorausberechnung (Juli 2019)

Abbildung 4.9 zeigt die Altersstruktur der Bevölkerung von Deutschland für den Annahmensest G2-L2-W1. Die nebenstehende Animation lässt sich mit allen 27 G-L-W-Kombinationen abspielen.



Animation
„Altersstruktur
für Deutschland“

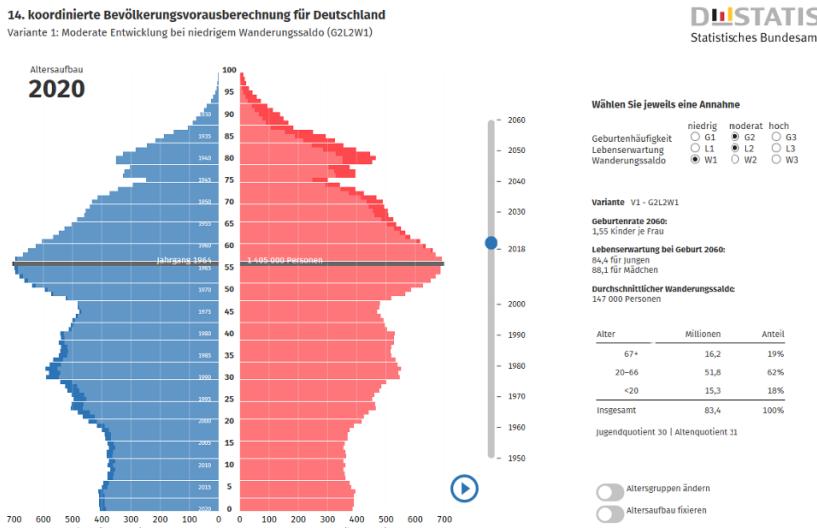


Abb. 4.9: Bevölkerungsstruktur von Deutschland im Jahr 2020 (Quelle: Statistisches Bundesamt; Basis: Annahmen G2-L2-W1)

Eine Darstellung des in Abbildung 4.9 wiedergegebenen Typs wird **Bevölkerungspyramide** genannt. Für Deutschland ähnelt diese Darstellung der Bevölkerungsstruktur allerdings schon längst eher einem Pilz. In der Abbildung ist der geburtenstärkste Jahrgang von 1964 betont, dem in 2020 etwa 1,4 Millionen Menschen zuzurechnen sind. Die Gesamtbevölkerung Deutschlands umfasst 2020 etwa 83,4 Millionen Personen, was zu diesem Zeitpunkt kaum mehr als 1% der Weltbevölkerung entspricht. Die dunkelviolette Fläche am oberen Rand des rechten Teilhistogramms weist den Frauenüberschuss bei der älteren Bevölkerung aus, während die dunkelblaue Fläche am unteren Rand des linken Teilhistogramms den



Interaktives Objekt
„Lebenserwartung
Neugeborener in
der EU“

leichten Männerüberschuss bei jüngeren Jahrgängen betont. Man erkennt noch die kriegsbedingten Einschnitte bei den Jahrgängen 1940 – 1945.

In Abbildung 4.9 sind neben der Grafik die für die Hochrechnung verwendeten Annahmen sowie die Bevölkerungsanteile dreier Altersklassen ausgewiesen (unter 20 Jahre, 20 bis einschließlich 66 Jahre, ab 67 Jahre). Die erste Kategorie repräsentiert die junge Bevölkerung, die zweite die Menschen im erwerbsfähigen Alter und die dritte die aus dem Erwerbsleben ausgeschiedenen Senioren. Der aus allen Alterswerten im Referenzjahr errechnete Median wird **Medianalter** genannt. Es ist das Alter, das von je 50% der Bevölkerung unter- bzw. überschritten wird. Im Jahr 2020 liegt das Medianalter der deutschen Bevölkerung bei ca. 45 Jahren.



Interaktives Objekt
„Bevölkerungsstruktur
zweier NRW-Städte“

Als **Jugendquotient** wird das Verhältnis „Anzahl jüngerer Menschen / Anzahl der Menschen im Erwerbsalter“ bezeichnet, als **Altenquotient** das Verhältnis „Anzahl der Ruheständler/ Anzahl der Menschen im Erwerbsalter“. Im Jahre 2020 gehören zur ersten Altersgruppe ungefähr 18% , zur zweiten Gruppe etwa 62% und zur dritten Gruppe etwa 19% der Bevölkerung. Hieraus ergibt sich ein Jugendquotient von 0,30 und ein Altenquotient von 0,31. Meist werden, wie in Abbildung 4.9, der Jugend- und auch der Altenquotient als Prozentwerte ausgewiesen. Dass die Bevölkerungsstruktur von Region zu Region stark variieren kann, spiegelt sich in unterschiedlichen Werten für den Jugend- und Altenquotienten wider. Für die NRW-Städte Herdecke und Wetter lag der Altenquotient im Jahr 2019 z. B. bei 41,7% resp. 34,7%. Die genannten Prozentwerte für den Altenquotienten beinhalten, dass auf 41,7 resp. 34,7 Personen im Ruhestand 100 Menschen im Erwerbsalter entfallen.

Anstieg des
Altenquotienten und
des Medianalters

Abbildung 4.10 zeigt, erneut auf der Basis der Annahmen G2-L2-W1, die Bevölkerungspyramide für 2030. Die Bevölkerungsstruktur von 2020 aus Abbildung 4.9 ist als Umriss eingeblendet. Auf diese Weise wird die Veränderung sehr deutlich. Die Gesamtbevölkerung wird unter den genannten Annahmen von 83,4 Millionen im Jahr 2020 auf 83,1 Millionen im Jahr 2030 und bis 2050 sogar auf etwa 77,6 Millionen zurückgehen. Während der Jugendquotient relativ konstant bleibt, erhöht sich der Altenquotient in den nächsten Jahrzehnten von 31 im Jahr 2020 auf 39 in 2030 und auf 49 im Jahr 2050. Damit einher geht eine Erhöhung des Medianalters. Man sieht diese Verschiebungen in Richtung Überalterung sehr deutlich. Der „Bauch“ des in Abbildung 4.10 nur in Umrissen wiedergegebenen Doppel-Histogramms für 2020 wandert schon bis 2030 deutlich nach oben. Der geburtenstärkste Jahrgang ist dann 66 Jahre alt und umfasst noch ca. 1,3 Millionen Menschen.

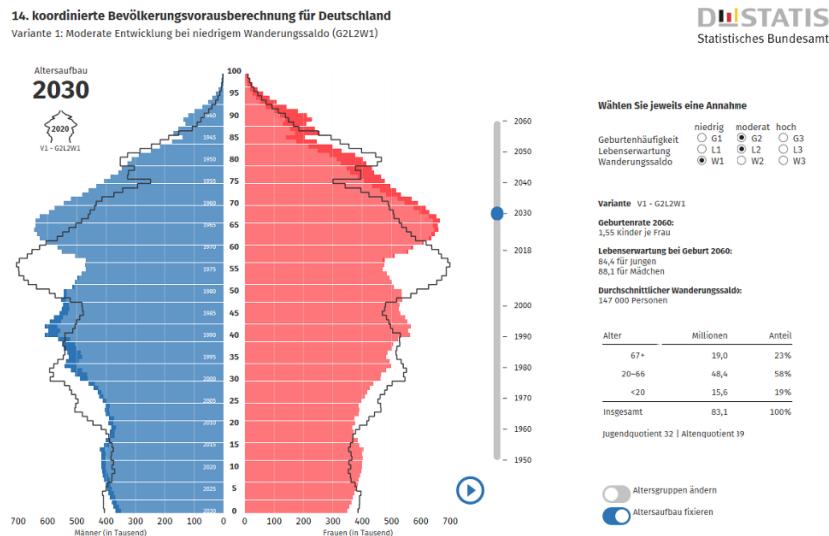


Abb. 4.10: Bevölkerungsstruktur von Deutschland im Jahr 2030 (Quelle: Statistisches Bundesamt; Basis: Annahmen G2-L2-W1)

Der Altenquotient wird sich jedenfalls in den nächsten Jahren spürbar erhöhen. Auch die steigende Lebenserwartung der Menschen stellt eine Herausforderung für die Renten- und Pensionskassen dar, weil sich damit die Bezugsdauer von Renten und Pensionen verlängert. Die bereits vollzogene Erhöhung des Ruhestand-Eintrittsalters von 65 auf 67 Jahre kann dies nur sehr bedingt auffangen. In einem Artikel vom 4. März 2019 widmete sich *Spiegel Online* dem Themenkomplex „Demografischer Wandel“ unter dem Blickpunkt regionaler Unterschiede.

Die Annahmen aus Tabelle 4.3 können durch politische und wirtschaftliche Ereignisse und Maßnahmen rasch obsolet werden. Wie sich die Zuwanderung nach Deutschland, die in 2015 unerwartet hoch war, in den nächsten Jahren entwickeln wird, weiß heute niemand. Geht man bei der Bevölkerungsvorausberechnung z. B. nicht von dem Annahmenset G1-L1-W2, sondern von G2-L2-W1 aus, ergibt sich für die Bevölkerung Deutschlands im Jahr 2050 anstelle von 77,6 Millionen ein Schätzwert von 75,2 Millionen. Die Animation des *Statistischen Bundesamts* zur Altersstruktur der Bevölkerung ist jedenfalls nur als Visualisierung von Wenn-Dann-Szenarien zu verstehen, die es ermöglicht, Auswirkungen denkbarer Entwicklungskorridore früh zu überblicken.



Interaktives Objekt
„Lebenserwartung
65-jähriger in
der EU“

Exkurs 4.1: Bevölkerungsstruktur und Migration

Wenn man vom demografischen Wandel spricht, sind Veränderungen des Bevölkerungsaufbaus gemeint, die mit dem Saldo aus Geburts- und Sterbefällen und der Veränderung des Anteils junger Menschen an der Gesamtbevölkerung zusammenhängen. Für Deutschland ist von einer weiter zunehmenden Überalterung auszugehen, die sich in Abbildung 4.10 schon widerspiegelt. Die Überalterungstendenz wäre noch viel ausgeprägter, wäre sie in den letzten Jahren und Jahrzehnten nicht durch Wellen starker Zuwanderungsbewegungen abgemildert worden. Es kamen zunächst Gastarbeiter aus der Türkei und Ländern der EU sowie Spätaussiedler aus Gebieten der ehemaligen Sowjetunion, später nach der Öffnung der Ostgrenzen und während des Bürgerkriegs im ehemaligen Jugoslawien Menschen aus Osteuropa und der Balkanregion, ab 2014 vermehrt Schutzsuchende aus Syrien, dem Irak und Afghanistan.

In Deutschland wird in der amtlichen Statistik zwischen Menschen mit und ohne Migrationshintergrund unterschieden. Eine Person hat einen Migrationshintergrund, wenn sie selbst oder zumindest ein Elternteil nicht mit deutscher Staatsangehörigkeit geboren wurde. Innerhalb der Gruppe der Menschen mit Migrationshintergrund – 20,8 Millionen Personen zählten Ende 2018 hierzu – wird noch danach differenziert, ob eigene Migrationserfahrung vorliegt oder nicht. Ende 2018 umfasste die Teilgruppe der Personen mit Migrationshintergrund und eigener Migrationserfahrung 13,5 Millionen Menschen (Zuwanderer), die Teilgruppe ohne eigene Migrationserfahrung 7,3 Millionen (Kinder von Zuwanderern). Zuwanderer sind heute nicht unbedingt mehr Ausländer. Viele sind inzwischen eingebürgert; ihre in Deutschland geborene Kinder erhalten bereits mit der Geburt die deutsche Staatsbürgerschaft.

Abbildung 4.11 zeigt eine Bevölkerungspyramide, die nicht nur nach Geschlecht differenziert, sondern auch danach, ob eine Person als Ausländer gilt oder die deutsche Staatsbürgerschaft besitzt. Bei den Deutschen wird noch zusätzlich unterschieden, ob ein Migrationshintergrund vorliegt oder nicht. In kräftigem Ocker ist innen der Bevölkerungsanteil der Ausländer, in hellem Grün im mittleren Bereich der Anteil der Deutschen mit Migrationshintergrund und außen in hellem Beige der Anteil der Deutschen ohne Migrationshintergrund wiedergegeben. Addiert man für jeden Jahrgang die drei farblich unterschiedlich gekennzeichneten Anteile, resultiert die Bevölkerungspyramide für die Gesamtbewölkerung. Die Bereiche in hellem Grün und kräftigem Ocker zusammen repräsentieren die 20,8 Millionen Menschen mit Migrationshintergrund.

Regionale Verteilung
von Ausländern
(interaktive Karte)



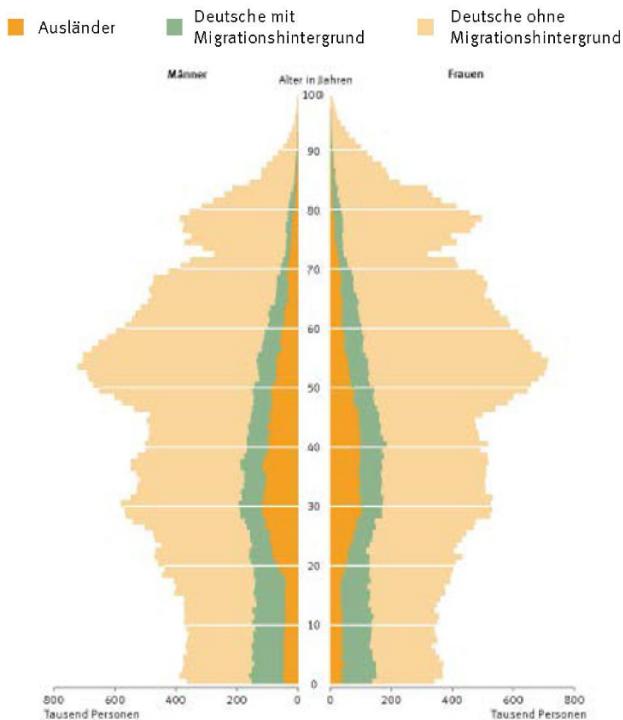


Abb. 4.11: Bevölkerungspyramide mit Differenzierung nach Migrationshintergrund; Stand: Ende 2018. Quelle: Statistisches Bundesamt

Vor allem die Menschen, die in der letzten großen Zuwanderungswelle aus Syrien, dem Irak und Afghanistan kamen, haben die Besetzung der jüngeren Jahrgänge in der Bevölkerungspyramide für Deutschland verstärkt. Abbildung 4.12 zeigt den Altersaufbau der Personengruppe, die 2018 als Schutzsuchende im Ausländerzentralregister erfasst waren. Man erkennt, dass es vor allem junge und hier überwiegend männliche Zuwanderer waren, die vor den Kriegswirren in ihren Ländern nach Deutschland flüchteten.

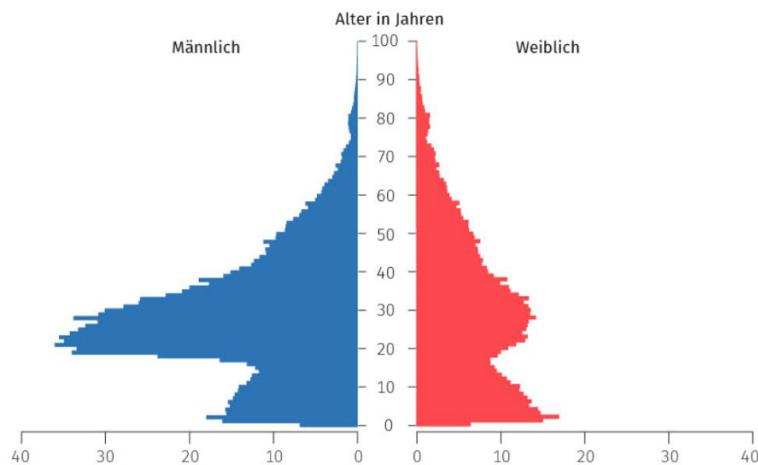


Abb. 4.12: Bevölkerungsaufbau registrierter Schutzsuchender, Besetzungshäufigkeiten der Jahrgänge in Tsd.; Stand: Ende 2018. Quelle: Statistisches Bundesamt

Das Thema „Demografischer Wandel und Migration“ wurde schon von SARRAZIN (2010) aufgegriffen. SARRAZIN warnte vor einer abnehmenden Konkurrenzfähigkeit Deutschlands im globalen Wettbewerb und brachte dies in Verbindung mit anhaltender Migration aus der Türkei, Nah- und Mittelost und Afrika und einem nach seiner Sicht hiermit verknüpften wachsendem Anteil von Menschen mit niedrigem Bildungsgrad.⁶ Seine Rechnungen, denen die unrealistische Annahme eines positiven Wanderungssaldos von nur 50 000 Personen pro Jahr zugrunde lag, basierten auf den Ergebnissen der 11. Bevölkerungsprognose für Deutschland von 2006.

Histogramme sind grafische Instrumente, mit denen die in umfangreichen Datensätzen enthaltenen Kerninformationen sichtbar werden können – z. B. bei Einkommensdaten die Asymmetrie von Einkommensverteilungen (stärkere Besetzung niedriger Einkommensklassen) oder beim Vergleich von Bevölkerungspyramiden die zunehmende Überalterung von Gesellschaften (stärkere Besetzung der Jahrgänge im Ruhestand).

- | | |
|---------------------------|---|
| Nachteil von Histogrammen | Der optische Eindruck eines Histogramms hängt von der Klasseneinteilung ab, also von der Breite und den Anfangspunkten der Klassen. Die Einkommensverteilungen in Abbildung 4.8 würden z. B. einen anderen Eindruck vermitteln, wenn man bei der Klassenbildung Intervalle von jeweils 1 000 oder 2 000 Euro wählte. Oft werden daher alternativ sog. Kerndichteschätzer verwendet, die man als Verallgemeinerung des Konzepts der Histogramme ansehen kann. Auf diese kann hier nicht näher |
|---------------------------|---|

⁶Der Bildungsgrad wird in amtlichen Statistiken anhand der International Standard Classification of Education (ISCED) erfasst, einem von der UNESCO zuletzt im November 2011 überarbeiteten Klassifikationsschema für Bildungssysteme.

eingegangen werden (vgl. aber z. B. FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 2.4.3). Es sei nur erwähnt, dass die Treppenfunktion, die den oberen Rand eines Histogramms darstellt, bei Kerndichteschätzern durch eine stetige Funktion ersetzt wird.



4.3 Die empirische Verteilungsfunktion

In Abschnitt 4.1 wurde dargelegt, dass sich ein *diskretes* Merkmal X mit k Ausprägungen a_1, \dots, a_k anhand der absoluten oder relativen Häufigkeiten $h(a_1), \dots, h(a_k)$ bzw. $f(a_1), \dots, f(a_k)$ beschreiben lässt und zwar für jeden Typ von Merkmalsskalierung. Die k Häufigkeiten repräsentieren die **absolute Häufigkeitsverteilung** resp. **relative Häufigkeitsverteilung** des Merkmals. Grafisch kann eine Häufigkeitsverteilung u. a. anhand eines Stab- oder Balkendiagramms veranschaulicht werden (vgl. Abbildung 4.4). Für *stetige* Merkmale kann man die Werte einer Urliste zu k Klassen zusammenfassen und die Klassenbesetzungshäufigkeiten, wie in Abbildung 4.8 illustriert, anhand eines Histogramms visualisieren.



Interaktives Objekt
„Relative
Häufigkeiten
(Augenzahlen)“

Wenn die Merkmalswerte metrisch oder zumindest ordinalskaliert sind, also eine natürliche Rangordnung erklärt ist, will man oft auch wissen, wie viele Werte unterhalb oder oberhalb eines Schwellenwertes x liegen. Eine Antwort auf solche Fragen liefert die kumulierte Häufigkeitsverteilung.

Übergang zu
kumulierten Häufig-
keitsverteilungen

Betrachtet sei also ein zumindest ordinalskaliertes Merkmal X mit Ausprägungen a_1, \dots, a_k , die nach aufsteigendem Rang bzw. aufsteigender Größe geordnet seien. Für das Merkmal liegen n Beobachtungen x_i vor ($i = 1, 2, \dots, n$). Die **absolute kumulierte Häufigkeitsverteilung** für X ergibt sich, wenn man für einen beliebigen reellen Wert x die Anzahl der Beobachtungen ermittelt, die x nicht überschreiten. Formal ergibt sich die kumulierte Häufigkeitsverteilung $H(x)$ als Summe der absoluten Häufigkeiten $h(a_i)$, die der Bedingung $a_i \leq x$ genügen. Formal lässt sich $H(x)$ wie folgt schreiben:

$$H(x) = \begin{cases} 0 & \text{für } x < a_1 \\ h_1 & \text{für } a_1 \leq x < a_2 \\ \vdots & \vdots \\ h_1 + h_2 + \dots + h_{k-1} & \text{für } a_{k-1} \leq x < a_k \\ n & \text{für } x \geq a_k. \end{cases} \quad (4.3)$$

Die Funktion $H(x)$ ist also für $x < a_1$ Null, springt in $x = a_1$ auf den Wert $h(a_1)$ und bleibt auf diesem Niveau bis zur Stelle $x = a_2$, an der sie auf den Wert $h(a_1) + h(a_2)$ springt usw. Die absolute kumulierte Häufigkeitsverteilung $H(x)$ für ein Merkmal X ist somit eine monoton steigende Treppenfunktion, die jeweils in $x = a_i$ um h_i nach oben springt.

Die **relative kumulierte Häufigkeitsverteilung** $F(x)$ resultiert, wenn man $H(x)$ durch den Umfang n des Datensatzes dividiert:

$$F(x) = \frac{H(x)}{n}. \quad (4.4)$$

Die Funktion (4.4) wird oft als **empirische Verteilungsfunktion** angesprochen. Sie besitzt in ausführlicher Schreibweise die Darstellung

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f_1 & \text{für } a_1 \leq x < a_2 \\ \vdots & \vdots \\ f_1 + f_2 + \dots + f_{k-1} & \text{für } a_{k-1} \leq x < a_k \\ 1 & \text{für } x \geq a_k, \end{cases} \quad (4.5)$$

repräsentiert also ebenfalls eine monoton steigende Treppenfunktion, die aber in $x = a_i$ ($i = 1, 2, \dots, k$) jeweils um f_i springt. Die Funktion $F(x)$ geht demnach aus (4.3) hervor, wenn man dort die absoluten Häufigkeiten h_i durch die relativen Häufigkeiten f_i ersetzt.



Interaktives Objekt
„Empirische
Verteilungsfunktion
(Augenzahlen)“

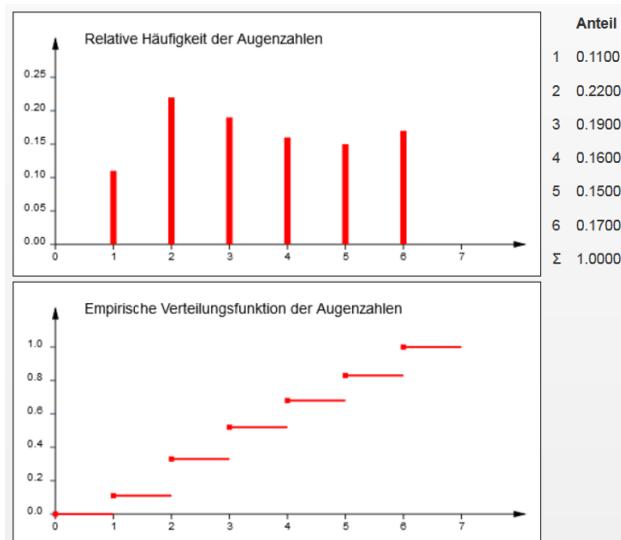


Abb. 4.13: Relative Häufigkeitsverteilung und relative kumulierte Häufigkeitsverteilung (100-faches Würfeln mit einem Würfel)

Abbildung 4.13 zeigt im oberen Teil die beobachteten relativen Häufigkeiten für die Ausprägungen $a_i = i$ ($i = 1, 2, \dots, 6$) des Merkmals „Augenzahl X “ bei einem Würfelexperiment, bei dem ein Würfel 100 Mal geworfen wurde. Die relativen Häufigkeiten $f(a_i) = \frac{h(a_i)}{100}$, die den Anteil

der einzelnen Augenzahlen an der Anzahl der Würfe repräsentieren, sind neben der Grafik in einer kleinen Tabelle wiedergegeben. Aus der Tabelle geht z. B. hervor, dass die Augenzahl 4 insgesamt 16 Mal auftrat.

Der untere Teil der Grafik veranschaulicht die empirische Verteilungsfunktion des Merkmals „Augenzahl X “. Die Funktion $F(x)$ kann nur an den Stellen $x = a_i = i$ Sprünge aufweisen. Sie springt bei dem hier durchgeföhrten Experiment an diesen Stellen um die Werte, die in der Tabelle neben der Grafik zu sehen sind. An der Stelle $x = 2$ beträgt die Sprunghöhe z. B. 0,02 und bei $x = 4$ hat sie den Wert 0,16. Zwischen zwei benachbarten Ausprägungen von X bleibt die empirische Verteilungsfunktion auf konstantem Niveau.

Wenn man das beschriebene Experiment mit *zwei* Würfeln durchführt und bei jedem Wurf die Augensumme ermittelt, kann die empirische Verteilungsfunktion des Merkmals „Augensumme X “ bis zu 11 Sprünge aufweisen, nämlich für $x = a_i = i$ mit $i = 2, 3, \dots, 12$. Abbildung 4.14 zeigt die Häufigkeitsverteilung der Augensumme für ein Experiment, bei dem zwei Würfel 100-mal geworfen wurden. Dabei wurde z. B. die Augensumme 2 insgesamt 2 Mal und die Summe 9 insgesamt 13 Mal beobachtet. Die Anteilswerte sind auch hier neben der Grafik tabelliert.



Interaktives Objekt
„Empirische
Verteilungsfunktion
(Augensummen)“

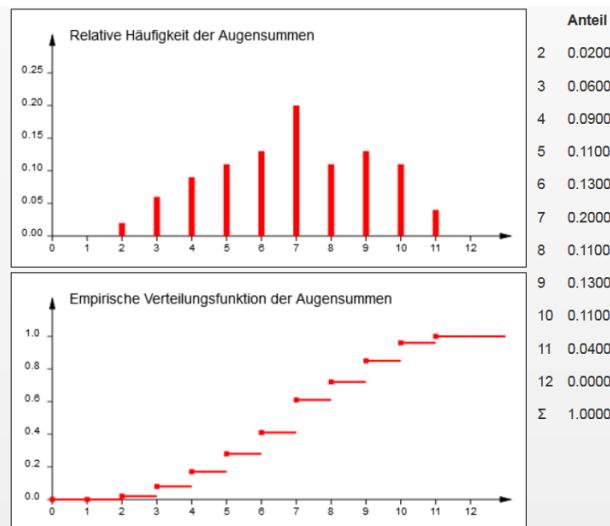


Abb. 4.14: Relative Häufigkeitsverteilung und relative kumulierte Häufigkeitsverteilung (100-faches Würfeln mit zwei Würfeln)



Bei Datensätzen für *stetige* Merkmale sind i. d. R. alle Elemente des Datensatzes verschieden. Hier wird man zweckmäßigerweise zu einer Klassenbildung übergehen und Klassenbesetzungshäufigkeiten darstellen, also ein Histogramm wählen. Die beiden Histogramme in Abbildung 4.8, die

Aufgabe 4.3

sich auf sehr große Datensätze für Bruttojahresverdienste von Arbeitnehmern beziehen, sind z. B. weitaus übersichtlicher als eine Darstellung der Häufigkeiten für die originären Verdienstdaten in Euro und Cent.



5 Kenngrößen empirischer Verteilungen



Vorschau auf
das Kapitel

Die in einem Datensatz für ein Merkmal enthaltene Information lässt sich zu Kenngrößen verdichten. Diese charakterisieren das Zentrum oder die Variabilität des Datensatzes. Man hat also Kenngrößen zur Beschreibung der „mittleren“ Lage der Elemente eines Datensatzes und solche zur Charakterisierung der Streuung. Als Lageparameter werden der Modalwert, der Median und das arithmetische Mittel vorgestellt, als Streuungsparameter die empirische Varianz bzw. die empirische Standardabweichung sowie die Spannweite. Welche Kenngröße in Betracht kommt, hängt davon ab, ob die Daten auf einer Nominal-, Ordinal- oder einer metrischen Skala erfasst wurden.

Abschließend wird der Lageparameter „Median“ verallgemeinert und der Begriff des Quantils eines Datensatzes eingeführt. Während der Median einen nach Größe geordneten Datensatz in zwei Anteile gleicher Größe $p = 0,5$ zerlegt, erfolgt bei einem p -Quantil eine Zerlegung in beliebige Anteile p und $1 - p$. Ein Visualisierungsinstrument für einen Datensatz, das von mehreren Quantilen Gebrauch macht, ist der Boxplot. In der einfachsten Variante veranschaulicht ein Boxplot die beiden Extremwerte und drei Quantile, nämlich das 0,25-, 0,5- sowie das 0,75-Quantil eines Datensatzes. Erwähnt wird auch das Kerzendiagramm – eine im Finanzsektor verbreitete Modifikation des Boxplots.

5.1 Lagemaße

Häufigkeitsverteilungen für ungruppierte oder gruppierte Daten vermitteln einen Eindruck von der Gestalt der Verteilung eines Datensatzes. Die Histogramme in Abbildung 4.8 zur Verteilung von Bruttoverdiensten in zwei südeuropäischen Staaten zeigen z. B., dass die Verteilung der Daten in beiden Fällen eine deutliche Asymmetrie aufweist, also eine gewisse „Schiefe“ der Verteilung zu beobachten ist. Ferner sieht man bei beiden Teilgrafiken, dass das „Zentrum“ (oder der „Schwerpunkt“) der Einkommensverteilung für Portugal im Bereich kleinerer Werte liegt und auch die „Streuung“ hier geringer ist. Die Begriffe „Zentrum“, „Schwerpunkt“, „Streuung“ oder „Schiefe“ einer Verteilung sind zunächst unscharf und bedürfen der Präzisierung. Lage- und Streuungsparameter dienen dem Zweck, solche Befunde zu präzisieren und zu objektivieren. Es geht darum, die in einem Datensatz steckende Information zu wenigen Kenngrößen zu verdichten. Eine solche Informationsverdichtung ermöglicht eine unmissverständliche Beschreibung von Charakteristika eines Datensatzes, ist aber grundsätzlich mit Informationsverlust verbunden. So können zwei sehr unterschiedliche Datensätze einen ähnlichen Schwerpunkt oder eine vergleichbare Streuung aufweisen. Kenngrößen zur Beschreibung

Wofür werden
Kenngrößen von
Verteilungen
benötigt?

empirischer Verteilungen sind aber dennoch wichtig. Sie liefern für einen gegebenen Datensatz nämlich wertvolle zusätzliche Informationen, die sich visuell aus der grafischen Darstellung einer empirischen Verteilung nicht immer ohne weiteres erschließen.

Zur Charakterisierung des „Zentrums“ einer Verteilung werden Lageparameter herangezogen. Ein besonders leicht zu bestimmender Lageparameter ist der **Modus** oder **Modalwert** x_{mod} (lies: x -*mod*). Dieser lässt sich immer anwenden, also auch bei Merkmalen, deren Ausprägungen nur Kategorien sind (qualitative Merkmale). Er ist definiert als die Merkmalsausprägung mit der größten Häufigkeit.

Beispiel 5.1: Modus beim Datensatz zum ZDF-Politbarometer

In Beispiel 4.1 (ZDF-Politbarometer vom 8. Dezember 2017, Merkmal „Parteipräferenz“) war die Ausprägung a_1 (Präferenz für die CDU/CSU) mit der größten Häufigkeit verbunden, d. h. hier ist $x_{mod} = a_1$. Anhand von Abbildung 4.4 lässt sich der Modus leicht bestimmen, weil die Häufigkeit $h(a_1)$ deutlich größer als alle anderen Häufigkeiten war. Wären zwei Häufigkeiten, z. B. $h(a_1)$ und $h(a_2)$ gleich groß, hätte man eine zweigipflige Häufigkeitsverteilung und es gäbe zwei Modalwerte (Modi). Der Modus ist also nur dann eindeutig erklärt, wenn die Häufigkeitsverteilung ein eindeutig bestimmtes Maximum aufweist.

Ein weiterer Lageparameter ist der **Median** \tilde{x} (lies: x -*Schlange*), der gelegentlich mit x_{med} abgekürzt wird (lies: x -*med*) und für den man auch die Bezeichnung **Zentralwert** findet. Der Median ist nur bei mindestens ordinalskalierten Merkmalen anwendbar, also bei Merkmalen, für deren Werte eine natürliche Rangordnung erklärt ist. Betrachtet sei also ein – noch nicht notwendigerweise geordnet vorliegender – Datensatz x_1, x_2, \dots, x_n für ein solches Merkmal. Um zwischen dem ursprünglichen und dem geordneten Datensatz unterscheiden zu können, sei letzterer mit $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ bezeichnet. Der Median ist dann, grob gesprochen, der „mittlere“ Wert des geordneten Datensatzes. Bei ungeradem n ist dies der eindeutig bestimmte Wert $x_{(\frac{n+1}{2})}$. Bei geradem n gibt es hingegen zwei Werte $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$, die die Mitte des Datensatzes repräsentieren. In diesem Falle ist der Median bei einem ordinalskalierten Merkmal nicht eindeutig bestimmt, sofern sich die beiden Werte $x_{(\frac{n}{2})}$ und $x_{(\frac{n}{2}+1)}$ voneinander unterscheiden. Bezieht sich der Datensatz hingegen auf ein metrisch skaliertes Merkmal, so kann man eine eindeutige Festlegung des Medians erreichen, in dem man aus den beiden zentralen Werten den Mittelwert bildet. Der Median ist dann definiert durch

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \cdot (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{falls } n \text{ gerade.} \end{cases} \quad (5.1)$$

Der bekannteste Lageparameter ist das **arithmetische Mittel**, das im Folgenden auch als **Mittelwert** angesprochen und mit \bar{x} abgekürzt wird (lies: *x-quer*). In der Umgangssprache findet man häufig die Bezeichnung **Durchschnitt**. Der Mittelwert ist nur bei metrisch skalierten Merkmalen anwendbar und ergibt sich, indem man alle Werte x_1, x_2, \dots, x_n eines Datensatzes addiert und die resultierende Summe durch n dividiert:¹

$$\bar{x} := \frac{1}{n} \cdot (x_1 + x_2 + \dots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^n x_i. \quad (5.2)$$

Der Mittelwert berücksichtigt demnach alle Werte eines Datensatzes mit gleichem Gewicht $\frac{1}{n}$, während in die Formel für den Median nur ein oder zwei zentrale Elemente eines Datensatzes eingehen. Wenn man also bei einem Datensatz den größten Wert $x_{\max} = x_{(n)}$ deutlich vergrößert, hat dies nur auf den Mittelwert einen Effekt. Der Mittelwert reagiert demnach, anders als der Median, empfindlich gegenüber extremen Werten. Man spricht in diesem Zusammenhang von einer höheren *Sensitivität* oder von einer geringeren *Robustheit* des Mittelwerts gegenüber Ausreißern.

Wenn man von jedem der Elemente x_1, x_2, \dots, x_n eines Datensatzes den Mittelwert subtrahiert und aufsummiert, resultiert 0:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0. \quad (5.3)$$

Gleichung (5.3) beinhaltet, dass sich der Mittelwert als Schwerpunkt des Datensatzes interpretieren lässt.



Interaktives Objekt
„Lageparameter“

Beispiel 5.2: Median und Mittelwert für Energieverbrauchsdaten

Die *Internationale Energieagentur (IEA)* veröffentlicht Daten zu den Themenfeldern Energieproduktion, Energieverbrauch und Umwelt. Tabelle 5.1 enthält ausgewählte Daten der IEA für das Jahr 2015, die sich auf 5 Länder beziehen. Man erkennt, dass die USA vergleichsweise großzügig Energie verbrauchen und CO_2 emittieren. Gedanklich stelle man sich 5 Personen vor, je eine Person aus den Ländern USA, Deutschland, Japan, China und Indien, für die jeweils die in Tabelle 5.1 angegebenen Verbrauchs- und Emissionswerte zutreffen, die also bezüglich der drei Merkmale als typische Vertreter ihrer Länder gelten können. Für diese kleine Personengruppe lässt sich dann der „mittlere“ Pro-Kopf-Verbrauch für Öl und Strom bzw. eine „mittlere“ CO_2 -Emission ermitteln, wobei man jeweils den Median oder den Mittelwert heranziehen kann.



Interaktives Objekt
„Emission von
Treibhausgasen“

¹Das Zeichen Σ und andere mathematische Symbole sind in Abschnitt 21.3 erklärt.

Land	Erdölverbrauch (in t/Kopf)	Stromverbrauch (in 1000 kWh/Kopf)	CO_2 -Emissionen (in t/Kopf)
USA	2,52	12,83	15,5
Japan	1,43	7,86	9,0
Deutschland	1,22	7,01	8,9
China	0,38	4,05	6,6
Indien	0,14	0,86	1,6



Tab. 5.1: Umweltrelevante Daten für fünf Staaten (2015)



Interaktives Objekt
„Erneuerbare
Energien in der EU“

Es seien hier für das metrisch skalierte Merkmal „Stromverbrauch/Kopf“ (in 1000 kWh) die Daten in der mittleren Spalte von Tabelle 5.1 betrachtet. Um den Median zu errechnen, sind die Werte $x_1 = 12,83$, $x_2 = 7,01$, $x_3 = 7,86$, $x_4 = 4,05$, $x_5 = 0,86$ zunächst nach Größe zu ordnen. Aus der resultierenden Folge $x_{(1)} = 0,86$, $x_{(2)} = 4,05$, $x_{(3)} = 7,01$, $x_{(4)} = 7,86$, $x_{(5)} = 12,83$ ergibt sich der Median für den hier vorliegenden Fall $n = 5$ nach (5.1) als $\tilde{x} = x_{(3)} = 7,01$. Würde man bei dem ursprünglichen Datensatz den Wert $x_5 = 0,86$ für Indien unberücksichtigt lassen, den Median also nur auf der Basis der Datenreihe x_1, \dots, x_4 ermitteln, erhielte man für \tilde{x} den Wert $\tilde{x} = \frac{1}{2} \cdot (7,01 + 7,86) = 7,435$.

Bestimmt man mit den 5 Ausgangsdaten den Mittelwert, so erhält man nach (5.2) den Wert $\bar{x} = \frac{1}{5} \cdot 32,61 = 6,522$. Würde man für x_1 anstelle von 12,83 den 10-fach größeren Wert 128,3 einsetzen, bliebe der Median bei $\tilde{x} = 7,01$, während sich für den Mittelwert nun $\bar{x} = \frac{1}{5} \cdot 148,08 = 29,616$ ergäbe.

Alternative
Berechnung des
Mittelwerts

Die Berechnung des arithmetischen Mittels kann einfacher bewerkstelligt werden, wenn Merkmalswerte mehrfach auftreten. Hat man für ein diskretes Merkmal X mit den Ausprägungen a_1, \dots, a_k insgesamt n Beobachtungswerte x_1, \dots, x_n ($n > k$), so würde die Anwendung von (5.2) implizieren, dass n Werte zu addieren sind. Anstelle der Urliste kann man hier für die Berechnung des Mittelwerts auch die relative Häufigkeitsverteilung $f(a_1), \dots, f(a_k)$ verwenden und \bar{x} nach

$$\bar{x} := a_1 \cdot f_1 + a_2 \cdot f_2 + \dots + a_k \cdot f_k = \sum_{i=1}^k a_i \cdot f_i \quad (5.4)$$

als Summe von nur k Termen berechnen. Das arithmetische Mittel \bar{x} lässt sich also alternativ als Summe der mit den relativen Häufigkeiten f_i gewichteten Ausprägungen a_i ermitteln ($i = 1, 2, \dots, k$). Erzielt man bei 10-maligem Würfeln z. B. vier Mal die 1, zwei Mal die 4, drei Mal die 5 und einmal die 6, ergibt sich sowohl nach (5.2) als auch nach (5.4) der Wert $\bar{x} = 3,3$.

Die Formel (5.4) lässt sich in leicht modifizierter Fassung auch zur Berechnung des Mittelwerts bei *gruppierten Daten* verwenden. Man hat nur

die Ausprägungen a_i durch die Mitte m_i der Klassen zu ersetzen und die Häufigkeiten f_i sind dann die relativen Klassenbesetzungshäufigkeiten.

Welchen der vorgestellten Lageparameter sollte man aber verwenden? Hierzu gibt es keine allgemeingültige Ausage. Die Antwort hängt sowohl von der Skalierung des Merkmals ab als auch von der jeweiligen Fragestellung. Bei einem nominalskalierten Merkmal kann man nur den Modalwert verwenden. Bei einem metrisch skalierten Merkmal hat man schon drei Alternativen, nämlich den Modalwert, den Median und das arithmetische Mittel, und es ist zu überlegen, wie robust die zu berechnende Kenngröße gegenüber Extremwerten sein soll. Bei einem kleinen Datensatz für das Merkmal „Bruttoverdienst“ (in Euro/Stunde) kann z. B. ein einziger Extremwert das arithmetische Mittel erheblich beeinflussen. Hier kann dann der Median aussagekräftiger sein, während der Modalwert i. Allg. wenig Information liefert. Bei metrisch skalierten Daten wird oft nicht nur ein Lageparameter berechnet, weil ein zweiter Parameter, etwa der Median zusätzlich neben dem Mittelwert, noch zusätzliche Information über die empirische Verteilung eines Datensatzes vermitteln kann. Bei einer Einkommensverteilung kann man z. B. \bar{x} und \tilde{x} vergleichen und hieraus Aussagen zur Symmetrie oder Asymmetrie der Verteilung ableiten.

Gibt es einen „besten“ Lageparameter?

Beispiel 5.3: Haushaltseinkommen in Großbritannien

Im März 2005 veröffentlichte das *Institute for Fiscal Studies* (IFS), ein unabhängiges Wirtschaftsforschungsinstitut in Großbritannien, einen Bericht „*Poverty and Inequality in Britain*“, in dem u. a. angeführt wurde, dass das mittlere verfügbare Hauseinkommen („average take-home income“) im Land im Zeitraum 2003/04 gegenüber dem Vorjahreszeitraum um 0,2 % abgenommen habe auf nunmehr 408 £ (Britische Pfund). Dieser Befund wurde von der Presse kritisch kommentiert, so dass schließlich Gordon BROWN, der damalige Schatzkanzler und spätere Premierminister, Stellung beziehen musste.

Die von den Medien aufgegriffene Information bezog sich auf den *Mittelwert* der Variablen „verfügbares Hauseinkommen“. Der Bericht führte aber auch an, ohne dass dies allerdings von den Journalisten aufgegriffen wurde, dass der *Median* im fraglichen Zeitraum um 0,5 % gestiegen war und jetzt 336 £ betrug. Der Median wäre aber zur Charakterisierung des „durchschnittlichen“ Haushaltseinkommens weitaus geeigneter als das arithmetische Mittel, weil Einkommensverteilungen asymmetrisch sind und der Mittelwert hier durch extrem hohe und für die Grundgesamtheit eher untypische Werte stark beeinflusst werden kann (vgl. auch Abbildung 4.8). Der Anstieg des Medians um 0,5 % war bei dem IFS-Bericht die weitaus aussagekräftigere und positiv zu bewertende Information. Sie beinhaltete nämlich, dass der Wert, der die unteren 50 % der Haushaltseinkommen von den oberen 50 % trennte, sich leicht nach oben verschoben hatte, d. h. die Ungleichheit der Verteilung der Haushaltseinkommen hatte leicht abgenommen.



Gordon BROWN.
Quelle: World Economic Forum



Video von Full Fact
zum EU-Nettobeitrag
Großbritanniens

Methodenkompetenz ist jedenfalls eine Voraussetzung dafür, statistische Sachverhalte in den Medien sachadäquat bewerten und unscharfe oder gar manipulative Darstellungen erkennen zu können. In Großbritannien gibt es mit **Full Fact** eine vielbeachtete unabhängige Institution, die von Medien oder Politikern verbreitete fragwürdige statistische Informationen öffentlich richtigstellt. Aussagen, die gezielt auf Desinformation setzen („Fake News“, „Alternative Fakten“), haben damit eine geringere Chance, Wirkungskraft zu entfalten. Richtiggestellt wurde von *Full Fact* z. B. die 2016 im Vorfeld des Brexit-Referendums von namhaften Brexit-Befürwortern, u. a. den Tories Boris JOHNSON und Michael GOVE, zur Erreichung des Austrittsziels wiederholt vorgetragene Falschaussage, dass Großbritannien wöchentlich 350 Millionen £ an die EU zahle.

Eigenschaften von
Mittelwert und
Median

Arithmetisches Mittel (Mittelwert) und Median sind Lösungen unterschiedlicher Minimierungsprobleme. Der Mittelwert hat die Eigenschaft, für einen Datensatz x_1, x_2, \dots, x_n denjenigen Wert z zu repräsentieren, der die Summe der quadrierten Abweichungen $(x_i - z)^2$ minimiert:

$$z = \bar{x} : \quad \sum_{i=1}^n (x_i - z)^2 \rightarrow \text{Min.}$$

Der Median minimiert die Summe der absoluten Abweichungen $|x_i - z|$:

$$z = \tilde{x} : \quad \sum_{i=1}^n |x_i - z| \rightarrow \text{Min.}$$

Beweise findet man bei SCHLITTGEN (2012, Abschnitt 3.1).

Weitere
Lageparameter

Für metrisch skalierte Merkmale gibt es noch weitere Lageparameter. Zu nennen ist das **gewichtete arithmetische Mittel**, bei dem die Werte x_1, x_2, \dots, x_n eines Datensatzes mit unterschiedlichen Gewichten versehen werden. Will man z. B. den mittleren Stromverbrauch für alle Einwohner der in Tabelle (5.1) aufgeführten 5 Länder berechnen, nicht nur für eine modellhafte Gruppe von 5 Ländervertretern, so bezöge sich die Mittelwertbildung auf einen Datensatz, dessen Umfang n durch die Summe $n_1 + n_2 + n_3 + n_4 + n_5$ der Bevölkerungszahlen aller 5 Länder gegeben wäre. Um die unterschiedlichen Bevölkerungsstärken zu berücksichtigen, wird der Wert x_i für jedes Land mit dem Gewichtungsfaktor n_i multipliziert.

Zu erwähnen ist auch das **getrimmte arithmetische Mittel**. Dieses lässt einen kleineren Anteil der Randdaten $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ eines nach aufsteigender Größe geordneten Datensatzes unberücksichtigt. Wenn dieser Anteil α beträgt, spricht man auch von einem α -getrimmten Mittelwert und kürzt diesen mit \bar{x}_α ab. Bei der Berechnung von \bar{x}_α werden die unteren und oberen $\frac{\alpha}{2} \cdot 100\%$ des geordneten Datensatzes vor der

Mittelwertberechnung eliminiert. Das führt dazu, dass getrimmte Mittelwerte, ähnlich wie der Median, robuster gegenüber Extremwerten (Ausreißerdaten) sind.

Das mit \bar{x}_g bezeichnete **geometrische Mittel** wird für Datensätze x_1, x_2, \dots, x_n verwendet, die Veränderungsraten repräsentieren, z. B. Wachstumsraten. Es errechnet sich als

$$\bar{x}_g = \sqrt[q]{x_1 \cdot x_2 \cdot \dots \cdot x_n}.$$

Bei Zeitreihendaten verwendet man zur Glättung häufig lokale arithmetische Mittel, die **gleitende Durchschnitte** genannt werden. Der Durchschnitt wird hier aus den Werten innerhalb eines gleitenden „Fensters“ gebildet, das außer dem aktuellen Zeitpunkt t je q Werte vor und nach t berücksichtigt.

Exkurs 5.1: Einkommensungleichheit und Armut

Armut ist ein multidimensionales Phänomen. Es bezeichnet einen Mangel an Einkommen, lebenswichtigen Gütern und Dienstleistungen (Nahrung, Kleidung, Obdach, medizinische Versorgung, Bildung). Unter *absoluter* oder *extremer* Armut versteht man einen existenzbedrohenden Mangelzustand, bei dem der lebenswichtige Grundbedarf nicht gesichert ist. Diese Form von Armut wird in etlichen Schwellen- und Entwicklungsländern beobachtet.

Wenn Statistische Ämter in Europa über Armut berichten, ist etwas anderes gemeint, nämlich *relative Armut*. Diese setzt das verfügbare Einkommen einer Person in einem Staat in Beziehung zum Einkommen, das andere Personen in diesem Staat im „Mittel“ ausgeben können. Es wird also nicht Armut im Sinne eines existenzbedrohenden Mangelzustands gemessen. In Deutschland und auch in anderen Ländern der Europäischen Union verwendet man in der amtlichen Statistik häufig den Begriff der *Armutgefährdungsgrenze*. Eine in einem Ein-Personen-Haushalt lebende Person wird als „armutsgefährdet“ angesehen, wenn ihr weniger als 60% des Medians der Einkommensverteilung in einem Land oder einer Region zur Verfügung stehen. Mit „Einkommen“ ist hier das Nettoeinkommen unter Einbezug staatlicher Transferleistungen gemeint. Nimmt man ein ganzes Land als Bezugsgröße, ist die Armutgefährdungsgrenze durch 60% des Medians der nationalen Einkommensverteilung definiert. Da sich die Einkommensverteilung eines Landes im Zeitverlauf ändert, ändert sich auch der als Bezugsgröße verwendete Wert von 60% des Medianeinkommens. Im Jahr 2011 lag dieser Wert für Deutschland bei 849 Euro/Monat, 2014 bei 917 Euro/Monat und 2017 bei 999 Euro/Monat.

Bei Mehrpersonenhaushalten werden fiktive Pro-Kopf-Haushaltseinkommen errechnet, sogenannte *Äquivalenzeinkommen*. In deren Berechnung gehen die Mitglieder eines Haushaltes mit unterschiedlichen Gewichten ein, um Einspareffekte abzubilden, die beim Zusammenleben mehrerer Personen erzielt werden können. Wo Armutgefährdung in Armut übergeht, ist nicht einheitlich

definiert, bestimmt sich aber wieder über einen Prozentsatz des Medians der zugrunde gelegten Einkommensverteilung. Eurostat und nationale Statistikämter in Europa verwenden 40% des Medians als Schwellenwert, der die Kategorien „arm/nicht-arm“ trennt.

Die Verwendung der nationalen Einkommensverteilung bei der Definition von Armut und Armutgefährdung für Deutschland impliziert, dass regionale Einkommensunterschiede, etwa solche zwischen Bundesländern, ausgeblendet sind. Dies hat zur Folge, dass in Regionen mit hohen Lebenshaltungskosten – z. B. in München – die Quote der als „arm“ geltenden Menschen unterschätzt wird, d. h. die offizielle Armutssquote kann hier deutlich nach unten verfälscht sein. Um die Messung von Armut und Armutgefährdungsquoten besser auf unterschiedliche Regionen zu beziehen, veröffentlicht das *Statistische Bundesamt* alljährlich auch Armutgefährdungsgrenzen für die 16 Bundesländer. Die Schwellenwerte für Bremen liegen z. B. durchweg unter denen von Baden-Württemberg, weil der Median der Einkommensverteilung für das Bundesland Bremen niedriger liegt als der für Baden-Württemberg. Für 2017 lag die Armutgefährdungsgrenze für Bremen z. B. bei 914 Euro/Monat, somit unter dem für Gesamtdeutschland geltenden Wert, und für Baden-Württemberg bei 1 091 Euro/Monat. Quantitative Informationen zum Anteil der Armutgefährdeten sind demnach nur sinnvoll interpretierbar, wenn man die Bezugsregion kennt.

Die in Europa gängige Verwendung nationaler Einkommensverteilungen bei Analysen zur Armutgefährdung bedingt, dass eine Person, die in Deutschland als armutgefährdet gilt, nicht unbedingt in einem Nachbarland zu dieser Personengruppe zählt. Einer am 27. März 2012 veröffentlichten Pressemitteilung des Statistischen Bundesamts entnahm man z. B., dass 60% des Medians der nationalen Einkommensverteilung in der Tschechischen Republik im Jahr 2009 bei 353 Euro/Monat lag. Trotz dieses im Vergleich zu Deutschland viel niedrigeren Schwellenwerts lag der Anteil der armutgefährdeten Personen an der Gesamtbevölkerung in Tschechien bei nur 9,0% und damit deutlich unter der damals für Deutschland ermittelten Quote von 15,6%.

Die Daten zur Armutgefährdung werden im Rahmen einer Erhebung über Einkommen und Lebensbedingungen in Europa gewonnen. Die Erhebung ist unter dem Kürzel *EU-SILC* bekannt (*European Union Statistics on Income and Living Conditions*), in Deutschland unter „Leben in Europa“.

5.2 Streuungsmaße

Ein Datensatz definiert eine empirische Verteilung eines Merkmals. Das „Zentrum“ einer solchen Verteilung kann man anhand einer oder mehrerer Kenngrößen charakterisieren. Bei einem metrisch skalierten Merkmal stehen vor allem der Modalwert, der Median und der Mittelwert zur Verfügung, wobei man hier i. Allg. den Mittelwert oder den Median verwenden wird. Die Kenntnis des Schwerpunktes reicht aber nicht aus, um

einen Datensatz zu beschreiben. Zwei Datensätze können in den Lageparametern übereinstimmen und sich dennoch bezüglich der Variation der Merkmalswerte deutlich unterscheiden. Hat man z. B. einen Datensatz x_1, x_2, \dots, x_n mit Mittelwert \bar{x} , so lässt die alleinige Kenntnis von \bar{x} offen, ob die einzelnen Elemente des Datensatzes alle sehr nahe am Mittelwert liegen, mit ihm gar alle übereinstimmen oder von \bar{x} stark abweichen und sich nur „ausmitteln“. Zur Charakterisierung von Merkmalen, für die Abstände zwischen Merkmalsausprägungen erklärt sind, also bei quantitativen Merkmalen, muss man noch Kenngrößen heranziehen, die die Streuung innerhalb des Datensatzes messen.

Warum braucht man auch Kenngrößen für die Streuung?

Ein besonders einfaches Streuungsmaß für metrisch skalierte Merkmale ist die **Spannweite** R eines Datensatzes (engl.: *range*). Um diese zu berechnen, ordnet man – wie bei der Berechnung des Medians \tilde{x} – den Datensatz zunächst nach aufsteigender Größe. Die Spannweite ergibt sich dann aus dem geordneten Datensatz $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ als Differenz aus dem größten Wert $x_{(n)}$ und dem kleinsten Wert $x_{(1)}$:

$$R := x_{(n)} - x_{(1)}. \quad (5.5)$$

Die Spannweite hat den Nachteil, dass sie eine hohe Empfindlichkeit bzw. eine geringe Robustheit gegenüber Ausreißern besitzt. Ändert man in einem Datensatz den maximalen oder den minimalen Wert stark, wirkt sich dies massiv auf den Wert von R aus.

Ein häufig verwendetes Maß für die Streuung eines Datensatzes ist die **empirische Varianz**

$$s^2 := \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2, \quad (5.6)$$

die im Kontext der beschreibenden Statistik oft auch nur **Varianz** genannt wird.² In die Varianz s^2 gehen die Abweichungen $x_i - \bar{x}$ der Merkmalswerte vom arithmetischen Mittel \bar{x} wegen (5.3) in quadrierter Form ein. Man findet daher für (5.6) auch die Bezeichnung **mittlere quadratische Abweichung**.

Bei der Berechnung der Varianz (5.6) kann die nachstehende Darstellung nützlich sein, bei der $\bar{x^2}$ das arithmetische Mittel der quadrierten Werte

²Zufallsvariablen werden in den Kapiteln 12 - 13 anhand von Modellen (Wahrscheinlichkeitsverteilungen) charakterisiert. Hier spricht man von *theoretischen Verteilungen*, deren Streuung durch eine *theoretische* Varianz beschrieben wird.

x_1^2, \dots, x_n^2 des Datensatzes bezeichnetnet: ³

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2. \quad (5.7)$$

Die Darstellung (5.7) geht aus (5.6) hervor, wenn man dort den quadrierten Term $(x_i - \bar{x})^2$ hinter dem Summenzeichen ausmultipliziert und die Summierung dann gliedweise vornimmt. Die Varianz s^2 ist ein *quadratisches* Streuungsmaß. Sind die Originaldaten z. B. Werte in *cm* oder in *sec*, so wird die Varianz in *cm*² bzw. in *sec*² gemessen.

Die Kenngröße (5.6) geht in ein *lineares* Streuungsmaß über, wenn man die Wurzel zieht. Man erhält so die **Standardabweichung** oder, genauer, die **empirische Standardabweichung**

$$s := \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\overline{x^2} - \bar{x}^2} \quad (5.8)$$

des Datensatzes. Diese wird in der Einheit ausgewiesen, in der die Ausgangsdaten gemessen werden. Die Standardabweichung ist daher im Vergleich zur Varianz ein anschaulicheres Streuungsmaß.

Uneinheitliche Bezeichnungen Die Bezeichnungen für Varianz und Standardabweichung sind in der Lehrbuchliteratur nicht immer einheitlich. Manchmal wird bei der Definition der Varianz in (5.6) anstelle des Bruchterms $\frac{1}{n}$ der Term $\frac{1}{n-1}$ verwendet. Das resultierende und hier mit s^{*2} abgekürzte Streuungsmaß

$$s^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \cdot s^2. \quad (5.9)$$

wird in diesem Manuskript **korrigierte Varianz** oder **Stichprobenvarianz** genannt. Nach Wurzelziehen resultiert die **korrigierte Standardabweichung** s^* , die auch als **Stichprobenstandardabweichung** angesprochen wird.

Die korrigierte Varianz wird beim Schätzen und Testen anstelle von (5.6) bevorzugt verwendet, weil sie – wie mit (15.8) und (15.9) noch gezeigt wird – günstigere Eigenschaften besitzt. Die Division durch $n - 1$ wird erst im Kontext der schließenden Statistik nachvollziehbar; sie lässt sich im Rahmen der beschreibenden Statistik nicht motivieren. Wichtig ist aber, dass man weiß, welche Formel der Berechnung zugrunde lag.

In diesem Manuskript werden die Bezeichnungen „Varianz“ und „Standardabweichung“ für Kenngrößen eines Datensatzes stets auf (5.6) bzw.

³Sind mehrere Merkmale im Spiel, etwa X und Y , so kann man zwischen den empirischen Varianzen und Standardabweichungen durch Verwendung tiefgestellter Indizes differenzieren, etwa s_x^2 und s_y^2 im Falle der Varianzen.

(5.8) bezogen und mit s^2 bzw s abgekürzt. Aus der Varianz s^2 kann man wegen $s^{*2} = \frac{n}{n-1} \cdot s^2$ leicht die korrigierte Varianz s^{*2} berechnen und umgekehrt. Die Unterschiede zwischen beiden Größen verschwinden mit zunehmendem n , können aber bei kleinem n ins Gewicht fallen.

Beispiel 5.4: Streuung bei Stromverbrauchsdaten

Geht man erneut vom Datensatz in Beispiel 5.2 zum Pro-Kopf-Strom-Verbrauch in den USA, Deutschland, Japan, China resp. Indien aus (mittlere Spalte in Tabelle 5.1), so ist dieser für die Berechnung der Spannweite R zunächst in die geordnete Folge $x_{(1)} = 0,86, x_{(2)} = 4,05, x_{(3)} = 7,01, x_{(4)} = 7,86, x_{(5)} = 12,83$ zu überführen. Es errechnet sich $R = 12,83 - 0,86 = 11,97$.

Bei der Berechnung der empirischen Varianz nach (5.6) werden die Originaldaten um den Mittelwert $\bar{x} = 6,522$ vermindert und die resultierenden Mittelwertabweichungen quadriert, aufsummiert und durch $n = 5$ dividiert. Man erhält so bei Rundung auf drei Nachkommastellen

$$s^2 = \frac{1}{5} \cdot [6,308^2 + 0,488^2 + 1,338^2 + (-2,472)^2 + (-5,662)^2] \approx 15,998.$$

Geht man alternativ von (5.7) aus, erhält man die etwas kürzere Rechnung

$$s^2 = \frac{1}{5} \cdot 292,6707 - 6,522^2 \approx 58,534 - 32,536 = 15,998.$$

Aus s^2 gewinnt man durch Wurzelziehen nach (5.8) die Standardabweichung s .

Die korrigierte empirische Varianz (5.9) errechnet sich als $s^{*2} = \frac{5}{4} \cdot s^2 \approx 19,997$. Der Unterschied zu $s^2 \approx 15,998$ ist deutlich, weil n hier klein ist.

Auch bei der Berechnung der Varianz kann man im Falle mehrfach auftretender Merkmalswerte auf relative Häufigkeiten zurückgreifen. Liegt für ein diskretes Merkmal X mit den Ausprägungen a_1, \dots, a_k eine größere Anzahl n von Beobachtungswerten x_1, \dots, x_n vor ($n > k$), so wären bei der Anwendung von (5.6) n Mittelwertabweichungen $x_i - \bar{x}$ zu quadrieren. Statt der Abweichungen $x_i - \bar{x}$ der Urwerte vom Mittelwert kann man alternativ die Abweichungen $a_i - \bar{x}$ der Merkmalsausprägungen vom Mittelwert heranziehen und deren Quadrate mit den Elementen f_i der relativen Häufigkeitsverteilung $f_1 = f(a_1), \dots, f_k = f(a_k)$ gewichten. Man erhält so für s^2 analog zu (5.4) die alternative Berechnungsformel

$$s^2 = \sum_{i=1}^k (a_i - \bar{x})^2 \cdot f_i, \quad (5.10)$$

bei der sich die Summenbildung auf nur k Terme bezieht. Hat man z. B. als Ergebnis eines Würfelexperiments mit 10 Würfen den Datensatz $\{1, 1, 1, 1, 4, 4, 5, 5, 5, 6\}$ erhalten, so errechnet man mit $\bar{x} = 3,3$ für die

Alternative
Berechnung der
Varianz



Aufgabe 5.1

Varianz s^2 sowohl nach (5.7) als auch nach (5.10) bei Rundung auf zwei Dezimalstellen den Wert $s^2 = 3,81$. Bei der Varianzberechnung im Falle *gruppierter Daten* sind die Ausprägungen a_i durch die Mitte m_i der Klassen zu ersetzen und die Häufigkeiten f_i entsprechen dann den relativen Besetzungshäufigkeiten der einzelnen Klassen.

Standardisierung von Datensätzen Will man Datensätze x_1, x_2, \dots, x_n vergleichen, die sich auf unterschiedliche Grundgesamtheiten beziehen oder mit unterschiedlichen Messinstrumenten gewonnen wurden, kann man von jedem Element den jeweiligen Mittelwert \bar{x} subtrahieren und die Differenz durch die Standardabweichung s oder die korrigierte Standardabweichung s^* dividieren. Es resultieren neue Datensätze y_1, y_2, \dots, y_n mit Mittelwert $\bar{y} = 0$ und Standardabweichung $s = 1$ resp. $s^* = 1$. Solche Transformationen sind z. B. sinnvoll, wenn man Intelligenzmessungen in unterschiedlichen Grundgesamtheiten durchführen oder schulische Leistungen anhand unterschiedlicher Fragebögen messen will. Die beschriebene Transformation wird in der *Psychologie* und in den *Sozialwissenschaften* auch **z-Transformation** genannt. Sie ist das empirische Analogon zu der in Abschnitt 13.2 dieses Manuskripts vorgestellten z-Transformation zur Standardisierung von Zufallsvariablen.

Ein weiteres Streuungsmaß Varianz s^2 und Standardabweichung s sind Streuungsmaße, die sich auf Abweichungen $x_i - \bar{x}$ vom *Mittelwert* eines Datensatzes beziehen. Ein alternatives Streuungsmaß ist die **mittlere absolute Abweichung vom Median**. Dieses oft mit d abgekürzte Maß basiert auf den Absolutbeträgen der Abweichungen $x_i - \tilde{x}$ vom Median:

$$d := \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}|.$$

Maßstabsinvariante Kenngrößen Unterzieht man die Daten x_i für ein quantitatives Merkmal einer Transformation $y_i = a + b \cdot x_i$, werden Median, Mittelwert und Standardabweichung in gleicher Weise transformiert, d. h. es gilt z. B. für den Mittelwert \bar{y} der transformierten Daten die Beziehung $\bar{y} = a + b \cdot \bar{x}$. Auf die Varianz und die Standardabweichung wirkt sich die Niveauverschiebung a nicht aus; nur der Wert von b ist relevant. Bezeichnet man die empirische Varianz des ursprünglichen Merkmals X mit s_x^2 und die des transformierten Merkmals Y mit s_y^2 , so gilt $s_y^2 = b^2 \cdot s_x^2$ und $s_y = |b| \cdot s_x$.

Medianen, Mittelwerte und Standardabweichungen von Datensätzen sind also vom Maßstab abhängig. Für quantitative Merkmale mit nicht-negativen Ausprägungen wird oft der **Variationskoeffizient**

$$v := \frac{s}{\bar{x}}$$

verwendet. Dieser ist ein *maßstabsunabhängiges* Streuungsmaß.

5.3 Quantile und Boxplots

Der für ein metrisch oder mindestens ordinalskaliertes Merkmal erklärte Median \tilde{x} hat die Eigenschaft, dass mindestens 50 % der nach Größe geordneten Elemente $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ eines Datensatzes kleiner oder gleich und mindestens 50 % größer oder gleich \tilde{x} sind. Bei den 5 Werten in der mittleren Spalte von Tabelle 5.1 war der Median z. B. durch $\tilde{x} = x_{(3)} = 7,01$ gegeben und je 3 der 5 Elemente in dieser Spalte, d. h. 60 % der Werte, waren kleiner oder gleich resp. größer oder gleich \tilde{x} . Bei ordinalskaliertem Merkmal ist \tilde{x} nicht immer eindeutig bestimmt. Bei metrischer Skalierung lässt sich über (5.1) eine eindeutige Festlegung erreichen.

Eine Verallgemeinerung des Medians ist das **p-Quantil**. Auch dieses setzt wieder ein metrisch oder zumindest ordinalskaliertes Merkmal voraus. Ein p -Quantil wird mit x_p abgekürzt und hat die Eigenschaft, dass mindestens $p \cdot 100\%$ der Elemente der geordneten Folge $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ kleiner oder gleich und mindestens $(1 - p) \cdot 100\%$ größer oder gleich x_p sind.⁴ Abbildung 5.1 veranschaulicht diese Definition.

Verallgemeinerung
des Medians

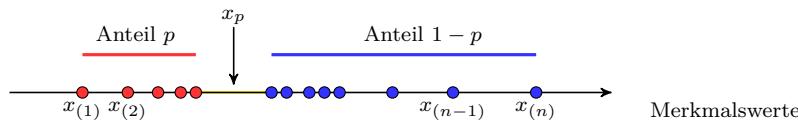


Abb. 5.1: Veranschaulichung des p -Quantils

Auch das p -Quantil ist bei einem ordinalskalierten Merkmal i. d. R. nicht eindeutig bestimmt. Bei metrischer Merkmalsskalierung kann Eindeutigkeit erreicht werden, wenn zur Berechnung das arithmetische Mittel der zwei Merkmalsausprägungen herangezogen wird, zwischen denen das p -Quantil liegt. Bezeichne $[np]$ die größte ganze Zahl, die kleiner oder gleich np ist. Es ist dann $[np] + 1$ die kleinste ganze Zahl, die größer als np ist.⁵ Mit dieser Notation kann x_p bei einem metrisch skalierten Merkmal in Verallgemeinerung von (5.1) eindeutig definiert werden durch

$$x_p = \begin{cases} x_{([np]+1)} & \text{falls } np \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{(np)} + x_{(np+1)}) & \text{falls } np \text{ ganzzahlig.} \end{cases} \quad (5.11)$$

Der Median ist demnach ein spezielles Quantil, nämlich das 0,5-Quantil. Weitere wichtige Quantile sind das 0,25-Quantil und das 0,75-Quantil,

Spezielle Quantile

⁴Die Notation für Quantile ist in der Literatur nicht einheitlich. Man findet auch die Schreibweise \tilde{x}_p anstelle von x_p ; vgl. z. B. STELAND (2016, Abschnitt 1.6.4).

⁵Die auf Carl Friedrich GAUSS zurückgehende Funktion $f(x) = [x]$ wird *Gauß-Klammer-Funktion* oder *Abrundungsfunktion* genannt. Sie ist eine für alle reellen Zahlen erklärte Treppenfunktion mit Sprungstellen bei jeder ganzen Zahl (Sprunghöhe 1). Es ist z. B. $[3,8] = 3$.

die **unteres Quartil** resp. **oberes Quartil** genannt werden. Abbildung 5.2 visualisiert diese drei Spezialfälle.

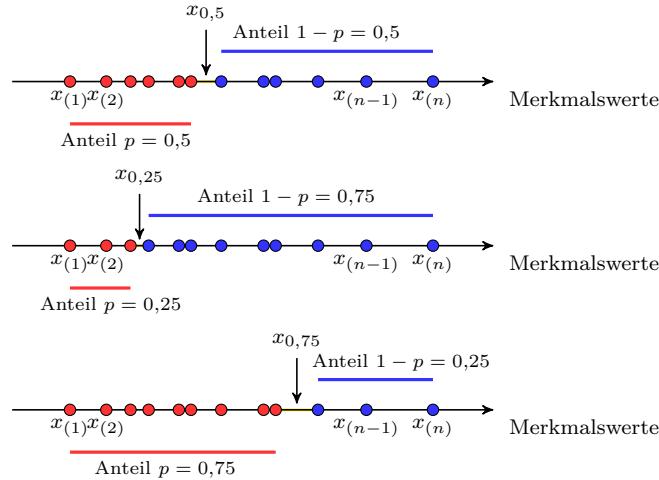


Abb. 5.2: Median $x_{0,5}$, unteres Quartil $x_{0,25}$ und oberes Quartil $x_{0,75}$

Die häufig mit Q abgekürzte Differenz der Quartile $x_{0,75}$ und $x_{0,25}$, also

$$Q := x_{0,75} - x_{0,25}, \quad (5.12)$$

wird **Quartilsabstand** genannt. Sie wird auch als **Interquartilsabstand IQR** angesprochen (engl: *interquartile range*). Ferner sind noch die **Dezile** zu nennen, die sich bei Wahl von $p = 0,1, p = 0,2, \dots, p = 0,9$ ergeben und oft mit $D1, D2, \dots, D9$ abgekürzt werden. Der Median $\tilde{x} = x_{0,5}$ stimmt also mit dem Dezil $D5$ überein.

In Abbildung 4.8 waren für spanische und portugiesische Arbeitnehmer Bruttojahresverdienste in Form von Histogrammen visualisiert, wobei über den Histogrammen jeweils die aus den Originaldaten (ungruppierte Daten) errechneten Dezile $D1$ und $D9$ sowie der Median $D5 = \tilde{x}$ und der Mittelwert \bar{x} wiedergegeben war. Das ebenfalls ausgewiesene Verhältnis $\frac{D9}{D1}$ der extremen Dezile liefert eine Information über den Grad der Ungleichheit der Verdienste in der betrachteten Grundgesamtheit von Arbeitnehmern – hohe Werte des Quotienten sprechen für eine ausgeprägte Ungleichheit. Man erkennt schon anhand der Grafiken, dass sich der überwiegende Teil der in Abbildung 4.8 veranschaulichten Verdienste in den unteren Einkommensbereichen bewegen, d. h. der überwiegende Teil der Daten ist linksseitig konzentriert – hier sind höhere Klassenbesetzungshäufigkeiten und ein steilerer Abfall der Verteilung zu beobachten. Man spricht dann von einer **linkssteilen** oder **rechtsschiefen Verteilung**. Eine **rechtssteile** oder **linksschiefe Verteilung** würde hingegen an der rechten Flanke steiler abfallen. In beiden Fällen liegt eine **asym-**

Wie erkennt man
eine asymmetrische
Verteilung?

metrische Verteilung vor. Die Nicht-Übereinstimmung von Median und Mittelwert einer empirischen Verteilung ist stets ein Indiz für eine Asymmetrie dieser Verteilung.

Ein aussagekräftiges grafisches Instrument zur Beurteilung einer empirischen Verteilung (Zentrum, Streuung, Asymmetrie) ist der sog. **Boxplot** („Schachtelzeichnung“). Dieser fasst in seiner einfachsten Form fünf Charakteristika eines Datensatzes zusammen, nämlich die beiden Extremwerte $x_{\min} = x_{(1)}$ und $x_{\max} = x_{(n)}$, die beiden Quartile $x_{0,25}$ und $x_{0,75}$ sowie den Median $x_{0,5}$. Die beiden Quartile definieren die Länge einer Box („Schachtel“). Innerhalb der Box ist der Median in Form eines Strichs oder Punktes eingezeichnet. Die Box wird mit den Extremwerten durch Linien verbunden (sog. „whiskers“, übersetzt: Schnurrhaare), deren Ende durch einen Strich markiert wird. Die Länge der Box entspricht also dem Quartilsabstand Q . Innerhalb der Box liegen etwa 50 % der Daten, unterhalb und oberhalb der Box jeweils ca. 25 %. Der Median liefert eine Information zum Zentrum des Datensatzes.

Basisvariante eines Boxplots



Aufgaben 5.2-3

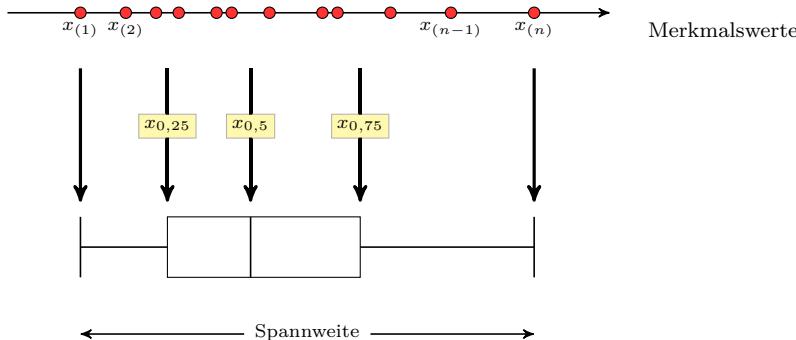


Abb. 5.3: Aufbau eines Boxplots (Basisversion)

Manchmal wird neben dem Median auch noch der Mittelwert innerhalb der Box dargestellt. Bei einer symmetrischen Verteilung liegt der Median genau in der Mitte der Box.

Abbildung 5.3 zeigt nur die einfachste Boxplot-Variante. Häufig wird eine andere, hier nur der Vollständigkeit halber erwähnte Version mit gleichem Aufbau der Box, aber anderer Begrenzung der an der Box angebrachten Linien verwendet. Statt die Linien stets genau bis zu den Extremwerten zu führen, kann man auch so verfahren, dass man die Linien nur dann bis zu den Extremwerten zeichnet, wenn deren Abstand zur Box nicht größer ist als das 1,5-fache des Quartilsabstands Q . Die an der Box angesetzten Linien werden andernfalls auf die Länge $1,5 \cdot Q$ begrenzt und weiter entfernt liegende Werte separat eingezeichnet. So lassen sich auffällige Datenpunkte („Ausreißer“) hervorheben.

Boxplots mit Visualisierung von Ausreißern

Beispiel 5.5: Boxplots zu relativen Militärausgaben von Ländern



Interaktives Objekt
„Militärausgaben“

Abbildung 4.1 verglich anhand eines Balkendiagramms, welchen Anteil ihres Bruttoinlandsprodukts 24 ausgewählte Länder im Jahr 2019 für Militärausgaben verwendeten. Die Prozentwerte für 2019 oder ein anderes Referenzjahr lassen sich zu je einem Boxplot aggregieren, der die 24 länderspezifischen Anteilswerte zu den in Abbildung 5.3 wiedergegebenen 5 Charakteristika aggregiert. In der nachstehenden Grafik sind Boxplots für die Jahre 2003 – 2019 eingezeichnet. Über jedem Boxplot ist der Code desjenigen Landes angegeben, das im jeweiligen Jahr den höchsten Prozentsatz seines BIP für militärische Zwecke ausgab. In dem hier betrachteten Zeitraum waren dies stets Oman (OM) oder Saudi-Arabien (SA).



Abb. 5.4: Militärausgaben von 24 Ländern in % des BIP für den Zeitraum 2003 – 2019 (Boxplots). Quelle: ;Datenextraktion: April 2020

Man erkennt, dass der Median $x_{0,5}$ der Militärausgaben der ausgewählten Länder in der gesamten Periode bei etwas über 2 % des BIP liegt. Auch die Grenzen der Boxen, d. h. die Quartile $x_{0,25}$ und $x_{0,75}$ der Jahresdatensätze, variieren nicht übermäßig. Die Maxima variieren hingegen erheblich und unterscheiden sich auffällig von den Werten innerhalb der Boxen.

Die Entwicklung der Militärausgaben im Zeitraum 2003 – 2019 für die 24 Staaten hätte man auch anhand von 24 Tabellen oder Zeitreihengraphen wiedergeben können. Abbildung 5.4 aggregiert die in diesen Tabellen bzw. Graphen enthaltende Information. So werden zentrale „Botschaften“ und Auffälligkeiten sichtbar, die sich aus langen Zahlenkolonnen („Zahlenfriedhöfen“) nicht leicht erschließen lassen.

Im Finanzsektor werden Kursschwankungen von Aktien oder Währungen während eines Börsentages häufig anhand von **Kerzendiagrammen** (engl.: *candlestick charts*) visualisiert. Diese sind den Boxplotdiagrammen des in Abbildung 5.4 wiedergegebenen Typs sehr ähnlich. Während ein Boxplot in der Basisvariante einen Datensatz auf 5 Charakteristika verdichtet (Extremwerte, unteres und oberes Quartil, Median), entfällt bei den Kerzendiagrammen die Ausweisung des Medians. Die quergestellten Striche, die bei den Boxplots das Niveau der Extremwerte betonen, werden ebenfalls weggelassen. Außerdem wird die Box nicht durch das untere und obere Quartil definiert, sondern durch die Kurswerte bei Eröffnung und bei Schließung der Börse. Die Differenz der beiden Werte definiert somit die Länge der Box. Da die Schlusswerte sowohl über als auch unter den Eröffnungskursen liegen können, werden die beiden Fälle oft durch eine andere Färbung der Boxen unterschieden - meist rote Färbung im Falle eines Börsentags, bei denen der Schlusskurs unter dem Eröffnungskurs lag, und grüne bei beobachteten Tagesgewinnen.

Modifikation von
Boxplots im
Finanzsektor

Exkurs 5.2: Kursschwankungen bei Kryptowährungen

Die Schwankungsintensität (Volatilität) von Kursen innerhalb eines Börsentages lässt sich mit Kerzendiagrammen gut veranschaulichen. Gerade bei den seit 2017 stark nachgefragten Kryptowährungen können diese Schwankungen extrem sein. Anfang 2018 gab es fast 1 500, Anfang 2020 bereits über 2 000 Kryptowährungen. Diese sind entweder als allgemein verwendbares digitales Zahlungsmittel konzipiert sind oder als Zahlungsmittel für spezielle Dienstleistungen. Aus der erstgenannten Kategorie sind Bitcoin, Dash („*Digital Cash*“) und NEM („*New Economic Movement*“) zu nennen, aus der zweiten Ether oder Ripple.

Zu Beginn des Jahres 2018 waren die Kryptowährungen Bitcoin, Ether und Ripple diejenigen mit der höchsten Marktkapitalisierung. Da der undurchsichtige Markt mit exorbitanten Gewinnen lockt, zog er ab 2017 auch unerfahrene Kleinanleger an. Diesen ist oft nicht bewusst, dass nicht nur die Gewinnchancen sehr hoch sind, sondern auch die Verlustrisiken. Wie stark die Kursschwankungen bei Kryptowährungen schon innerhalb eines einzigen Tages sein können, veranschaulicht die mit *R* erstellte Abbildung 5.5 anhand von Kerzendiagrammen für Bitcoin und Ether. Die Grafik bezieht sich auf die ersten 6 Wochen des Jahres 2018. Börsentage mit Kursverlusten sind anhand der rot eingefärbten Boxen zu erkennen. Besonders hoch war bei beiden Währungen der Tagesverlust am 16. Januar 2018. Der Schlusskurs lag bei Bitcoin (BTC) an diesem Tag etwa 23,7%, bei Ether (ETH) 22,5% unter dem Eröffnungskurs.



Abb. 5.5: Kerzendiagramme für den Kurs der Kryptowährungen Bitcoin und Ether im Zeitraum vom 2. Januar 2018 - 13. Februar 2018.
Datenquelle: ARIVA.

Als weitere Risiken sind Sicherheitsprobleme zu nennen, die mit der Speicherung von virtuellem Geld verbunden sind und Totalverluste nach sich ziehen können. Die Tokioter Online-Handelsplattform *Mt. Gox*, über die jahrelang weit mehr als die Hälfte des weltweiten Bitcoin-Handels abgewickelt wurde, musste Anfang 2014 nach einem Hackerangriff Insolvenz anmelden. Bei dem Angriff waren 650 000 Bitcoins gestohlen worden, die zum Zeitpunkt der Attacke einen Wert von ca. 390 Millionen Euro hatten. Es gab in der Folgezeit noch weitere spektakuläre Fälle. So wurde die Kryptowährung NEM, wie die *Wirtschaftswoche* berichtete, in der zweiten Januarhälfte 2018 von einem Hackerangriff betroffen, bei dem von der Tokioter Online-Börsenplattform *Coincheck* NEM-Einheiten im Wert von ca. 430 Millionen Euro gestohlen wurden.

6 Analyse von Ereignisdaten



Vorschau auf
das Kapitel

Bei der Auswertung von Ereignisdaten geht es um die Dauer, in denen Objekte einer Beobachtungsstudie verweilen, bis sie von einem Ausgangszustand in einen anderen Zustand übergehen. Es werden zunächst unterschiedliche Anwendungsfelder der Ereignisdatenanalyse aufgelistet, der Schwerpunkt dann aber auf den Bereich der Medizin gelegt. Typisch für Ereignisdaten ist, dass die Verweildauer der untersuchten Objekte nicht bei allen Objekten bekannt ist. Dieser Fall unvollständiger Information tritt ein, wenn Objekte während der Studie aus der Untersuchung ausscheiden und nicht weiter beobachtbar sind oder bis zum Ende des Beobachtungszeitraums keinen Zustandswechsel erfahren haben. Bei Vorliegen solcher, als zensiert bezeichneten Beobachtungen kann z. B. der Median des Merkmals „Verweildauer“ nicht mehr in klassischer Weise bestimmt werden.

Ein Ansatz, der die Bestimmung mittlerer Verweildauern bei Ereignisdaten ermöglicht, ist das Kaplan-Meier-Verfahren. Dieses wird im zweiten Teil des Kapitels beschrieben und anhand eines Fallbeispiels aus der Medizin illustriert. Dabei wird auch verdeutlicht, dass die bei Anwendung des Verfahrens errechneten mittleren Verweildauern stark von den zeitlichen Abständen zwischen den Beobachtungspunkten abhängen.

6.1 Anwendungsfelder und Grundbegriffe

Ereignisdaten beinhalten Informationen über die Länge von Zeitintervallen zwischen dem Wechsel von Zuständen, die eine Menge von Untersuchungsobjekten annehmen kann. Da die Abstände zwischen den Beobachtungszeitpunkten bei Ereignisdaten im Vergleich zu Paneldaten vergleichsweise kurz sind, beinhalten Ereignisdaten Informationen zu den Verweildauern der betrachteten Objekte im Ausgangszustand, die über Paneldaten nicht vermittelt werden. In diesem Manuskript wird zwischen nur zwei Zuständen unterschieden und nur der Fall betrachtet, dass der Übergang in den interessierenden zweiten Zustand nicht umkehrbar ist.

Hier einige Beispiele aus unterschiedlichen Bereichen:

- *Medizin*: Bei Krebspatienten kann innerhalb eines Beobachtungszeitraums die Dauer der Zustände „keine Progression“ (kein Fortschreiten der Erkrankung / Heilung) und „Progression“ (Ereignis: Verschlechterung / Tod) festgehalten werden;
- *Sozialwissenschaften*: Bei kinderlosen Frauen kann in einer Beobachtungsperiode verfolgt werden, ob und wann der Zustand „kinderlos“ verlassen wird (Ereignis: Geburt eines Kindes).

Wo fallen
Ereignisdaten an?

- *Soziologie; Ökonometrie*: Bei erwerbstätigen Personen ist die Dauer der Erwerbstätigkeit von Interesse (Ereignis: Ausscheiden aus dem Erwerbsleben) oder die Dauer vom Eintritt in ein Unternehmen bis zum Wechsel zu einem anderen Arbeitgeber (Ereignis: Wechsel des Arbeitgebers).
- *Ingenieurwissenschaften*: Bei industriell gefertigten Serienprodukten werden Daten zu deren Lebensdauer erhoben (Ereignis: Übergang vom Zustand „gebrauchstauglich“ in den Zustand „defekt“).

Bereichsspezifische Terminologien	Die Auswertung von Ereignisdaten wird in einzelnen Anwendungsfeldern mit unterschiedlichen Bezeichnungen belegt. In der Medizin spricht man Überlebenszeitanalyse (engl.: <i>survival analysis</i>), in den Sozialwissenschaften von Verweildaueranalyse (engl.: <i>duration analysis</i>) und in den Ingenieurwissenschaften von Zuverlässigkeitssanalyse (engl.: <i>reliability analysis</i>). Hier werde der neutrale Begriff „Ereigniszeitanalyse“ verwendet und als „Überlebenszeit“ sei die Zeitspanne bis zum Eintritt des interessierenden Ereignisses gemeint. Letztere kann den Todeszeitpunkt eines Patienten beinhalten, den Wirkungszeitpunkt eines Medikaments oder den Ausfallzeitpunkt einer Maschine. Innerhalb der Beobachtungsperiode ist für jedes an der Untersuchung beteiligte Objekt, für das das Ereignis noch nicht eingetreten ist, ein Zustandswechsel möglich. Man sagt auch, dass es „unter Risiko“ steht.
Ereignisdaten in der Medizin	In der Medizin, auf die sich die folgenden Ausführungen konzentrieren, sind bei der Auswertung von Ereignisdaten folgende Fragestellungen typisch: <ul style="list-style-type: none"> - Wie groß ist der Anteil der Personen aus einer Patientengruppe, bei denen zu einem bestimmten Zeitpunkt das interessierende Ereignis noch nicht eingetreten ist? - Wie lange dauert es bei einer einem bestimmten Therapieansatz unterzogenen Patientengruppe im Mittel, bis das interessierende Ereignis eintritt? Wie lange dauert es im Mittel bei Anwendung eines alternativen Therapiekonzepts?

Bei der Berechnung von mittleren Überlebenszeiten tritt das Problem auf, dass es Personen geben kann, bei denen die Dauer bis zum Eintritt des interessierenden Ereignisses innerhalb des Untersuchungszeitraums nicht festgestellt werden kann. Dies ist bei Personen der Fall, bei denen das Ereignis erst nach dem Ende eines Beobachtungszeitraums eintritt. Man spricht in diesem Falle von **zensierten Daten**. Zensierte Daten fallen auch an, wenn Personen während einer Studie aus unbekannten Gründen oder auf eigenen Wunsch ausscheiden und ihr Zustand nicht mehr beobachtbar ist.

Abbildung 6.1 zeigt nach Größe geordnete fiktive Überlebenszeiten für 7 Krebspatientinnen in Form eines Stabdiagramms mit horizontal ausgerichteten Stäben. Für die Frauen wurden ab Beginn einer Therapie über 2 Jahre allmonatlich Daten darüber erhoben, ob die Erkrankung fortschreitet. Das interessierende Ereignis ist hier „Progression“, was sowohl eine deutliche Zustandsverschlechterung als auch Tod meinen kann. „Überleben“ beinhaltet demnach nur, dass die Erkrankung nicht fortgeschritten ist. Der Zeitpunkt $t_0 = 0$ markiert den Anfangspunkt des 2-jährigen Beobachtungszeitraums der Studie. Die Stäbe enden jeweils zu den Zeitpunkten t_i , an denen entweder der Eintritt des Ereignisses „Progression“ entdeckt wird oder eine Zensierung erfolgt. Bei zensierten Daten ist der Zensierungszeitpunkt t_i durch einen kleinen vertikalen Strich markiert. Das erste Ereignis wird bei $t_2 = 9$ beobachtet und betrifft die Patientin mit der Identifikationsnummer 2. Das Ereignis „Progression“ kann hier natürlich schon zwischen den Beobachtungszeitpunkten t_1 und t_2 eingetreten sein.

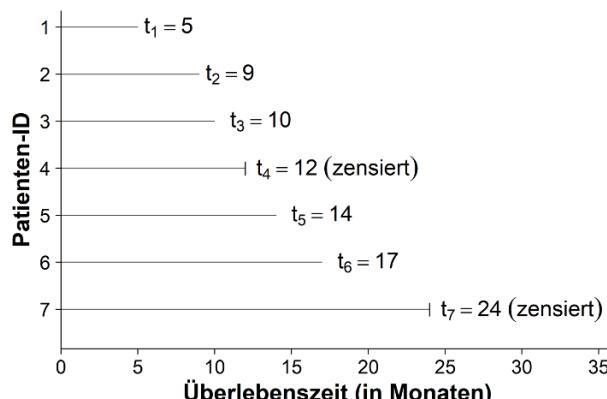


Abb. 6.1: Überlebenszeitverläufe mit censierten und uncensierten Daten

Die Abbildung weist insgesamt 5 uncensierte Beobachtungen aus. Die Patientin mit der Nummer 4 schied aus der Studie nach 12 Monaten auf eigenen Wunsch aus; ihr Zustand war danach nicht mehr beobachtbar. Von ihr weiß man nur, dass ihre Überlebenszeit mindestens 12 Monate betrug. Die sie betreffende Beobachtung ist censiert. Dies gilt auch für die Patientin mit der ID-Nummer 7, bei der am Ende der Studie noch keine Progression aufgetreten war. Auch hier ist nur eine untere Schranke für die Überlebenszeit bekannt (24 Monate).

Zur Bestimmung mittlerer Überlebenszeiten verwendet man den Median. Die klassische Bestimmung des Medians nach (5.1) ist nicht anwendbar. Beim dem in Abbildung 6.1 wiedergegebenen Beispiel wäre der Median z. B. durch die Überlebenszeit der Patientin mit der ID 4 gegeben, wenn das Ereignis bei ihr vor dem Zeitpunkt t_5 eintrüte. Ob dies zutrifft, ist aber

nicht bekannt. Zöge man nur die innerhalb des Untersuchungszeitraums beobachteten Ereignisse bei der Bestimmung des Medians oder anderer Quantile der beobachteten Überlebenszeiten heran, würde die Kenntnis der unteren Schranken für die zensierten Daten nicht verwertet. Man benötigt daher für die Ereignisdatenanalyse spezielle Auswertungsverfahren. Dabei kann man mit den empirischen Befunden arbeiten und auf eine Modellierung der Überlebenszeiten als Ausprägungen von Zufallsvariablen mit bestimmten Verteilungen (Lebensdauerverteilungen) verzichten. Das im Folgenden vorgestellte Kaplan-Meier-Verfahren repräsentiert einen solchen verteilungsfreien Ansatz.



Ereignisdaueranalysen auf der Basis von Lebensdauerverteilungen werden bei TOUTENBURG / HEUMANN (2008, Kapitel 12) ausführlicher behandelt. Erwähnt sei, dass bei der Auswertung von Ereignisdaten auch die im noch folgenden Abschnitt 17.6 vorgestellten binären Reponsemodelle herangezogen werden können. Dabei bleibt allerdings Information zu den Verweildauern im Ausgangszustand unberücksichtigt.

6.2 Das Kaplan-Meier-Verfahren



Edward Lynn
KAPLAN

Ein Ansatz, der es erlaubt, auch zensierte Überlebenszeiten bei der Auswertung von Ereignisdaten zu berücksichtigen, wurde 1958 von dem US-amerikanischen Mathematiker Edward Lynn KAPLAN (1920 – 2006) und seinem Landsmann und Statistiker Paul MEIER (1920 – 2006) präsentiert. Das von Kaplan und Meier vorgeschlagene Verfahren verwendet eine absteigende Treppenfunktion, deren Graph **Kaplan-Meier-Kurve** genannt wird. Sie ermöglicht es, Überlebenszeiten für Personen oder – allgemeiner – für Objekte, die an einer Untersuchung beteiligt sind, zu visualisieren und den Anteil der Objekte abzulesen, für die das interessierende Ereignis zum Zeitpunkt t noch *nicht* aufgetreten ist. Dieser Anteil hat den Zeitpunkt t in dem anfangs definierten Sinne „überlebt“, steht zum Zeitpunkt t demnach noch „unter Risiko“. Die **Überlebensfunktionsfunktion $S(t)$** (engl: *survival function*) weist den Wert des genannten Anteils in Abhängigkeit von t aus.

Da die beobachteten Ereigniszepunkte nicht exakt die Zeitpunkte wiedergeben, an denen die Ereignisse tatsächlich eintreten (verzögerte Erfassung aufgrund von Abständen zwischen den Messungen), hat man in der Praxis nur eine Schätzung $\hat{S}(t)$ der Überlebenszeitfunktion. Die Schätzung trifft die „wahre“ Funktion $S(t)$ um so besser, je enger die Messzeitpunkte zusammenliegen.

Die Kaplan-Meier-Kurve ist der Graph von $\hat{S}(t)$. Die Abszissenachse repräsentiert die Zeit, die Ordinatenachse den Anteil der an der Untersuchung beteiligten Objekte unter Risiko, d. h. die **Überlebensrate**. Die

Ordinatenwerte werden oft als Überlebenswahrscheinlichkeiten aufgefasst, obwohl das Verfahren lediglich deskriptive Häufigkeitsaussagen liefert. Der Startzeitpunkt t_0 muss nicht für alle Untersuchungsobjekte identisch sein. In der Medizin können Personen z. B. zu unterschiedlichen Zeitpunkten in eine Studie aufgenommen werden (z. B. nach Diagnose einer bestimmten Krankheit oder nach erstmaliger Einnahme eines Medikaments). In diesem Fall ist t_0 für diese Personen durch den Zeitpunkt der Aufnahme in die Untersuchung definiert.

Vom Startzeitpunkt t_0 der Untersuchung bis zum Zeitpunkt t_1 , an dem erstmalig das interessierende Ereignis beobachtet wird, sind n_1 Objekte beteiligt und alle stehen zunächst unter Risiko. Die Treppenkurve verläuft in diesem Zeitintervall auf dem Niveau 1, weil die Quote der Objekte unter Risiko noch bei 1 liegt. Tritt das Ereignis in t_1 bei genau einem Objekt ein, reduziert sich die Anzahl der unter Risiko stehenden Objekte ab t_1 um 1 auf $n_2 = n_1 - 1$ und das Niveau der Treppenfunktion fällt von 1 auf den Wert $1 - \frac{1}{n_1}$. Beobachtet man in t_2 ein weiteres Ereignis bei einem Objekt, fällt der Graph auf das Niveau $(1 - \frac{1}{n_1}) \cdot (1 - \frac{1}{n_2})$ und es stehen nur noch $n_3 = n_2 - 1$ Objekte unter Risiko, usw. Wenn zu einem Zeitpunkt das interessierende Ereignis bei mehreren Objekten gleichzeitig beobachtet wird, vergrößert sich der Sprung nach unten entsprechend. Tritt zu einem Zeitpunkt bei einem Objekt eine Zensierung auf, vermindert sich die Anzahl der Objekte unter Risiko um 1, weil das Objekt aus dem Untersuchungsrahmen herausfällt. Das Niveau der Treppenfunktion bleibt aber unverändert, weil es kein beobachtetes Ereignis gab. Zensierte Beobachtungen werden durch einen kurzen, senkrechten Strich auf der Kurve markiert.

Konstruktion der Kaplan-Meier-Kurve

Die Anzahl der Sprungstellen der Kaplan-Meier-Kurve entspricht der Anzahl der Beobachtungspunkte, an denen mindestens ein Ereignis registriert wird. Man erhält für die Treppenfunktion $\hat{S}(t)$ eine allgemeine Formel, wenn man einige Abkürzungen einführt. Es bezeichne

- t_i die Zeitpunkte, an denen ein Ereignis / eine Zensierung eintritt;
- d_i die Anzahl der zum Zeitpunkt t_i beobachtbaren Ereignisse;
- n_i die bis zum Zeitpunkt t_i unter Risiko stehenden Objekte.

Mit dieser Notation gilt

$$\hat{S}(t) = \begin{cases} 1 & \text{für } t \in [0; t_1), \\ 1 - \frac{d_1}{n_1} & \text{für } t \in [t_1; t_2), \\ (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2}) & \text{für } t \in [t_2; t_3), \\ (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})(1 - \frac{d_3}{n_3}) & \text{für } t \in [t_3; t_4), \\ \dots & \end{cases} \quad (6.1)$$

Die Formel (6.1) kann man unter Verwendung des Produktoperators kürzer schreiben:¹

$$\hat{S}(t) = \begin{cases} 1 & \text{für } t \in [0; t_1), \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) & \text{für } t \geq t_1. \end{cases} \quad (6.2)$$

Die Werte sind Schätzungen für den Anteil der Untersuchungsobjekte, die den Zeitpunkt t überleben. Die Schätzwerte lassen sich aus der Kaplan-Meier-Kurve direkt ablesen. Zu jedem Wert t gehört nach (6.1) ein eindeutig bestimmter Wert $\hat{S}(t)$. Die Kurve, die eigentlich an den Sprungstellen Unterbrechungen aufweisen müsste, wird nur deswegen durchgehend gezeichnet, weil dies die ebenfalls mögliche Ablesung von Überlebenszeiten t erleichtert, die zu einer geschätzten Überlebensrate $\hat{S}(t)$ gehören. Die letztgenannte Ableserichtung führt nicht zwingend zu einem eindeutig bestimmten Wert, wie man den noch folgenden Abbildungen 6.3 und 6.4 entnimmt.

Beispiel 6.1: Fallbeispiel zur Kaplan-Meier-Kurve

Das folgende Beispiel illustriert anhand der in Abbildung 6.1 dargestellten Daten, wie die Treppenfunktion $\hat{S}(t)$ zu berechnen ist. Die Funktion ist für den 24 Monate umfassenden Beobachtungszeitraum erklärt und besitzt in diesem Zeitrahmen 5 Sprünge. Diese liegen in t_1, t_2, t_3, t_5 und t_6 , weil nur hier jeweils das interessierende Ereignis „Progression“ (Verschlechterung / Tod) beobachtet wurde. Aus (6.1) leitet man ab, dass die Kaplan-Meier-Kurve



Aufgaben 6.1-2

- in t_1 das Niveau 1 verlässt und danach bis Erreichen von t_2 den Wert $1 - \frac{1}{7} = \frac{6}{7}$ hält;
- in t_2 von $\frac{6}{7}$ auf den Wert $\frac{5}{7}$ abfällt und diesen bis t_3 beibehält;
- in t_3 auf $\frac{4}{7}$ geht und dort bis t_5 bleibt;
- nach Erreichen von t_5 von $\frac{4}{7}$ auf das Niveau $\frac{8}{21}$ springt und hier bis einschließlich t_6 bleibt;
- nach Erreichen von t_6 das Endniveau $\frac{4}{21}$ annimmt.

Abbildung 6.2 zeigt die zugehörige Kaplan-Meier-Kurve. Aus dieser lässt sich insbesondere der Median der geschätzten Überlebenszeit ablesen. Zur Ablesung wird, wie in der Grafik realisiert, von der Ordinatenachse parallel zur Abszissenachse eine gepunktete Linie auf dem Niveau 0,5 gezogen (Überlebensrate 50 %). Von der Stelle, an der sie auf die Kaplan-Meier-Kurve trifft, führt man die gepunktete Linie nach unten in Richtung der Zeitachse weiter. Der Median

¹Der Produktoperator \prod und andere mathematische Symbole sind in Tabelle ?? erklärt. Die Angabe unter dem Produktzeichen in (6.2) bedeutet, dass im hinter dem Operator stehenden Term nacheinander alle „i“ eingesetzt werden, für die $t_i \leq t$ gilt. Die resultierenden Klammerterme sind dann miteinander zu multiplizieren.

liegt bei 14 Monaten. Das ist der Zeitpunkt t , bis zu dem die Hälfte (50 %) der Patientinnen progredient oder verstorben ist.

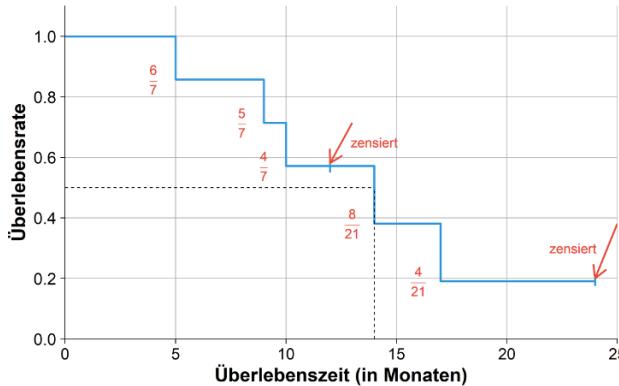


Abb. 6.2: Kaplan-Meier-Kurve mit censierten und unzensierten Daten

In der Medizin wird kaum beachtet, dass der Wert der an einer Kaplan-Meier-Kurve zu jedem Zeitpunkt t ablesbaren Schätzung $\hat{S}(t)$ der Überlebensrate erheblich davon abhängt, in welchen zeitlichen Abständen die Daten erhoben werden. Ist das interessierende Ereignis – anders als etwa beim Ereignis „Tod“ – nicht unmittelbar beobachtbar, sondern muss erst durch ein bildgebendes oder ein anderes aufwändiges Diagnoseverfahren bestätigt werden, kann der Beobachtungs- und Diagnosezeitpunkt erheblich vom Zeitpunkt des Ereignisses abweichen. In klinischen Studien, die das progressionsfreie Überleben von Krebspatienten verfolgen, werden bildgebende Verfahren wie die Magnetresonanztomographie (MRT) oder die Computertomographie (CT) schon aus Kostengründen nicht kontinuierlich, sondern nur in größeren Abständen eingesetzt. Die Abstände zwischen den Befunderhebungen werden mittels der sogenannten Scan-Frequenz angegeben.



Kritisch
nachgefragt

Die folgenden Abbildungen verdeutlichen, wie stark sich eine zunehmende Zeitspanne zwischen den Untersuchungen auf den Schätzwert für den Median der Überlebensraten auswirkt. In Abbildung 6.3 sind zwei Kaplan-Meier-Kurven dargestellt, die sich auf simulierte unzensierte Überlebensdaten beziehen. Die blaue glattere Kurve wurde bei Annahme einer maximalen Scan-Frequenz gewonnen (tägliche Beobachtung), während der rot dargestellten Treppenfunktion eine 4-wöchige Scan-Frequenz zugrunde liegt. Für den Median des Merkmals „Überlebenszeit“ liest man im ersten Fall 4 Monate als Schätzwert ab, im zweiten Fall hingegen 5 Monate.

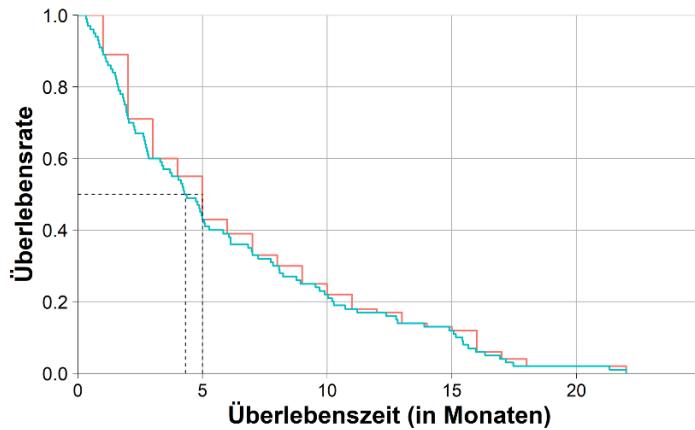


Abb. 6.3: Kaplan-Meier-Kurven bei maximaler und bei 4-wöchiger Scan-Frequenz

Eine erneute Verringerung der Scan-Frequenz auf 8 Wochen erhöht den Schätzwert für den Median sogar auf 6 Monate. Im Vergleich mit der „Wahrheit“, für die hier die kontinuierliche Beobachtung steht, wird die mittlere Überlebenszeit also um 50 % (!) überschätzt. Das Problem resultiert nicht aus der Methodik der Datenanalyse, sondern aus der systematischen Verzerrung der Daten durch die Methodik der Datenerhebung. Dies zeigt, wie wichtig es ist, Informationen zur Herkunft der Daten und der Art ihrer Gewinnung zu haben, wenn man die Belastbarkeit der aus den Daten gezogenen Schlüsse bewerten will.

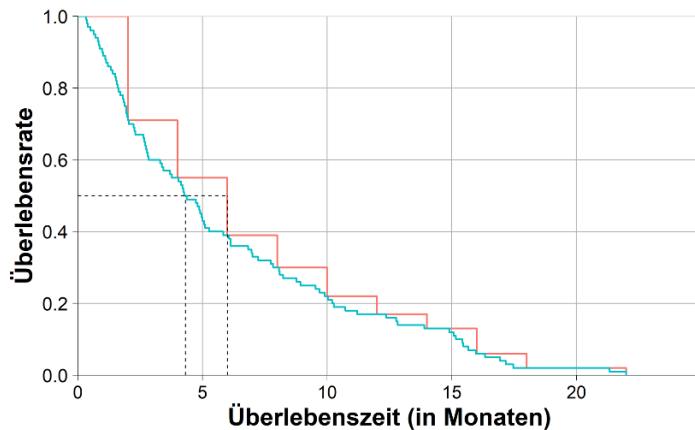


Abb. 6.4: Kaplan-Meier-Kurven bei maximaler und bei 8-wöchiger Scan-Frequenz



7 Konzentration von Merkmalswerten



Vorschau auf
das Kapitel

Bei metrisch skalierten Merkmalen mit nicht-negativen Ausprägungen – etwa Einkommen von Arbeitnehmern oder Marktanteilen von Unternehmen – ist oft von Interesse, wie sich die Summe aller Merkmalswerte innerhalb einer Menge von n Merkmalsträgern verteilt. Eine gleichmäßige Verteilung liegt vor, wenn alle Merkmalswerte übereinstimmen. Man spricht hier von fehlender Konzentration. Maximale Konzentration liegt hingegen vor, wenn ein einziger Merkmalsträger die gesamte Merkmalsumme auf sich vereint.

Ein grafisches Instrument zur Beurteilung von Konzentration ist die Lorenzkurve. Diese ist bei fehlender Konzentration durch die direkte Verbindung der Punkte $(0; 0)$ und $(1; 1)$ gegeben. Bei vorhandener Konzentration ist die Lorenzkurve hingegen ein von $(0; 0)$ bis $(1; 1)$ verlaufender „durchhängender“ Polygonzug, wobei sich das „Durchhängen“ mit zunehmender Konzentration verstärkt. Die Fläche zwischen dem Polygonzug und der im konzentrationsfreien Fall resultierenden Strecke kann zur Quantifizierung von Konzentration herangezogen werden. Man verwendet den Gini-Koeffizienten G , der durch das Zweifache der genannten Fläche definiert ist. Da die obere Schranke für G von der Anzahl n der Merkmalsträger abhängig ist, dividiert man G noch durch die obere Schranke und erhält ein normiertes Konzentrationsmaß G^* . Der normierte Gini-Koeffizient G^* wird u. a. bei der Analyse nationaler Einkommensverteilungen zur Quantifizierung von Einkommensungleichheiten verwendet.

Erwähnt wird noch der Herfindahl-Index. Dieser ist ein Konzentrationsmaß, das bei sehr kleiner Anzahl n von Merkmalsträgern Vorteile bietet.

7.1 Die Lorenzkurve

Bei metrisch skalierten Merkmalen mit nicht-negativen Ausprägungen – z. B. Umsätzen oder Marktanteilen von Firmen – interessiert man sich häufig dafür, wie sich die Summe aller Merkmalswerte innerhalb einer Grundgesamtheit verteilt. Konzentration bezüglich des jeweiligen Merkmals liegt vor, wenn sich die Merkmalsumme ungleichmäßig auf die betrachteten statistischen Einheiten verteilt.

Was bedeutet
„Konzentration“?

Fragen, die auf die Identifikation von Konzentrationsphänomen abzielen, sind etwa:

- Gibt es beim Vergleich ausgewählter Staaten größere Unterschiede hinsichtlich des Pro-Kopf-Energieverbrauchs?
- Wie ist das Einkommen von Arbeitnehmern in einer Volkswirtschaft oder einem Wirtschaftszweig verteilt?

- Gibt es innerhalb der Gruppe der weltweit größten Chip-Hersteller oder auf dem europäischen Automarkt einen marktbeherrschenden Produzenten?
- Gibt es in der Landwirtschaft eine Tendenz zu immer größeren Betrieben?

Beispiel 7.1: Energieverbrauch und CO₂-Emissionen



Interaktives Objekt
„Verbrauch von Primärenergie“

In Tabelle 5.1 waren Daten der *Internationalen Energieagentur* zum Pro-Kopf-Verbrauch von Erdöl und Strom sowie zu den CO₂-Emissionen pro Kopf für die USA, Deutschland, Japan, China und Indien wiedergegeben. Zum Datensatz für den Stromverbrauch (in t / Kopf) wurden in den Beispielen 5.2 und 5.5 bereits Kenngrößen berechnet, die sich für die Beschreibung des Zentrums oder der Streuung des Datensatzes eignen.

Bei der Konzentrationsmessung geht es nicht mehr darum, die Lage und Streuung eines Datensatzes zu charakterisieren. Vielmehr steht hier die numerische Bewertung von Ungleichheiten bei der Verteilung von Merkmalswerten auf die einzelnen Merkmalsträger im Vordergrund. Bezogen auf die Umweltdaten aus Tabelle 5.1 heißt dies z. B., dass man sich dafür interessiert zu quantifizieren, wie sich der gesamte Erdölverbrauch oder die gesamte CO₂-Emission aller fünf Länder innerhalb der 5 Elemente umfassenden Grundgesamtheit verteilt.



Video

„Lorenzkurve“

Ein grafisches Instrument zur Beurteilung von Konzentrationsphänomenen ist die **Lorenzkurve**. Sie ist nach dem amerikanischen Statistiker Max Otto LORENZ (1876 - 1959) benannt, der sie 1905 erstmals zur Veranschaulichung von Einkommensungleichheit einsetzte. Ausgangspunkt für die Herleitung einer Lorenzkurve ist eine Grundgesamtheit mit n Merkmalsträgern. Die zugehörigen Merkmalswerte konstituieren eine Urliste x_1, \dots, x_n . Wenn man deren Elemente nach *zunehmender Größe* sortiert, resultiert eine geordnete Liste $x_{(1)}, \dots, x_{(n)}$. Die über dem Intervall [0; 1] definierte Lorenzkurve visualisiert, wie sich die Summe aller Merkmalswerte innerhalb der Grundgesamtheit verteilt. Markiert man im Intervall [0; 1] die Punkte

$$u_i := \frac{i}{n}; \quad i = 1, \dots, n, \quad (7.1)$$

resultiert eine Zerlegung in n gleich lange Teilintervalle. Jeder Wert u_i lässt sich interpretieren als Anteil der ersten i Werte der Liste an der Gesamtzahl n der Elemente der Urliste. Bezeichnet man nun noch die Summe der kleinsten i Merkmalswerte mit

$$p_i := x_{(1)} + x_{(2)} + \dots + x_{(i)}; \quad i = 1, \dots, n \quad (7.2)$$

und den Anteil der zugehörigen Merkmalsträger an der Merkmalssumme p_n mit

$$v_i := \frac{p_i}{p_n}; \quad i = 1, \dots, n, \quad (7.3)$$

so ist die Lorenzkurve ein aus n Teilstrecken bestehender Polygonzug, der monoton steigt und den Punkt $(0; 0)$ mit den Punkten $(u_1; v_1), \dots, (u_n; v_n)$ verbindet. Offenbar ist $(u_n; v_n) = (1; 1)$, d. h. die Lorenzkurve endet in $(1; 1)$. Wenn alle Merkmalswerte gleich groß sind (fehlende Merkmalskonzentration), stimmen u_i und v_i jeweils überein. Die Lorenzkurve verbindet dann die Punkte $(0; 0)$ und $(1; 1)$ direkt. Um Konzentration anhand einer Lorenzkurve zu beurteilen, empfiehlt es sich auch die im konzentrationsfreien Fall resultierende Diagonale zu zeichnen. Je stärker die Lorenzkurve von der Diagonalen abweicht, d. h. je stärker sie „durchhängt“, desto größer ist die Konzentration.

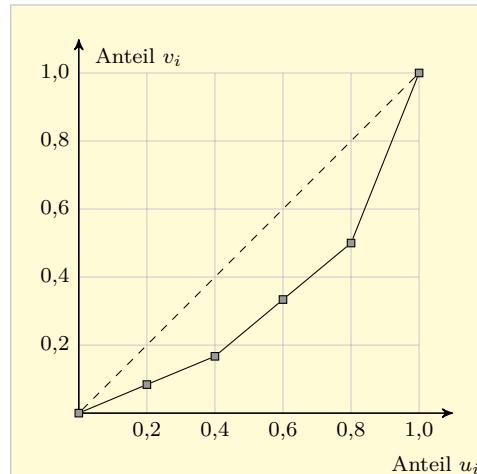


Abb. 7.1: Beispiel einer Lorenzkurve im Falle $n = 5$

Abbildung 7.1 zeigt die Lorenzkurve, die sich für eine Urliste mit den Werten 20, 20, 40, 40 und 120 ergibt. Für den Wert $(u_3; v_3)$ der Lorenzkurve errechnet man mit (7.1) - (7.3), dass $u_3 = 0,6$ und $v_3 = \frac{80}{240} \approx 0,333$. Dies beinhaltet, dass die kleinsten drei Werte der Urliste (60 % aller Merkmalswerte) nur insgesamt ca. 33,3 % der Merkmalssumme $p_5 = 240$ auf sich vereinen. Bei einer gleichmäßigen Verteilung der Merkmalssumme auf alle Merkmalsträger wäre $v_3 = 0,6$. Dies ist der Wert, den die in Abbildung 7.1 eingezeichnete Diagonale an der Stelle $u_3 = 0,6$ annimmt. Die Stützpunkte der Lorenzkurve bleiben unverändert, wenn man die Werte der Urliste mit einem positiven Faktor multipliziert.

Es sei erwähnt, dass die Berechnung von Lorenzkurven auch bei gruppierten Daten möglich ist. Der Polygonzug besteht bei Gruppierung zu k



Klassen aus k Teilstrecken. Details zur Berechnung der Stützpunkte der Lorenzkurve im Falle gruppierter Daten findet man z. B. bei FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 2.3.1) oder TOUTENBURG / HEUMANN (2009, Abschnitt 3.5.1).

Exkurs 7.1: Anwendung der Lorenzkurve in der Materialwirtschaft

Die Lorenzkurve findet unter der Bezeichnung **ABC-Analyse** auch in der *Materialwirtschaft* Anwendung. Ziel ist es hier, den Ressourceneinsatz auf die wertmäßig wichtigsten Materialien zu konzentrieren und Einsparpotenziale zu nutzen. Die für einen Produktionsprozess benötigten Materialien werden drei Gruppen A, B, C zugeordnet. Die A-Gruppe umfasst Materialien mit vergleichsweise niedrigem Mengen- und hohem Wertanteil. Zur B-Gruppe werden Materialien gerechnet, die mengen- und wertmäßig einen mittleren Anteil ausmachen. Die C-Gruppe fasst Materialien zusammen, denen ein hoher Mengen- und ein geringer Wertanteil zukommt.

Die Klassenbildung orientiert sich am Anteil der einzelnen Materialpositionen am Gesamtwert aller Materialien. Sie wird in der Praxis nicht immer einheitlich vorgenommen. Meist wird ein Anteil von ca. 0,75 am Gesamtwert (75 %), für die Materialien der A-Gruppe und Anteilswerte von etwa 0,15 resp. 0,10 für die Materialien der B- und C-Gruppe verwendet. Man geht so vor, dass man zunächst den Wert aller Materialpositionen für eine bestimmte Verbrauchsperiode – z. B. ein Jahr – ermittelt und die einzelnen Werte als Anteil am Gesamtwert aller in der Periode verbrauchten Materialien ausweist.

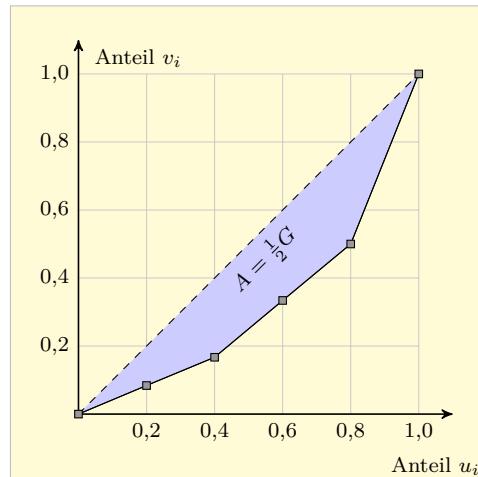
Wenn man die Anteilswerte dann nach *absteigender* Größe ordnet und bis zum Erreichen eines Wertes $v_1 \approx 0,75$ kumuliert, sind die Elemente der A-Gruppe identifiziert. Kumuliert man weiter bis zum Erreichen eines Anteilswertes $v_2 \approx 0,90$, hat man die Verbrauchspositionen der B-Gruppe gefunden. Die restlichen Positionen gehören zur C-Gruppe. Bezeichnet man den Wert aller Positionen der A-Gruppe mit p_1 , den Gesamtwert aller Elemente der A- und B-Gruppe mit p_2 und den Gesamtwert aller drei Gruppen mit p_3 , kann man den Nullpunkt $(0; 0)$ mit den Punkten $(u_1; v_1)$, $(u_2; v_2)$ und $(u_3; v_3)$ verbinden, wobei $u_i = \frac{i}{3}$ und $v_i = \frac{p_i}{p_3}$ gemäß (7.3). Die resultierende Lorenzkurve entspricht der in Abbildung 7.1, wenn man diese auf den Fall $n = 3$ bezieht und die Kurve an der gestrichelten Diagonalen spiegelt. Die Lorenzkurve ist hier nach oben gewölbt, weil die Anordnung der Wertanteile der einzelnen Materialposten bei der ABC-Analyse nach *abnehmender* Größe erfolgt.

7.2 Konzentrationsmaße

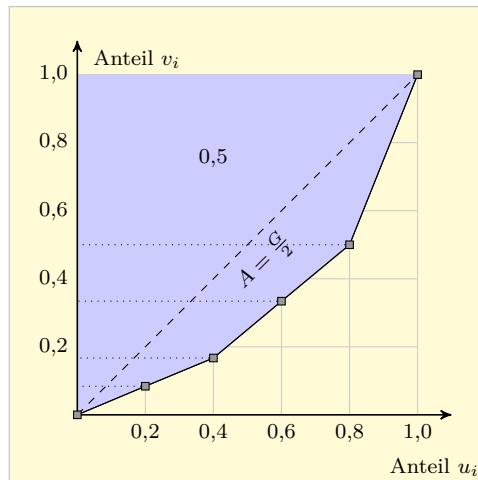
Die Lorenzkurve visualisiert Konzentrationsphänomene, repräsentiert aber noch kein Maß für die Stärke von Konzentration. Da sie sich mit zunehmender Konzentration immer mehr von der bei fehlender Konzentration resultierenden Diagonalen entfernt, liegt es nahe, die Fläche A zwischen der Diagonalen im Einheitsquadrat und der Lorenzkurve zur Konzentrationsmessung heranzuziehen. Der auf den italienischen Statistiker GINI (1884 - 1965) zurückgehende **Gini-Koeffizient** G ist ein solches Konzentrationsmaß. Er wurde zuerst für die Quantifizierung von Ungleichheiten bei Einkommensverteilungen herangezogen und ergibt sich aus dem Flächeninhalt A , indem man diesen mit dem Inhalt 0,5 eines der beiden Dreiecke vergleicht, in die das Einheitsquadrat durch die Diagonale zerlegt wird. Der Vergleich erfolgt durch Bildung des Quotienten $G = \frac{A}{0,5} = 2 \cdot A$ beider Flächeninhalte. Abbildung 7.2 weist erneut die Lorenzkurve aus Abbildung 7.1 für die Urwerte 20, 20, 40, 40 und 120 aus, nun mit Hervorhebung der Fläche $A = \frac{G}{2}$.



Corrado GINI

Abb. 7.2: Veranschaulichung von $\frac{G}{2}$ im Falle $n = 5$

Um $G = 2 \cdot A$ zu berechnen, ist es zweckmäßig, den Inhalt B der in Abbildung 7.3 betonten Fläche zu betrachten, die aus einem Dreieck mit dem Flächeninhalt 0,5 und einer Fläche mit dem Inhalt A besteht. Es gilt somit $B = \frac{G}{2} + 0,5$, d. h. $G = 2B - 1$. Die Fläche mit dem Flächeninhalt B lässt sich nun, wie in Abbildung 7.3 für den Fall $n = 5$ anhand gepunkteter horizontaler Linien angedeutet, in n Teilflächen zerlegen (ein Dreieck und $n - 1$ Trapeze), deren Flächeninhalte sich elementar bestimmen lassen. Für den Gini-Koeffizienten gilt also die Darstellung $G = 2 \cdot (\text{Summe der Inhalte der } n \text{ Teilflächen}) - 1$.

Abb. 7.3: Geometriegestützte Herleitung einer Formel für G

Man erhält bei Anwendung elementarer Flächeninhaltsformeln mit p_n aus (7.2) und der gewichteten Merkmalssumme

$$q_n := 1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \dots + n \cdot x_{(n)} \quad (7.4)$$

nach einigen Umformungen für den Gini-Koeffizienten die Darstellung¹

$$G = \frac{2 \cdot q_n}{n \cdot p_n} - \frac{n+1}{n} = \frac{1}{n} \left(\frac{2 \cdot q_n}{p_n} - 1 \right) - 1. \quad (7.5)$$

Für die Urliste mit den Elementen 20, 20, 40, 40 und 120, deren Lorenzkurve in Abbildung 7.1 dargestellt wurde, errechnet man $p_5 = 240$ und $q_5 = 940$ und hieraus $G \approx 0,367$. Die Berechnung von G setzt also nicht die Kenntnis der Stützpunkte $(u_i; v_i)$ der Lorenzkurve voraus.

In Abbildung 7.4 ist der Fall maximaler Konzentration dargestellt. Zu grunde gelegt wurde erneut eine Urliste mit $n = 5$ Elementen, bei der aber nur ein Wert positiv ist, etwa $x_{(5)} = 120$, und die anderen Werte Null sind. Die gesamte Merkmalssumme p_n konzentriert sich hier auf einen einzigen Merkmalsträger. Die Fläche A und damit auch der Gini-Koeffizient $G = 2A$ nehmen dann im hier betrachteten Spezialfall $n = 5$ den maximalen Wert $A_{\max} = 0,4$ resp. $G_{\max} = 0,8$ an.

Bei beliebigem n ist $A_{\max} = \frac{n-1}{2n}$, wie man anhand einfacher geometrischer Überlegungen verifizieren kann. Der Gini-Koeffizient $G = 2A$ ist also durch $G_{\max} = \frac{n-1}{n}$ nach oben begrenzt. Bei fehlender Konzentration ist $A = 0$; der Gini-Koeffizient G nimmt dann sein Minimum $G_{\min} = 0$ an.

¹Vgl. etwa BAMBERG / BAUR / KRAAPP (2017, Abschnitt 2.3.1).

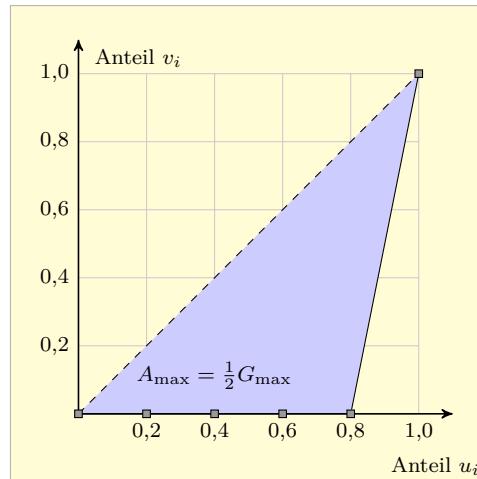


Abb. 7.4: Veranschaulichung von $\frac{G}{2}$ bei maximaler Konzentration ($n = 5$)

Für den Gini-Koeffizienten gilt also stets

$$0 \leq G \leq \frac{n-1}{n}. \quad (7.6)$$

Dass die obere Schranke von G von der Länge n der Urliste abhängt, ist ein Nachteil. Dieser lässt sich durch Einführung des **normierten Gini-Koeffizienten**

$$G^* := \frac{G}{G_{\max}} = \frac{n}{n-1} \cdot G \quad (7.7)$$

beheben. Für den normierten Gini-Koeffizienten hat man also



Interaktives Objekt
„Gini-Koeffizient“

$$0 \leq G^* \leq 1, \quad (7.8)$$

wobei die untere Schranke bei fehlender und die obere bei maximaler Merkmalskonzentration erreicht wird. Im Falle $0 < G^* \leq 0,5$ spricht man von mäßiger, im Falle $0,5 < G^* < 1$ von deutlicher Konzentration.

Besondere Bedeutung kommt dem Gini-Koeffizienten bei der Quantifizierung von Einkommensungleichheiten zu (s. hierzu den Exkurs 7.1). Für diese Zielsetzung kann man alternativ auch Quantilsquotienten empirischer Einkommensverteilungen heranziehen. Wenn man z. B., wie in Abbildung 4.8 illustriert, das Verhältnis $\frac{D_9}{D_1}$ aus oberem Dezil D9 und unterem Dezil D1 betrachtet, erhält man ebenfalls eine Information über Asymmetrien bei Einkommensverteilungen. Die Quantile D9 und D1 weisen aus, unterhalb welcher Schwelle 90% bzw. 10% der Einkommen liegen. Wie groß die oberhalb von D9 liegenden Einkommen sind, spielt keine Rolle. Die in den Daten steckende Information wird also – wie bei jeder Informationsverdichtung – nicht voll ausgeschöpft.

Quantifizierung von
Einkommens-
ungleichheit

Relative und absolute
Konzentration

Letzteres gilt auch für den Gini-Koeffizienten. Zum einen können unterschiedliche Urlisten der Länge n zum gleichen Gini-Koeffizienten führen. Die Hauptkritik am Gini-Koeffizienten bezieht sich aber auf die Konzentrationsmessung bei kleinen Datensätzen. Der Gini-Koeffizient zeigt fehlende Konzentration an ($G = G_{\min} = 0$), wenn alle Merkmalsträger einer Urliste übereinstimmen. Die Länge n der Urliste spielt dabei keine Rolle. Dies bedeutet, dass die Lorenzkurve, aus der sich der Gini-Koeffizient G ableitet, Aussagen des Typs „ $x\%$ der Merkmalsträger teilen sich $y\%$ der Merkmalssumme“ liefert, nicht aber Aussagen der Art „ x Merkmalsträger sind für $y\%$ der Merkmalssumme verantwortlich“. Je nachdem, ob man Aussagen für einzelne Merkmalsträger oder für Anteile in der Grundgesamtheit formuliert, wird absolute Konzentration bzw. relative Konzentration bewertet. Der Gini-Koeffizient misst *relative* Konzentration. Wenn aber z. B. ein Markt für ein bestimmtes Produkt oder eine bestimmte Dienstleistung von nur sehr wenigen Unternehmen beherrscht wird, kann man auch bei einem Wert von $G = 0$ nicht mit Berechtigung von fehlender Marktkonzentration sprechen. In diesem Falle lassen sich Maße für absolute Konzentration heranziehen.

Ein Maß für *absolute* Merkmalskonzentration ist der nach dem US-Ökonomen Orris C. HERFINDAHL (1918 - 1972) benannte **Herfindahl-Index**. Dieser ist definiert durch die Summe

$$H := \sum_{i=1}^n \left(\frac{x_i}{p_n} \right)^2 = \frac{1}{p_n^2} \cdot \sum_{i=1}^n x_i^2. \quad (7.9)$$

Der Wert dieser Summe hängt nicht davon ab, ob die Werte x_i der Urliste geordnet vorliegen, d. h. bei der Berechnung des Herfindahl-Indexes ist es – anders als beim Gini-Koeffizienten – nicht unbedingt erforderlich, die Elemente der Ausgangsurliste nach Größe zu ordnen.

Wenn vollständige Konzentration vorliegt, die gesamte Merkmalssumme also auf ein einziges Element entfällt, ist der Anteil dieses Elements an p_n offenbar 1 und der der anderen Elemente Null. Der Herfindahl-Index nimmt dann den Wert 1 an. Bei gleichmäßiger Merkmalsverteilung besitzen hingegen alle Anteile den Wert $\frac{1}{n}$ und der Index H nimmt sein Minimum $H_{\min} = n \cdot \left(\frac{1}{n}\right)^2 = \frac{1}{n}$ an. Es gilt demnach

$$\frac{1}{n} \leq H \leq 1. \quad (7.10)$$

Der Herfindahl-Index besitzt folglich, anders als der Gini-Koeffizient, eine positive untere Schranke, die mit abnehmender Länge n der Urliste größer wird ($H_{\min} = 0,5$ im Falle $n = 2$). Für die Urliste mit den Werten 20, 20, 40, 40 und 120, für die sich $G \approx 0,367$ und $G^* = \frac{5}{4} \cdot G \approx 0,458$ ergibt, errechnet man mit $p_5 = 240$ den Wert $H = \frac{1}{240^2} \cdot 18400 \approx 0,319$. Beim Herfindahl-Index können auch Werte, die nicht weit von H_{\min} entfernt

liegen (im Falle $n = 5$ also nicht weit vom Wert 0,2), bereits deutliche Konzentration beinhalten.

Der Herfindahl-Index wird u. a. von Kartellbehörden zur Messung unerwünschter Anbieterkonzentration eingesetzt, so z. B. in Deutschland von der Monopolkommission bei kartellrechtlichen Entscheidungen oder in den USA vom Antitrust Department.

Anwendungsfeld für den Herfindahl-Index

Beispiel 7.2: Konzentrationsmessung bei Stromverbrauchsdaten

Will man für die Daten zum Stromverbrauch in Beispiel 5.2 die Stützpunkte der Lorenzkurve sowie den Gini-Koeffizienten und den Herfindahl-Index berechnen, empfiehlt sich die Anlage einer kleinen Arbeitstabelle. Die Abszissenwerte der Stützpunkte $(u_i; v_i)$ der Lorenzkurve sind nach (7.1) durch $u_i = \frac{i}{5}$ gegeben, also durch 0,2, 0,4, …, 1,0; die Ordinatenwerte v_i errechnen sich nach (7.3). Für die Ermittlung des Gini-Koeffizienten G benötigt man noch die in (7.4) eingeführte gewichtete Merkmalssumme q_5 und für den Herfindahl-Index die Summe der quadrierten Urwerte. Wollte man nur den Herfindahl-Index berechnen, wäre die Ordnung der Urliste nach Größe nicht erforderlich.

i	x_i	$x_{(i)}$	p_i	v_i	$i \cdot x_{(i)}$	$x_{(i)}^2$
1	12,83	0,86	0,86	0,026	0,86	0,74
2	7,01	4,05	4,91	0,151	8,10	16,40
3	7,86	7,01	11,92	0,365	21,03	49,14
4	4,05	7,86	19,78	0,607	31,44	61,78
5	0,86	12,83	32,61	1,0	64,15	164,61
Summe:				$q_5 = 125,58$	292,67	

Tab. 7.1: Berechnung des Gini-Koeffizienten (Stromverbrauchsdaten)

Stellt man sich, analog zu Beispiel 5.2, wieder gedanklich eine Gruppe von 5 Personen vor, je eine Person aus den Ländern USA, Deutschland, Japan, China und Indien, und nimmt man an, dass für diese jeweils der in Tabelle 5.1 angegebene mittlere Jahresstromverbrauch ihres Landes zutrifft, so besagt z. B. der Punkt $(u_2; v_2) = (0,4; 0,151)$ der Lorenzkurve, dass 40% der Gruppe (die beiden Personen aus Indien und China mit dem niedrigsten Stromverbrauch) nur für etwa 15,1% des Gesamtstromverbrauchs der Gruppe verantwortlich sind, d. h. die restlichen 60% der Länder verbrauchen 84,9%. Entsprechend lässt sich aus $(u_4; v_4) = (0,8; 0,607)$ ableiten, dass die USA allein bereits 39,3% des Gesamtstromverbrauchs verursachen. Für den normierten Gini-Koeffizienten G^* sollte man also hier einen Wert erwarten, der eine mäßige Merkmalskonzentration beinhaltet. In der Tat ergibt sich mit (7.5) und den Werten p_5 und q_5 aus Tabelle 7.1

$$G = \frac{1}{5} \left(\frac{2 \cdot 125,58}{32,61} - 1 \right) - 1 \approx 0,340$$

und hieraus nach (7.7)



$$G^* = \frac{5}{4} \cdot G \approx 0,425.$$

Für den Herfindahl-Index erhält man nach (7.9) den Wert

Aufgaben 7.1-2

$$H = \frac{1}{32,61^2} \cdot 292,67 \approx 0,275.$$

Exkurs 7.2: Messung und Bewertung von Einkommensungleichheit

Die *OECD* veröffentlicht zur Charakterisierung von Einkommensungleichheit in ihren Mitgliedsländern Listen mit Gini-Koeffizienten. Für 2016 und 2017 waren die Werte für die skandinavischen Länder durchweg niedrig (im Bereich von 0,26 bis 0,27) und auch Deutschland hatte mit ca. 0,29 noch einen vergleichsweise niedrigen Gini-Koeffizienten. Deutlich höher lagen die Werte für Großbritannien (0,36) und die USA (0,39), Mexiko und Chile (je 0,46) sowie für Südafrika (0,62). Von den *Vereinten Nationen* und von der *Central Intelligence Agency (CIA)* der USA werden auch für andere Länder Listen mit Gini-Koeffizienten veröffentlicht, von der CIA im Rahmen des von ihr herausgegebenen *World Factbook*. Neben den Gini-Koeffizienten werden zur Beurteilung von Einkommensungleichheit auch Quotienten von Quantilen der nationalen Einkommensverteilungen eingesetzt.

Das durchschnittliche Einkommen sowie das anhand von Gini-Koeffizienten quantifizierte Ausmaß von Einkommensungleichheit in Staaten wird von WILKINSON / PICKETT (2010) mit Daten für unterschiedliche Merkmale verknüpft, die sich alle als Indikatoren für den Zustand einer Gesellschaft interpretieren lassen, u. a. die relative Häufigkeit von psychischen Störungen und Suchtproblemen, der Anteil der Schulabrecher oder Fettleibigen ($BMI > 30$) sowie die Quote der Inhaftierten oder Mörder. Die Autoren wollen mit dem u. a. aus Erhebungen der *OECD* und der *WHO* stammenden Datenmaterial belegen, dass in entwickelten Staaten weniger das absolute Einkommensniveau, sondern vielmehr die Einkommensverteilung ausschlaggebend für das soziale „Funktionieren“ einer Gesellschaft ist. Sie stellen heraus, dass Länder mit sehr ungleicher Einkommensverteilung – etwa die USA und Großbritannien – bezüglich der genannten Merkmale auffällig schlechter abschneiden als Länder mit weniger weit geöffneter Einkommensschere. Auch in Deutschland wird das Thema „Einkommensungleichheit“ stärker diskutiert, z. B. im *Handelsblatt* vom 7. Oktober 2019.



8 Indikatoren



Vorschau auf
das Kapitel

Im Zentrum dieses Kapitel stehen Indikatoren (Indexzahlen). Mit diesen versucht man komplexe gesellschaftsrelevante Entwicklungen abzubilden – etwa im Bereich Ökonomie, Gesundheit, Umwelt oder Bildung – und Vergleiche zwischen Regionen zu ermöglichen. Beispiele sind die Indikatoren „EU-Staatsschulden / Bruttoinlandsprodukt“ und „Militärausgaben / Kopf“ oder der Anteil der Erwerbstätigen an der Bevölkerung im erwerbsfähigen Alter.

Behandelt werden auch Indexzahlen, die durch Verknüpfung mehrerer Einzelindikatoren entstehen. Beispiele für solche zusammengesetzten Indikatoren sind der amtliche Verbraucherpreisindex oder der Human Development Index (HDI). Bei zusammengesetzten Indikatoren hängt der Indexwert von der Gewichtung der Einzelindikatoren ab.

8.1 Verhältniszahlen

In den Kapiteln 4 – 5 wurde dargestellt, wie man empirische Verteilungen für ein Merkmal anhand von Häufigkeiten sowie anhand weniger Kenngrößen zur Charakterisierung der Lage oder Streuung beschreiben kann. Zahlen, die einen Sachverhalt quantifizieren, nennt man allgemein **Maßzahlen**. Wenn man zwei Maßzahlen durch Quotientenbildung miteinander verknüpft, spricht man von einer **Verhältniszahl**. Verhältniszahlen sollen die Vergleichbarkeit statistischer Informationen für unterschiedliche Regionen oder Zeitpunkte ermöglichen. Es wäre z. B. wenig informativ, wenn man die registrierten Aids-Fälle in Deutschland und Luxemburg anhand der absoluten Häufigkeiten vergleiche, die sehr unterschiedlichen Bevölkerungszahlen also nicht in den Vergleich einbezöge. Beim Vergleich von Staatsschulden wird meist das Bruttoinlandsprodukt (BIP) anstelle der Bevölkerungszahl als Referenzwert herangezogen.

Sehr anschauliche Verhältniszahlen sind die in Abschnitt 4.1 behandelten relativen Häufigkeiten. Diese verknüpfen durch Anteilsbildung eine Teilgesamtheit mit einer Grundgesamtheit. Solche Verhältniszahlen, bei denen eine Grundgesamtheit durch Anteilsbildung bezüglich *eines Merkmals* strukturiert wird, nennt man auch **Gliederungszahlen**. Sie sind dimensionslos. Ein Beispiel für eine Gliederungszahl ist die Erwerbslosenquote p eines Landes. Sie verknüpft die Anzahl der Erwerbslosen im betreffenden Land mit der Anzahl aller Personen im erwerbsfähigen Alter. Eine Gliederungszahl p wird meist als Prozentwert ausgewiesen (Multiplikation mit 100).

Arten von
Verhältniszahlen

Es gibt Verhältniszahlen, die durch Quotientenbildung eine Verbindung zwischen *zwei* unterschiedlichen *Merkmale* herstellen. Man spricht dann von **Beziehungszahlen**. Die Verknüpfung der beiden Merkmale muss inhaltlich Sinn geben. Beispiele sind die Bevölkerungsdichte einer Region (Maßzahl: Einwohnerzahl / km^2), das Bruttoinlandsprodukt (Maßzahl: Euro / Einwohner), die Verschuldung eines Staates (Maßzahl: Euro / BIP oder Euro / Einwohner) oder der Body-Mass-Index.

Beispiel 8.1: Der Body-Mass-Index

Eine von der **World Health Organization (WHO)** und anderen Organisationen im Gesundheitsmonitoring verwendete Beziehungszahl ist der durch

$$\text{BMI} = \frac{\text{Körpergewicht (in kg)}}{[\text{Körpergröße (in m)}]^2}$$

definierte Body-Mass-Index. Als Faustregel gilt, dass erwachsene Personen mit einem BMI-Wert unter 18,5 als untergewichtig gelten, bei Werten von 18,5 bis unter 25,0 als normalgewichtig, von 25,0 bis unter 30,0 als übergewichtig und ab einem Wert von 30,0 als fettleibig. Der BMI-Wert ist einfach zu berechnen und eine grobe Methode, um zwischen Unter-, Normal- sowie Übergewichtigkeit und Fettleibigkeit zu unterscheiden. Er ist aber nicht perfekt, weil er weder zwischen Muskelmasse und Fettanteil unterscheidet und auch weder das Geschlecht noch das Alter eines Erwachsenen berücksichtigt.

Vor allem Fettleibigkeit ist heute ein die Gesundheitssysteme stark belastendes Problem vieler Länder. Das Problem wird sich in der Zukunft noch verschärfen, wenn nicht entschiedener gegengesteuert wird, z. B. durch Kennzeichnung stark zuckerhaltiger Lebensmittel sowie durch Maßnahmen zur Reduktion des Nikotin- und Alkoholkonsums und von Bewegungsarmut. Die Medien - u. a. die **Zeit** in einem Beitrag vom 11. Oktober 2017 – malen bereits Schreckensszenarien an die Wand, besonders mit Blick auf den weltweit stark wachsenden Anteil fettleibiger Kinder und Jugendlicher. Die WHO vermittelt **Basisinformationen** zur aktuellen Verbreitung und zur Bekämpfung von Fettleibigkeit.

In der Praxis wird manchmal der Quotient zweier Maßzahlen bestimmt, die sich zwar auf dasselbe Merkmal, aber auf Werte aus unterschiedlichen Beobachtungsperioden beziehen. Bei Zeitreihen, etwa für den Preis eines Produkts oder einer Dienstleistung, werden die Daten in der aktuellen Periode t ($t > 0$) durch die Werte einer Referenz- oder Basisperiode (Periode $t = 0$) geteilt. So werden Veränderungen gegenüber der Referenzperiode besser sichtbar. Das **Statistische Bundesamt** bezieht z. B. Preise für den privaten Verbrauch auf ein Referenzjahr. Der Preis x_t für Diesel-Kraftstoff im Jahr $t = 2019$ wird also nicht direkt, sondern in Form des Quotienten $I_t := \frac{x_t}{x_0}$ ausgewiesen, wobei x_0 den Preis im Referenzjahr bezeichnet. Verhältniszahlen, die die Werte für ein Merkmal

für *zwei Zeitpunkte* verknüpfen, werden **einfache Indexzahlen** genannt. Der Zusatz „einfach“ soll darauf verweisen, dass sich die Indexzahl nur auf ein einziges Merkmal bezieht.

Geeignete Maßzahlen werden als **Indikatoren** herangezogen, um komplexe Entwicklungen, etwa die Veränderung von objektiven Lebensbedingungen oder sozialer Kohäsion in einer Bevölkerung, möglichst repräsentativ abzubilden und Vergleiche zwischen Regionen zu ermöglichen. Es seien hier beispielhaft einige gesellschaftsrelevante Dimensionen genannt, für deren Messung unterschiedliche Indikatoren herangezogen werden:

- *Gesundheit*: Lebenserwartung Neugeborener; Anteil von Personen mit Fettleibigkeit, Anzahl der HIV-Fälle pro Million Einwohner, Letalität bzw. Mortalität von COVID-19 (Anteil der am Corona-Virus Verstorbenen an der Gesamtzahl der Infizierten resp. an der gesamten Bevölkerung);
- *Wohlstand*: BIP pro Kopf; Bruttoeinkommen von Arbeitnehmern pro Stunde; Erwerbstätigengquote; Anteil der nach amtlicher Definition als „arm“ geltenden Personen;
- *Bildung*: Abiturientenquote eines Jahrgangs; Anteil der Ausgaben für öffentliche und private Bildungseinrichtungen am BIP;
- *Innovationskraft*: Anzahl der Patente pro Einwohner; Bevölkerungsanteil mit Hochschulabschluss; Anteil der Staatsausgaben für Forschung und Entwicklung;
- *Öffentliche Sicherheit*: Polizeidichte; Aufklärungsquote bei Gewaltkriminalität; inhaftierter Bevölkerungsanteil;
- *Umwelt*: Anteil erneuerbarer Energien am Primärenergieverbrauch; Treibhausemissionen in CO_2 -Äquivalenten bezogen auf das Referenzjahr 1990;
- *Industrie*: Bewertung des Grades der Einhaltung von Toleranzen bei der Fertigung mit Prozessfähigkeitsindizes (vgl. Beispiel 15.2).

Das *Statistische Bundesamt* veröffentlicht Zeitreihen für Indikatoren und Indikatorensysteme für verschiedene Bereiche, u. a. **Indikatoren zur nachhaltigen Entwicklung** in Deutschland. Auf europäischer Ebene werden zahlreiche Indikatoren von *Eurostat* publiziert, z. B. Schlüsselindikatoren der Europa-2020-Strategie der EU. Die *Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS)* bietet ein umfassendes System sozialer Indikatoren für Deutschland und für europäische Länder an. Die *Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD)* hält zahlreiche Indikatoren auch für außereuropäische Länder bereit, u. a. für die Bereiche „Bildung“ und „Gesundheit“.

In den Medien werden Maß- oder Verhältniszahlen häufig für Vergleiche zwischen Ländern, Regionen oder auch Institutionen herangezogen. Das

Erfassung komplexer Entwicklungen anhand von Indikatoren



Interaktives Objekt „Erneuerbare Energien“



Interaktives Objekt „Treibhausgasemissionen pro Kopf“

Wo findet man Informationen über Indikatoren?

Erstellung von
Ranglisten

Stockholmer Institut für Friedensforschung und Rüstungskontrolle *SIPRI* (*Stockholm International Peace Research Institute*) veröffentlicht z. B. jedes Jahr die Militärausgaben von Staaten nicht nur anhand der in US-Dollar ausgewiesenen absoluten Werte, sondern auch anhand des Anteils der Militärausgaben am BIP sowie anhand der Beziehungszahl „Militärausgaben pro Kopf“. Je nachdem, welches Vergleichskriterium herangezogen wird, resultieren ganz unterschiedliche Ranglisten. Man muss daher bei Ranglisten aller Art schauen, auf welcher Vergleichsbasis sie fußen, ob der verwendete Indikator sachadäquat ist und ob eventuell die Verwendung mehrerer Indikatoren ein differenzierteres Bild liefert.

Beispiel 8.2: Vergleich der Militärausgaben von Ländern

Tabelle 8.1 weist für zwölf Länder die von *SIPRI* veröffentlichten Militärausgaben für 2019 aus. Die Tabelle zeigt in den letzten drei Spalten die absoluten Werte für Militärausgaben (in Milliarden US-Dollar, mittlerer Wechselkurs des Referenzjahrs), den Anteil dieser Ausgaben am BIP (in Prozent) sowie die Ausgaben pro Kopf (in vollen US-Dollar, ebenfalls zum mittleren Wechselkurs des Jahres 2018).



Interaktives Objekt
„Militärausgaben“

Rang	Nation	Militärausgaben		
		absolut (Mrd. US-Dollar)	in % des BIP	pro Kopf
1.	USA	731,8	3,4	2 224
2.	China	261,1	1,9	182
3.	Indien	71,1	2,4	52
4.	Russland	65,1	3,9	446
5.	Saudi-Arabien	61,9	8,0	1 805
6.	Frankreich	50,1	1,9	770
7.	Deutschland	49,3	1,3	590
8.	Großbritannien	48,7	1,7	720
9.	Japan	47,6	0,9	375
10.	Brasilien	26,9	1,5	128
11.	Israel	20,5	5,3	2 402
12.	Singapur	11,2	3,2	1 932

Tab. 8.1: Militärausgaben für 2019 im Ländervergleich (Quelle: *SIPRI*; Datenextraktion: April 2020)

Die Werte in den beiden ersten Datenspalten sind auf eine Dezimalstelle, die in der letzten Spalte auf ganze Zahlen gerundet. Die Länder sind nach absteigender Größe der absoluten Werte für die Militärausgaben geordnet. Ordnet man hingegen nach dem BIP-Anteil oder den Pro-Kopf-Ausgaben, ergeben sich andere Rangfolgen. In jeder Datenspalte ist der erste Rangplatz

durch normalen, der letzte Rangplatz durch kursiven Fettdruck betont. Jede der drei Datenspalten liefert eine andere Sicht auf dasselbe Thema. Die absoluten Werte vermitteln z. B. eine Vorstellung von der Größenordnung des Markts für militärische Güter und Dienstleistungen und von der Nachfragemacht einzelner Länder auf diesem Markt. Bei den Ausgaben pro Kopf wird die Wirtschaftskraft eines Landes ausgeblendet und nicht, wie in der vorletzten Spalte von Tabelle 8.1, mit dieser verknüpft. Für die NATO-Mitglieder sind 2,0% des Bruttoinlandsprodukts für Verteidigungsausgaben vereinbart. Auf Deutschland stieg seit dem US-Regierungswechsel 2017 der Druck, diese bisher deutlich verfehlte Zielmarke einzuhalten.

**Aufgabe 8.1**

In China starteten 2014 Pilotprojekte, die darauf abzielen, das Wohlverhalten der Einwohner des Landes anhand eines Indexwerts zu messen („individueller Sozialkredit“). Die meisten Daten stammen aus Überwachungskameras, die mit einer umfassenden biometrischen Datenbank verbunden sind. Jede unerwünschte Handlung, neben Verkehrsdelikten oder nicht vertragsgerechter Kredittilgung z. B. auch parteischädigendes Verhalten, kann zu einer Absenkung des Sozialkreditstands, zu einem Eintrag in einer „Schwarzen Liste“ und zu Sanktionen führen, etwa zur Streichung des Zugangs zu Flug- oder Bahntickets.

**Kritisch
nachgefragt**

Im Mai 2019 veröffentlichte die Menschenrechtsorganisation *Human Rights Watch* einen Bericht, der näher beschrieb, wie Massenüberwachung in der mehrheitlich von Uiguren bewohnten westchinesischen Provinz Xinjiang organisiert ist. Im November 2019 berichtete hierüber auch das *International Consortium of Investigative Journalists* und bezog sich auf geleakte Geheimdokumente der chinesischen Verwaltung, die sog. „China Cables“. Aus diesen geht hervor, wie modernste Technik – neben Software zur Gesichtserkennung auch Apps zur heimlichen Überwachung von Smartphones unter Nutzung von Methoden der künstlichen Intelligenz und der Statistik – in China systematisch und massiv zur Unterdrückung religiöser Minderheiten eingesetzt wird.



Video „Minderheiten in China“ (Human Rights Watch)

Die Anfang 2020 beginnende weltweite Corona-Krise wurde von China dazu genutzt, die im Land zur Überwindung der Endemie rigoros und scheinbar auch erfolgreich eingesetzten Überwachungspraktiken als Beleg für die Überlegenheit des Systems zu propagieren. Das chinesische Modell, in dem nicht offen über Fragen des Datenschutzes diskutiert werden kann, hat natürlich in einer Ausnahmesituation im Vergleich zum Demokratiemodell westlicher Länder weitreichendere Möglichkeiten der Bürgerkontrolle.

8.2 Zusammengesetzte Indexzahlen

Ranglisten erfreuen sich großer Aufmerksamkeit in den Medien – man denke etwa an das öffentliche Interesse an Ranglisten für Universitäten, an den Ergebnissen der Pisa-Studien oder an Produktbewertungen der *Stiftung Warentest*. Meist wird bei der Erstellung von Ranglisten aber nicht nur eine einzige Maß- oder Verhältniszahl herangezogen. Vielmehr werden oft mehrere Indikatoren zu einer einzigen Maßzahl verknüpft. Bei der Bewertung konkurrierender Produkte durch die *Stiftung Warentest* spielen z. B. neben dem Preis und technischen Eigenschaften auch Designaspekte und Aspekte der Umweltverträglichkeit eine Rolle. Die von Experten vorgenommene Gewichtung der in die Bewertung eingehenden Merkmale wird in den Testergebnissen ausgewiesen.

Aggregation mehrerer Indikatoren
Die Verknüpfung mehrerer Indikatoren zu einer einzigen Maßzahl ist jedenfalls in vielen Bereichen des gesellschaftlichen Lebens gängige Praxis. Die resultierenden Aggregate werden **zusammengesetzte Indexzahlen** oder **zusammengesetzte Indikatoren** genannt (engl.: *composite indices*). Die bei ihrer Konstruktion herangezogenen einzelnen Indikatoren können gleich oder unterschiedlich gewichtet sein.

Problem: Festlegung der Gewichte
Schon an der zunächst einfach erscheinenden Frage nach der sportlich erfolgreichsten Nation bei einer Olympiade anhand von Medaillenspiegeln zeigt sich ein grundsätzliches Problem, das mit der Erstellung von Ranglisten auf der Basis zusammengesetzter Indikatoren verbunden ist. Es ist das Problem der sachgerechten Festlegung der Gewichte für die einzelnen Indikatoren.

Beispiel 8.3: Medaillenspiegel bei der Olympiade

Der offizielle Medaillenspiegel von Olympiaden orientiert sich primär an der *Anzahl der Goldmedaillen*; nur bei Gleichstand wirkt sich die Anzahl der Silber- und Bronzemedaillen auf die Platzierung aus. Diese Art der Bewertung der sportlichen Leistung eines Landes wurde nach Abschluss der Sommerolympiade 2008 erstmals in verschiedenen Internetforen und in der Presse sehr kontrovers diskutiert, zumal amerikanische Zeitungen einen anderen Medaillenspiegel führten, bei dem die *Gesamtzahl der Medaillen* als Indikator für den sportlichen Erfolg einer Nation fungierte.

Tabelle 8.2 zeigt die ersten zehn Platzierungen des offiziellen Medaillenspiegels der Sommerolympiade 2008. China belegt hier den ersten Rangplatz.

Rang	Nation		Gold	Silber	Bronze	Gesamt
1.		China	51	21	28	100
2.		USA	36	38	36	110
3.		Russland	23	21	28	72
4.		Großbritannien	19	13	15	47
5.		Deutschland	16	10	15	41
6.		Australien	14	15	17	46
7.		Südkorea	13	10	8	31
8.		Japan	9	6	10	25
9.		Italien	8	10	10	28
10.		Frankreich	7	16	17	40

Tab. 8.2: Offizieller Medaillenspiegel der Sommerolympiade 2008 (Auszug)



In Tabelle 8.3 sind die ersten zehn Platzierungen zu sehen, die sich nach dem US-amerikanischen Bewertungsmodus ergaben. Die Rangnummern von Ländern, die hier einen anderen Rang belegen, sind durch Fettdruck markiert.



Rang	Nation		Gesamt	Gold	Silber	Bronze
1.		USA	110	36	38	36
2.		China	100	51	21	28
3.		Russland	72	23	21	28
4.		Großbritannien	47	19	13	15
5.		Australien	46	14	15	17
6.		Deutschland	41	16	10	15
7.		Frankreich	40	7	16	17
8.		Südkorea	31	13	10	8
9.		Italien	28	8	10	10
10.		Japan	25	9	6	10

Interaktives Objekt
„Sommerolympiade
2016“

Tab. 8.3: US-amerikanischer Medaillenspiegel der Sommerolympiade 2008

Im *Guardian* erschien ein Beitrag, der die Fragwürdigkeit des offiziellen Rankings thematisierte und weitere Bewertungsalternativen aufzeigte, z. B. eine Bewertung der sportlichen Leistung eines Landes nach der *Anzahl der Goldmedaillen pro Kopf*. Bei Verwendung dieses Ansatzes hätten 2008 Jamaika und Bahrain ganz vorne gelegen und die führenden Länder der amtlichen Liste wären auf weit hinten liegende Plätze gerückt. Es gab Vorschläge, auch die Wirtschaftskraft eines Landes einzubeziehen, weil diese die Trainingschancen von Sportlern beeinflussen kann.



Aufgabe 8.2

Über die Sinnhaftigkeit des offiziellen Medaillenspiegels, der Silber- und Bronzemedaillen nur hilfsweise berücksichtigt, lässt sich sicher streiten. Aber auch das in den USA praktizierte Addieren von Medaillen ohne Differenzierung zwischen Gold, Silber und Bronze erscheint willkürlich. Ein Kompromiss könnte darin bestehen, zwar alle Medaillen zu addieren, aber mit unterschiedlichen Gewichten für Gold, Silber und Bronze. Dabei wäre zu klären, wie die Gewichte festgelegt werden. Sind z. B. 3 Punkte für Gold, 2 für Silber und 1 Punkt für Bronze passender als die Abstufung 5 – 3 – 2? Zudem wäre zu diskutieren, ob es nicht angemessener wäre die *Anzahl* der Gold-, Silber- und Bronzemedaillen *pro Einwohner* eines Landes heranzuziehen.

Statistiker können allerdings die Frage nach der sachadäquatesten Operationalisierung des Merkmals „Sportlicher Erfolg einer Nation bei der Olympiade“ nicht beantworten. Welcher Ansatz die sportliche Leistung eines Landes am besten widerspiegelt, könnte z. B. von einem internationalen Sportkomitee per Mehrheitsbeschluss entschieden werden.

Zusammengesetzte
Indikatoren in der
Wirtschaft

Zusammengesetzte Indikatoren werden auch zur Beschreibung von Entwicklungen im ökonomischen Bereich herangezogen. Als Beispiel seien **Aktienindizes** angeführt, etwa der **Deutsche Aktienindex** (DAX) oder der **Dow Jones Index**. Ein weiteres bekanntes Beispiel ist der amtliche **Verbraucherpreisindex**. Der Verbraucherpreisindex ist ein gewichteter Mittelwert der auf eine Basisperiode bezogenen Preise für den Inhalt eines „repräsentativen“ Warenkorbs. Als Gewichte verwendet man die Ausgabenanteile der Güter und Dienstleistungen im Warenkorb in einer Referenzperiode – seit Anfang 2019 ist es das Jahr 2015 – für die der Index auf 100 gesetzt ist.

Beispiel 8.4: Der amtliche Verbraucherpreisindex



Video „Warenkorb“

Die Entwicklung der Verbraucherpreise für über 600 häufig nachgefragte Güter und Dienstleistungen wird vom **Statistischen Bundesamt** laufend verfolgt. Diese Güter und Dienstleistungen sollen das Konsumverhalten der Bevölkerung widerspiegeln. Sie bilden in ihrer Gesamtheit einen virtuellen Warenkorb. Die Veränderungen der Preise der Güter des Warenkorbs gehen in die Berechnung der *Inflationsrate* ein. Diese gibt die prozentuale Veränderung des Preises für den Warenkorb gegenüber dem Vorjahr an.



Inflationsrechner

Das Statistische Bundesamt bietet eine sehr benutzerfreundliche Darstellung der Inflationsrate anhand eines interaktiven *Inflationsrechners* an. Der interaktive Inflationsrechner zeigt die Entwicklung in Form eines Zeitreihengraphen für den Verbraucherpreisindex und zusätzlich für eine vom Betrachter frei wählbare Güterklasse – in Abbildung 8.1 ist es die Güterklasse „Kraftstoffe“. Auffälligkeiten bei der Preisentwicklung sind hier vor allem auf Schwankungen der Preise für Rohöl zurückzuführen. Fährt man mit der Maus oder, bei mobilen Endgeräten, mit dem Finger über einen Graphen, wird der Zeitpunkt ausgewiesen, auf den sich der jeweilige Kurvenpunkt bezieht. Wer bestimmte

Güter – etwa Tabakwaren – nicht oder nur in geringem Umfang benötigt, kann die Gütergruppe ausblenden oder ihr Gewicht reduzieren und sich auf der Basis dieses personalisierten Warenkorbs seinen individuellen Verbraucherpreisindex anzeigen lassen.



Aufgabe 8.3

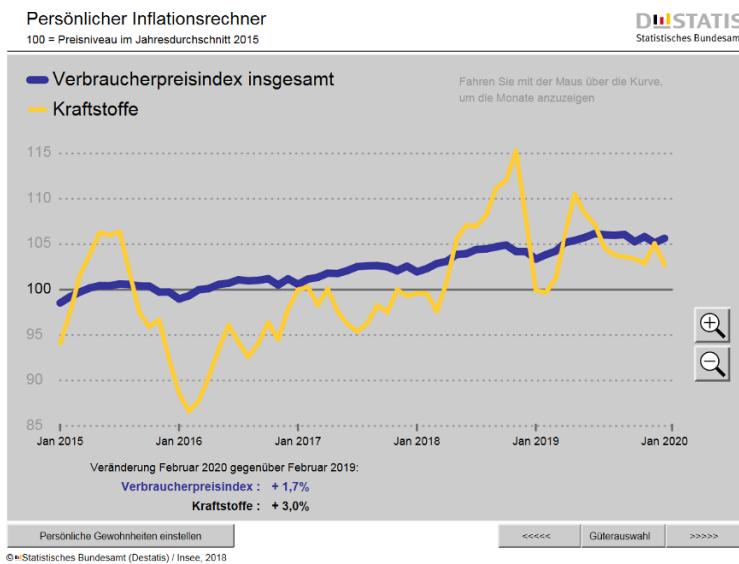


Abb. 8.1: *Inflationsrechner* (Statistisches Bundesamt; Februar 2020)

Wie sich der zur Berechnung des Verbraucherpreisindexes herangezogene Warenkorb zusammensetzt und wie groß die Gewichte der einzelnen Güter sind, veranschaulicht das Statistische Bundesamt anhand eines innovativen, als *Preiskaleidoskop* bezeichneten Visualisierungsinstruments. Der Warenkorb ist hier durch einen Kreis repräsentiert, während die Warengruppen und deren Komponenten mosaiksteinartig durch Anteile an der Kreisfläche dargestellt sind. Die Größe der „Mosaiksteine“ spiegelt jeweils den Ausgabenanteil der Warengruppe am Warenkorb wider. Die Flächeninhalte visualisieren somit das Gewicht, mit dem die Warengruppe oder eine bestimmte Komponente einer Warengruppe in den Verbraucherpreisindex eingeht. Die Gewichte für die einzelnen Komponenten der zusammengesetzten Indexzahl „Verbraucherpreisindex“ sind somit – anders als beim Medaillenspiegel von Olympiaden – durch Beobachtungsdaten eindeutig bestimmt. Durch unterschiedliche Färbungen werden beim Preiskaleidoskop auch die Veränderungen gegenüber dem Vorjahresmonat sichtbar gemacht. Geht man mit der Maus auf eine Mosaikfläche, werden der Name der Warengruppe bzw. der Komponente angezeigt sowie das Gewicht und die Preisänderung gegenüber dem Vorjahresmonat. In Abbildung 8.2 ist die Komponente „Speisefette und Speiseöle“ der Ausgabengruppe „Nahrungsmittel und alkoholfreie Getränke“ betont. Die



Preiskaleidoskop

Ausgaben hierfür gingen mit einem Gewicht von 0,2 % in den Warenkorb ein und die Preise lagen etwa 6,5 % unter dem Vorjahresniveau.

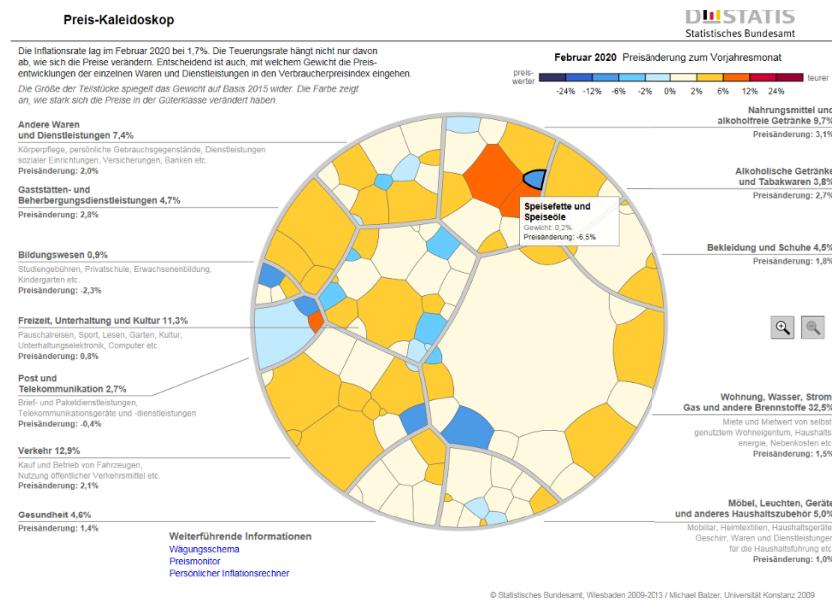


Abb. 8.2: Preiskaleidoskop (Statistisches Bundesamt; Februar 2020)

- Zusammengesetzte Indikatoren in der Politik** Zusammengesetzte Indexzahlen werden heute von verschiedenen supranationalen Institutionen wie der OECD, der Europäischen Kommission und den Vereinten Nationen eingesetzt, etwa zur Messung von Wohlfahrt oder zur Bewertung von Politiken und Fortschritten im Bereich der Entwicklungshilfe, des Umweltschutzes sowie der Technologieförderung. Genannt seien beispielhaft der **Human Development Index** und der **Human Poverty Index** der Vereinten Nationen. Beide bilden die Wohlfahrtsentwicklung in verschiedenen Ländern ab. Erwähnt sei auch der **Global Innovation Index**, an dessen Entwicklung die US-amerikanische Cornell University beteiligt ist.
- Vor- und Nachteile zusammen gesetzter Indikatoren** Die z. Z. verwendeten zusammengesetzten Indexzahlen repräsentieren additive Verknüpfungen eines Sets von Maß- und Verhältniszahlen, brechen also umfassende Indikatorensysteme auf eine einzige Variable herunter. Das gewachsene Interesse an ihnen erklärt sich daraus, dass sie
- eine eindimensionale Betrachtung multidimensionaler Phänomene ermöglichen;
 - einen direkten Ländervergleich gestatten und damit mehr Beachtung in den Medien finden als komplexe Systeme von Einzelwerten.
- Es gibt aber auch gewichtige Nachteile. Diese sind darin zu sehen, dass

- zusammengesetzte Indikatoren oft nur eine begrenzte Aussagekraft haben, weil ihre Werte von den Gewichten für die einfließenden Maß- und Verhältniszahlen abhängen und die Festlegung der Gewichte nicht immer unmittelbar nachvollziehbar oder motivierbar ist;
- die in sie eingehenden Einzelindikatoren im Zeitverlauf nicht selten geändert werden (Aufnahme neuer Indikatoren, Veränderung der Operationalisierung) und damit Rangplätze für Länder für verschiedene Zeitpunkte nicht unbedingt vergleichbar sind.

Die Rankings für Länder hängen jedenfalls davon ab, wie die Gewichte der einzelnen Indikatoren spezifiziert werden. Häufig werden alle in einen zusammengesetzten Indikator eingehenden Sub-Indikatoren mit gleichem Gewicht verknüpft, weil man keine Informationen hat, die eine unterschiedliche Gewichtung motivieren. Der Verbraucherpreisindex ist hier eine Ausnahme – das Gewichtungsschema ergibt sich bei diesem zusammengesetzten Indikator auf natürliche Weise aus Daten (s. Abbildung 8.2). Wenn man zusammengesetzte Indexzahlen verwendet, sollte man sie jedenfalls lediglich als grobe erste Orientierungsmarken verstehen. Man muss wissen, dass ihr Gebrauch eine genauere Betrachtung der in sie eingehenden Einzelindikatoren nicht ersetzen kann, weil nur diese eine differenzierte Bewertung komplexer Sachverhalte erlauben.

Kritische
Anmerkung zu
Ranglisten

Exkurs 8.1: Der Human Development Index der UN

Der **Human Development Index (HDI)** der Vereinten Nationen (UN) verknüpft drei Dimensionen, die den Entwicklungsstand eines Landes charakterisieren, nämlich *Gesundheit*, *Bildungsstand* und *Lebensstandard* der Bevölkerung. Der Gesundheitsstatus wird über die Lebenserwartung von Neugeborenen abgebildet, der Lebensstandard seit 2010 über das in Kaufkraftparitäten umgerechnete Bruttonationaleinkommen pro Einwohner. Zur Messung des Bildungsstands werden die Ausprägungen zweier Merkmale kombiniert, nämlich die in Jahren wiedergegebene durchschnittliche Dauer B1 des früheren Schulbesuchs von Erwachsenen im Alter von mindestens 25 Jahren und die erwartete Dauer B2 des Besuchs von Bildungseinrichtungen bei Kindern im Einschulungsalter. Aus den verwendeten Indikatoren wird ein Mittelwert gebildet (geometrisches Mittel), der so normiert wird, dass er stets Werte im Intervall [0; 1] annimmt. Ein HDI-Wert unter 0,5 wird als Indiz für einen geringen Entwicklungsstand des Landes interpretiert, Werte zwischen 0,5 und 0,8 als Zeichen für einen mittleren Stand und HDI-Werte ab 0,8 als Ausweis eines hohen Entwicklungsstandes. Europäische Länder finden sich regelmäßig im oberen Feld, während die untere Kategorie durchweg von afrikanischen Staaten belegt ist.



Video der UN zum
HDI-Report 2018

Tabelle 8.4 zeigt für das Jahr 2018 die besten sechs HDI-Werte und die beiden niedrigsten Werte. Die besten und schletesten Werte für die Indikatoren, aus denen sich der HDI zusammensetzt, sind ebenfalls ausgewiesen.

	HDI (Gesamtindex)	Gesundheit (Indikator <i>Lebenserwartung</i>)
1.	Norwegen (0,954)	1. Hongkong, China (84,7)
2.	Schweiz (0,946)	2. Japan (84,5)
3.	Irland (0,942)	3. Schweiz (83,6)
4.	Deutschland; Hongkong, China (0,939)	4. Singapur (83,5)
5.	Australien; Island (0,938)	5. Italien; Spanien (83,4)
6.	Schweden (0,937)	6. Australien (83,3)
	⋮	⋮
188.	Zentralafrik. Rep. (0,381)	188. Tschad (54,0)
189.	Niger (0,377)	189. Zentralafrik. Rep. (52,8)

	Bildungsstand (Indikatoren B1 / B2)	Lebensstandard (Indikator „Kaufkraft“)
1.	Deutschland (14,1) / Australien (22,1)	1. Qatar (110 489)
2.	Schweiz (13,4) / Belgien (19,7)	2. Liechtenstein (99 732)
3.	USA (13,4) / Finnland (19,3)	3. Singapur (83 793)
4.	Kanada (13,3) / Island (19,2)	4. Brunei (76 389)
5.	Israel (13,0) / Dänemark (19,1)	5. Kuwait (71 164)
6.	Litauen (13,0) / Irland (18,8)	6. Norwegen (68 389)
	⋮	⋮
188.	Niger (2,0) / Eritrea (5,0)	188. Zentralafrik. Rep. (777)
189.	Burkina Faso (1,6) / Südsudan (5,0)	189. Burundi (660)

Tab. 8.4: HDI-Werte und Sub-Indikatoren ausgewählter Länder für 2018
(Quelle: *Human Development Report 2019* der UN)

Man erkennt, dass man auf die Sub-Indikatoren bei einer Gesamtbeurteilung eines Staates nicht verzichten sollte, weil sich hier ein differenzierteres Bild ergibt. Man sieht insbesondere, dass sich die Werte für den HDI-Gesamtindex von Ländern mit benachbarten Rangplätzen – etwa die von Deutschland und Island – oft kaum unterscheiden. Kleinsten Messfehler oder minimale Veränderungen des Gewichtungsschemas können eine andere Rangfolge liefern.

Anhaltende kriegerische Konflikte spiegeln sich in abfallenden HDI-Rangplätzen wider. Die Ränge der Länder Jemen, Libyen und Syrien bezüglich des HDI-Gesamtindexes lagen im Jahr 2012 z. B. bei 158, 82 und 128, im Jahr 2018 nur noch bei 177, 110 resp. 154.

Welche Indikatoren oder Indikatorensysteme für die Erfassung einer gesellschaftsrelevanten Dimension, etwa „Wirtschaftswachstum“, besonders aussagekräftig sind, ist nicht immer leicht zu beantworten. Aus diesem Grund werden auch Alternativen zum Wohlfahrtsmaß „Bruttoinlandsprodukt (BIP)“ diskutiert, weil das BIP auch mit Umweltvernichtung einhergehendes Wirtschaftswachstum als Fortschritt bewertet und weder unbezahlte Arbeit noch Einkommensungleichheiten in einer Gesellschaft erfasst. Hinzu kommt, dass zusammengesetzte Indikatoren, die dasselbe zu messen scheinen, aufgrund unterschiedlicher Methodiken nicht unbedingt direkt vergleichbar sind. Manchmal ändern sich auch Operationalisierungen von Variablen. Die Umsetzung der letzten Aktualisierung des Europäischen Systems Volkswirtschaftlicher Gesamtrechnungen (ESVG) im September 2014 durch das Statistische Bundesamt hatte z. B. zur Folge, dass das Bruttoinlandsprodukt um ca. 3 % allein aufgrund geänderter Messvorschriften stieg.

Sind unterschiedliche Indikatorensysteme vergleichbar?

Exkurs 8.2: Weitere Wohlfahrtsindikatoren

Das Statistikamt von Großbritannien veröffentlicht einen als *National Well-Being-Index* bezeichneten zusammengesetzten Index, der makroökonomische Daten mit Daten zum subjektivem Wohlbefinden verknüpft. Die Ergebnisse werden anhand einer interaktiven Karte zugänglich gemacht. Genannt sei auch der von der OECD entwickelte *Better-Life-Index*, der ebenfalls über den klassischen Wohlfahrtsindikator „Bruttoinlandsprodukt“ hinausgeht. Er wird von der OECD berechnet, u. a. für Deutschland.

Außerhalb der amtlichen Statistik gibt es ebenfalls Ansätze zur Messung von Lebenszufriedenheit, etwa den vom US-amerikanischen Meinungsforschungsinstitut Gallup und der US-Firma Healthways geführten Well-Being-Index, der das Wohlbefinden von Menschen in verschiedenen Ländern widerspiegeln soll. Beide Indikatoren unterscheiden sich hinsichtlich der Sub-Indikatoren, die in den Index eingehen. Die für Deutschland berechneten Werte des Better-Life-Index und des Gallup-Healthways Well-Being-Indexes konkurrieren mit Daten, die im *Glücksatlas Deutschland* zusammengefasst sind. Die hier veranschaulichten und auf einer Likert-Skala von 0 bis 10 erhobenen Zufriedenheitswerte, in Form eines „Glücksindexes“ für 19 Regionen ausgewiesen, werden mit Recht kritisch hinterfragt. Als Beispiel sei ein Beitrag in der *FAZ* vom 18. November 2013 angeführt. Die Kritik bezieht sich vor allem auf die sehr geringen Stichprobenumfänge. Da die regionalen Zufriedenheitsunterschiede klein sind, kann schon der unvermeidliche Stichprobenfehler das Ranking determinieren.

Das *Weltwirtschaftsforum* veröffentlicht jährlich in seinem *Global Competitiveness Report* den *Global Competitiveness Index (GCI)*, mit dem die Wettbewerbsfähigkeit von Staaten verglichen wird. Der Index besteht – ähnlich wie der Human Development Index – aus drei Teilindizes, die noch weiter untergliedert sind. Der Teilindex *Basic Requirements* bezieht sich u. a. auf die

Infrastruktur eines Landes und den Zustand des Gesundheits- und Grundschulwesens, der Teilindex *Efficiency Enhancers* u. a. auf die Effizienz des Güter- und Arbeitsmarkts, der Teilindex *Innovation and Sophistication Factors* auf den Entwicklungsstand der Wirtschaft und die Innovationskraft eines Landes. Die Gewichtung der Teilindizes wird von der Höhe des Bruttoinlandsprodukts abhängig gemacht. Sowohl quantitative Faktoren wie die Höhe der Steuersätze, als auch aus Expertenbefragungen gewonnene qualitative Befunde, etwa zur Unabhängigkeit der Justiz und zur Verbreitung von Korruption, fließen in den Index ein. Im CGI-Report 2019 belegte Singapur den ersten und Deutschland den sechsten Rangplatz.

Erwähnt seien auch der *World Values Survey*, der Wohlfahrt, subjektives Wohlbefinden und soziokulturelle Wertemuster zu erfassen sucht und sich auf persönliche Interviews stützt (mindestens 1 000 pro Land), sowie der *World Press Freedom Index*. Letzterer bewertet weltweit den Grad der Pressefreiheit auf der Basis von Fragen, die sich auf sieben Einzelindikatoren beziehen.



9 Bivariate Häufigkeitsverteilungen



Vorschau auf
das Kapitel

Bei einem diskreten Merkmal X mit k Ausprägungen kann man die Häufigkeiten für die einzelnen Ausprägungen feststellen. Es resultiert eine univariate Häufigkeitsverteilung. Hat man *zwei* diskrete Merkmale X und Y mit k bzw. m Ausprägungen, kann man die absoluten oder relativen Häufigkeiten für die $k \cdot m$ Ausprägungskombinationen tabellarisch präsentieren. Die auch als Kontingenztafel bezeichnete Tabelle definiert eine bivariate Häufigkeitsverteilung. Ein Spezialfall einer Kontingenztafel ist die Vierfeldertafel, bei der X und Y jeweils nur zwei Ausprägungen aufweisen.

Eine Kontingenztafel kann man um die univariaten Häufigkeitsverteilungen ergänzen. Diese werden Randverteilungen genannt und ergeben sich durch Aufsummieren aller Werte einer jeden Zeile bzw. einer jeden Spalte. Die Randverteilungen werden benötigt, um bedingte Häufigkeiten zu berechnen. Letztere sind die für eine Ausprägungskombination beobachteten Häufigkeiten unter der Nebenbedingung, dass für X oder für Y eine bestimmte Ausprägung gilt. Randverteilungen und bedingte Häufigkeiten werden anhand von Daten des ZDF-Politbarometers veranschaulicht. Beide spielen eine zentrale Rolle bei der Untersuchung eines möglichen Zusammenhangs zwischen X und Y . Wenn kein Zusammenhang besteht, spricht man von empirischer Unabhängigkeit der beiden Merkmale.

Am Ende des Kapitels geht es um die Präsentation von Daten für zwei stetige Merkmale anhand von Streudiagrammen.

9.1 Empirische Verteilungen diskreter Merkmale

In Abschnitt 4.1 wurde beschrieben, wie man Daten für ein diskretes oder ein gruppiertes stetiges Merkmal X anhand von absoluten oder relativen Häufigkeitsverteilungen charakterisieren und grafisch präsentieren kann. In vielen Anwendungen interessiert man sich aber nicht nur für ein einziges, sondern gleichzeitig für zwei oder mehr Merkmale, für die ein Datensatz von je n Beobachtungswerten vorliegt. Diese Daten will man grafisch aufbereiten und Zusammenhänge zwischen den Merkmalen erfassen. Die folgenden Ausführungen beschränken sich auf den Fall *zweier* Merkmale, also auf die **bivariate Datenanalyse**. Als Beispiele für die gemeinsame Erhebung zweier Merkmale seien die simultane Erfassung der Merkmale „Parteipräferenz X von Wählern“ und „Geschlecht Y “ genannt oder „Jahresbruttoeinkommen X eines Arbeitnehmers“ und „Bildungsstand Y “, letzterer operationalisiert über den höchsten erreichten Bildungsabschluss einer Person. Wie man Datensätze für zwei Merkmale

aufbereitet und welches Zusammenhangsmaß verwendet werden kann, hängt von der Merkmalsskalierung ab.

Ausgangspunkt sei eine Erhebung, bei der für zwei *diskrete* Merkmale X und Y mit beliebiger Skalierung an n Untersuchungseinheiten jeweils die Merkmalsausprägung festgestellt wird. Die folgenden Ausführungen lassen sich auch auf *gruppierte stetige* Merkmale beziehen; die Ausprägungen entsprechen dann den Klassen. Das Merkmal X weise die Ausprägungen a_1, \dots, a_k , das Merkmal Y die Ausprägungen b_1, \dots, b_m auf. Die Merkmalswerte x_1, \dots, x_n und y_1, \dots, y_n repräsentieren eine **bivariate Urliste**. Diese lässt sich z. B. in der Form $(x_1, y_1), \dots, (x_n, y_n)$ schreiben, wobei Merkmalspaare (x_i, y_i) mehrfach auftreten können. Auch bei bivariaten Urlisten kann man die in den Rohdaten enthaltene Information aggregieren, hier durch Angabe von Häufigkeiten für das Auftreten von Ausprägungskombinationen oder – bei gruppierten Daten – für Kombinationen von Klassenbesetzungshäufigkeiten. Analog zu (4.1) bezeichne

$$h_{ij} := h(a_i, b_j) \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, m \quad (9.1)$$

die **absolute Häufigkeit** und analog zu (4.2)

$$f_{ij} := f(a_i, b_j) \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, m \quad (9.2)$$

Gemeinsame
Verteilung zweier
Merkmale

die **relative Häufigkeit** für die Ausprägungskombination (a_i, b_j) . Die $k \cdot m$ Häufigkeiten h_{ij} und f_{ij} definieren die gemeinsame **absolute Häufigkeitsverteilung** resp. **relative Häufigkeitsverteilung** der Merkmale X und Y . Man kann diese übersichtlich in tabellarischer Form wiedergeben. Die resultierende Tabelle heißt **Kontingenztafel** oder **Kontingenztabelle**, gelegentlich auch **Kreuztabelle**. Sie definiert die gemeinsame **empirische Verteilung** der beiden Merkmale. Die Dimension einer Kontingenztafel wird durch die Anzahl k und m der Ausprägungen für X und Y bestimmt. Meist gibt man die Dimension mit an und spricht im Falle von $k \cdot m$ Ausprägungskombinationen von einer $(k \times m)$ -Kontingenztabelle. Tabelle 9.1 zeigt eine solche für absolute Häufigkeiten.

Tabellen für
bivariate
Häufigkeits-
verteilungen

Die Tabelle weist in einer Vorspalte die Ausprägungen von X und in einer Kopfzeile die von Y aus.

		Ausprägung von Y					
		b_1	b_2	\dots	b_j	\dots	b_m
Ausprägung von X	a_1	h_{11}	h_{12}	\dots	h_{1j}	\dots	h_{1m}
	a_2	h_{21}	h_{22}	\dots	h_{2j}	\dots	h_{2m}
	\vdots	\vdots		\ddots			\vdots
	a_i	h_{i1}	h_{i2}	\dots	h_{ij}	\dots	h_{im}
	\vdots	\vdots				\ddots	\vdots
	a_k	h_{k1}	h_{k2}	\dots	h_{kj}	\dots	h_{km}

Tab. 9.1: $(k \times m)$ -Kontingenztafel für absolute Häufigkeiten

Kontingenztafeln werden üblicherweise noch um je eine Zeile und Spalte ergänzt, wobei die zusätzliche *Spalte* bei einer Tabelle für absolute Häufigkeiten die k Zeilensummen

$$h_{i \cdot} := h_{i1} + h_{i2} + \dots + h_{im} = \sum_{j=1}^m h_{ij} \quad i = 1, 2, \dots, k \quad (9.3)$$

und analog bei relativen Häufigkeiten die Summen

$$f_{i \cdot} := f_{i1} + f_{i2} + \dots + f_{im} = \sum_{j=1}^m f_{ij} \quad i = 1, 2, \dots, k \quad (9.4)$$

ausweist (lies: *h-i-Punkt* resp. *f-i-Punkt*). Die Summe (9.3) bzw. (9.4) entspricht der absoluten bzw. relativen Häufigkeit derjenigen Merkmalskombinationen, bei denen X die Ausprägung a_i und Y eine beliebige der m Ausprägungen b_1, \dots, b_m hat. Letzteres bedeutet, dass Y nicht berücksichtigt wird. Die Häufigkeiten $h_{1 \cdot}, h_{2 \cdot}, \dots, h_{k \cdot}$ werden **absolute Randhäufigkeiten** von X genannt, die Häufigkeiten $f_{1 \cdot}, f_{2 \cdot}, \dots, f_{k \cdot}$ **relative Randhäufigkeiten** von X . Durch sie ist die sog. **Randverteilung** von X definiert.

Die zusätzliche *Zeile*, um die man eine Kontingenztafel erweitert, enthält die m Spaltensummen

$$h_{\cdot j} := h_{1j} + h_{2j} + \dots + h_{kj} = \sum_{i=1}^k h_{ij} \quad j = 1, 2, \dots, m \quad (9.5)$$

resp.

$$f_{\cdot j} := f_{1j} + f_{2j} + \dots + f_{kj} = \sum_{i=1}^k f_{ij} \quad j = 1, 2, \dots, m. \quad (9.6)$$



(lies: *h-Punkt-j* bzw. *f-Punkt-j*). Die Häufigkeiten $h_{.1}, h_{.2}, \dots, h_{.m}$ und $f_{.1}, f_{.2}, \dots, f_{.m}$ sind die absoluten Randhäufigkeiten bzw. die relativen Randhäufigkeiten von Y . Sie konstituieren die Randverteilung von Y .

Aufgabe 9.1 Randverteilungen sind nichts anderes als die Häufigkeitsverteilungen der Einzelmerkmale. Die Summe jeder der beiden Randverteilungen besitzt im Falle absoluter Häufigkeiten offenbar den Wert n und im Falle relativer Häufigkeiten den Wert 1.

		Ausprägung von Y						Randverteilung von X
		b_1	b_2	\dots	b_j	\dots	b_m	
Ausprägung von X	a_1	h_{11}	h_{12}	\dots	h_{1j}	\dots	h_{1m}	$h_{1.}$
	a_2	h_{21}	h_{22}	\dots	h_{2j}	\dots	h_{2m}	$h_{2.}$
	\vdots	\vdots	\ddots				\vdots	\vdots
	a_i	h_{i1}	h_{i2}	\dots	h_{ij}	\dots	h_{im}	$h_{i.}$
	\vdots	\vdots			\ddots	\ddots	\vdots	\vdots
	a_k	h_{k1}	h_{k2}	\dots	h_{kj}	\dots	h_{km}	$h_{k.}$
		$h_{.1}$	$h_{.2}$	\dots	$h_{.j}$	\dots	$h_{.m}$	n
		Randverteilung von Y						

Tab. 9.2: Vollständige $(k \times m)$ -Kontingenztafel für absolute Häufigkeiten

Durch die Randverteilungen wird eine Verbindung zwischen uni- und bivariaten Häufigkeitsverteilungen hergestellt. Aus den gemeinsamen Häufigkeiten (9.1) bzw. (9.2) zweier Merkmale X und Y lassen sich stets gemäß (9.3) und (9.5) bzw. (9.4) und (9.6) die Randhäufigkeiten beider Merkmale bestimmen. Die Umkehrung gilt aber nicht, d. h. durch zwei gegebene Randverteilungen kann man i. Allg. nicht eindeutig auf die gemeinsamen Häufigkeiten zurückschließen. Dies ist plausibel, denn die Summenbildung beinhaltet Verdichtung von Information und damit auch Informationsverlust.

Beispiel 9.1: Geschlechtsspezifische Ergebnisse des Politbarometers

In Tabelle 4.2 waren nach Geschlecht differenziert zur „Sonntagsfrage“ des ZDF-Politbarometers vom 8. Dezember 2017 für 1451 befragte Personen wiedergegeben. Tabelle 9.3 zeigt die Daten erneut, nun in Form einer Kontingenztabelle für absolute und darunter für relative Häufigkeiten. Wiedergegeben sind auch die beiden Randverteilungen für das Merkmal „Parteipräferenz“. Vergleicht man in beiden Teiltabellen die Randverteilung von X mit den univariaten Häufigkeitsverteilungen aus Tabelle 4.1, stellt man fest, dass beide übereinstimmen. Die Randverteilung des Merkmals „Parteipräferenz X “ ist also identisch mit der Häufigkeitsverteilung, welche sich bei Verzicht auf die Differenzierung nach Frauen und Männern ergibt.

		Ausprägungen von Y		Randverteilung von X
		σ b_1	φ b_2	
Ausprägungen von X	 a ₁	235	226	461
	 a ₂	219	134	353
	 a ₃	86	105	191
	 a ₄	86	75	161
	 a ₅	88	45	133
	 a ₆	93	30	123
	Sonstige a ₇	17	12	29
		824	627	1 451
		Randverteilung von Y		

		Ausprägungen von Y		Randverteilung von X
		σ b_1	φ b_2	
Ausprägungen von X	 a ₁	0,162	0,156	0,318
	 a ₂	0,151	0,092	0,243
	 a ₃	0,059	0,072	0,132
	 a ₄	0,059	0,052	0,111
	 a ₅	0,061	0,031	0,092
	 a ₆	0,063	0,021	0,085
	Sonstige a ₇	0,012	0,008	0,020
		0,568	0,432	1
		Randverteilung von Y		

Tab. 9.3: (7×2) -Kontingenztafel für absolute und für relative Häufigkeiten

Ein Spezialfall einer Kontingenztabelle ist die **Vierfeldertafel**, die sich für $k = m = 2$ ergibt und in Tabelle 9.4 für den Fall absoluter Häufigkeiten wiedergegeben ist. Vierfeldertafeln werden bei der Untersuchung von Zusammenhängen zwischen zwei Merkmalen verwendet, die je nur zwei Ausprägungen aufweisen. Solche Merkmale nennt man **binäre Merkmale** oder **dichotome Merkmale**. Beispiele sind etwa „Geschlecht“ und

Spezialfall: (2×2) -Kontingenztafel

„Prüfungserfolg“, wenn man beim letztgenannten Merkmal nur zwischen „Bestehen“ und „Nicht-Bestehen“ differenziert.

	b_1	b_2	Zeilensummen
a_1	h_{11}	h_{12}	$h_{1\cdot}$
a_2	h_{21}	h_{22}	$h_{2\cdot}$
Spaltensummen	$h_{\cdot 1}$	$h_{\cdot 2}$	n

Tab. 9.4: Vierfeldertafel für absolute Häufigkeiten

In Zeitungen findet man oft Informationen, die sich zwar in einer Vierfeldertafel zusammenfassen lassen, aber nicht direkt in dieser Form gegeben sind. Die Übertragung der veröffentlichten Information in eine Vierfeldertafel kann dadurch erschwert sein, dass die Informationen sich teilweise auf absolute und teilweise auf relative Häufigkeiten beziehen. In solchen Fällen kann es zweckmäßig sein, anstelle einer Vierfeldertafel zunächst ein **Baumdiagramm** zu entwickeln. Letzteres ist eine Darstellung mit hierarchischer Struktur – analog zu einem Stammbaum mit sich verzweigenden Ästen. Anstelle der Darstellung in Tabelle 9.4 könnte man z. B. das folgende Baumdiagramm wählen:

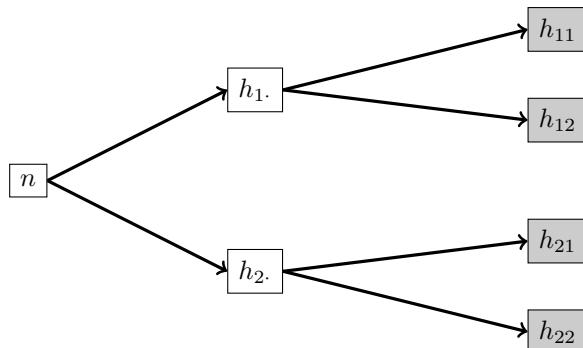


Abb. 9.1: Baumdiagramm als Alternative zu einer Vierfeldertafel

Beispiel 9.2: Baumdiagramm zu amtlichen Bevölkerungsdaten

Einer Pressemitteilung des Statistischen Bundesamtes vom 27. Juni 2019 war zu entnehmen, dass die Gesamtbevölkerung in Deutschland Ende 2018 bei 83,0 Millionen lag. Der Anteil der Personen weiblichen Geschlechts betrug zu diesem Zeitpunkt 51,7 %. Bei den Frauen lag der Anteil der Erwerbstätigen Ende 2018 bei 46,4 %, bei den Männern bei 54,6 %. Kinder sind hier jeweils einbezogen und der Kategorie „nicht erwerbstätig“ zugeordnet. Aus diesen Informationen lässt sich z. B. nicht unmittelbar ablesen, wie groß Ende 2018 die Anzahl der Männer und Frauen ohne Erwerbstätigkeit waren.

Bevor man eine Vierfeldertafel für absolute Häufigkeiten ableitet, ist es hilfreich, die vorstehenden Angaben erst einmal in ein Baumdiagramm zu übertragen. Dieses ist in Abbildung 9.2 wiedergegeben, wobei die oben genannten Informationen durch Fettdruck betont sind.

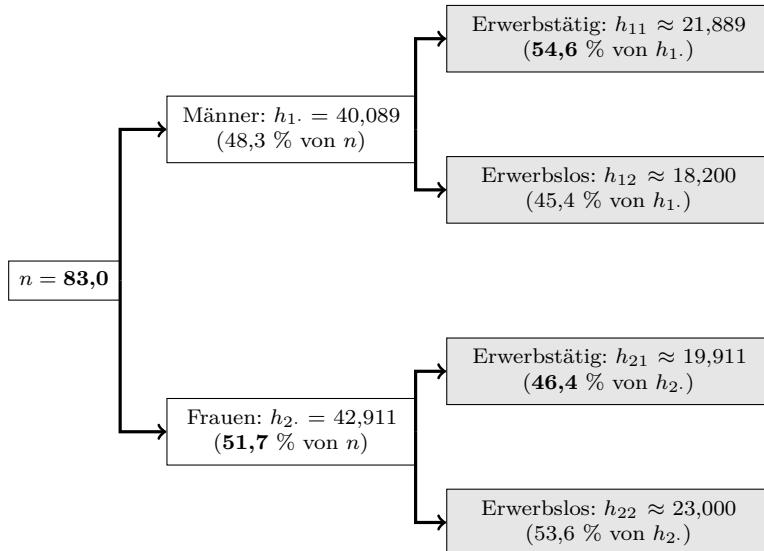


Abb. 9.2: Baumdiagramm zur Erwerbstätigkeit von Männern und Frauen

Die in obigem Baumdiagramm enthaltenen Informationen lassen sich übersichtlich in Form eines Mosaik-Plots veranschaulichen. Die Flächen, die die Anteile an erwerbstätigen und erwerbslosen Frauen repräsentieren, sind im Vergleich zu den korrespondierenden Flächen für die Männer jeweils etwas heller gefärbt.

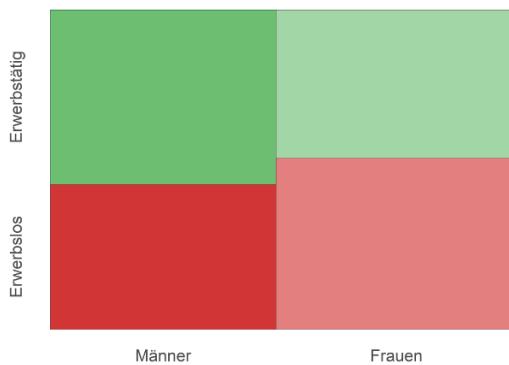


Abb. 9.3: Mosaik-Plot zur Erwerbstätigkeit von Männern und Frauen

Aus den im Baumdiagramm wiedergegebenen Zahlen leitet sich bei Rundung auf volle Hunderter die in Tabelle 9.5 wiedergegebene Vierfeldertafel für absolute Häufigkeiten ab. Die Ausprägungen des Merkmals „Geschlecht“ sind hier – anders als bei der Kontingenztabelle aus Beispiel 9.1 – vertikal aufgelistet.

	Erwerbstätige	Erwerbslose	Zeilensummen
Männer	21 889	18 200	40 089
Frauen	19 911	23 000	42 911
Spaltensummen	41 800	41 200	83 000

Tab. 9.5: Vierfeldertafel für absolute Häufigkeiten

9.2 Empirische Unabhängigkeit diskreter Merkmale

Aus den gemeinsamen Häufigkeiten für zwei Merkmale X und Y kann man noch nicht direkt Aussagen über Zusammenhänge zwischen den Merkmalen ableiten. Aus der Tatsache etwa, dass bei der „Sonntagsfrage“ des ZDF vom 8. Dezember 2017 insgesamt 15,1% der Personen der Stichprobe weibliche Wähler mit SPD-Präferenz waren (134 von 1 451 Personen), lässt sich noch keine Aussage über eine geschlechtsspezifische Präferenz dieser Partei gewinnen. Zur Herleitung einer solchen Aussage benötigt man auch die Information, wie groß die Teilmenge *aller* Befragten in der Stichprobe war, die sich für die SPD aussprach. Diese Information wird durch eine Randhäufigkeit vermittelt (hier: $h_{2\cdot} = 353$). Eine geeignete Verknüpfung der gemeinsamen Häufigkeiten für zwei diskrete Merkmale X und Y mit den Randhäufigkeiten führt zu **bedingten relativen Häufigkeiten**. Diese sind der Ausgangspunkt für die Untersuchung von Zusammenhängen zwischen zwei diskreten Merkmalen.

		Ausprägung von Y						Randverteilung von X
		b_1	b_2	\dots	b_j	\dots	b_m	
Ausprägung von X	a_1	h_{11}	h_{12}	\dots	h_{1j}	\dots	h_{1m}	$h_{1\cdot}$
	a_2	h_{21}	h_{22}	\dots	h_{2j}	\dots	h_{2m}	$h_{2\cdot}$
\vdots	\vdots	\ddots					\vdots	\vdots
a_i	h_{i1}	h_{i2}	\dots	h_{ij}	\dots	h_{im}	$h_{i\cdot}$	
\vdots	\vdots				\ddots		\vdots	\vdots
a_k	h_{k1}	h_{k2}	\dots	h_{kj}	\dots	h_{km}	$h_{k\cdot}$	
	$h_{\cdot 1}$	$h_{\cdot 2}$	\dots	$h_{\cdot j}$	\dots	$h_{\cdot m}$	n	
Randverteilung von Y								

Tab. 9.6: Absolute Häufigkeiten für die Ausprägungen von Y unter der Bedingung $X = a_i$

Um das Konzept der bedingten Häufigkeiten verständlich zu machen, ist in Tabelle 9.6 nochmals eine $(k \times m)$ -Kontingenztafel veranschaulicht, nun aber mit Hervorhebung der i -ten Zeile (Betonung durch Umrahmung). Man findet im hervorgehobenen Bereich neben der Angabe der Ausprägung a_i für das Merkmal X die m gemeinsamen absoluten Häufigkeiten $h_{ij} = h(a_i, b_j)$ beider Merkmale, welche der Bedingung $X = a_i$ genügen. Am Ende des betonten Bereichs steht die durch Aufsummieren der m genannten Häufigkeiten resultierende Randhäufigkeit $h_{i \cdot}$ von X .

Bedingte
Häufigkeits-
verteilung
für Y

Dividiert man nun jedes der m Elemente $h_{i1}, h_{i2}, \dots, h_{im}$ durch die Randhäufigkeit $h_{i \cdot}$, so erhält man die relativen Häufigkeiten für das Auftreten der Ausprägungen b_1, b_2, \dots, b_m bei Gültigkeit von $X = a_i$. Das Ergebnis sind bedingte relative Häufigkeiten für Y . Wenn man diese mit $f_Y(b_j|a_i)$ abkürzt, gilt also

$$f_Y(b_j|a_i) := \frac{h_{ij}}{h_{i \cdot}} \quad j = 1, 2, \dots, m. \quad (9.7)$$

Diese m bedingten relativen Häufigkeiten definieren die **bedingte Häufigkeitsverteilung** für Y unter der Bedingung $X = a_i$.

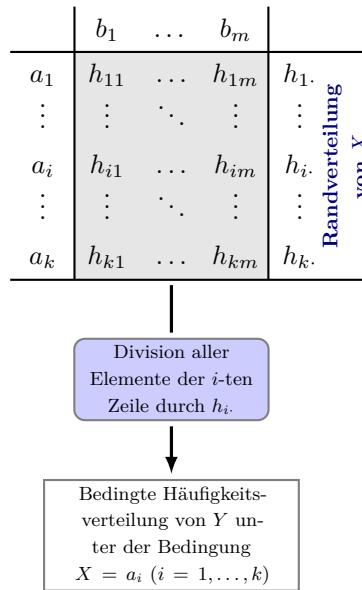


Abb. 9.4: Bestimmung der bedingten Häufigkeitsverteilung für Y

Analog kann man, wie in Tabelle 9.7 illustriert, in der $(k \times m)$ -Kontingenztafel die j -te Spalte hervorheben. In der Kopfzeile steht dann für Y die Ausprägung b_j . Darunter folgen die k gemeinsamen absoluten Häufigkeiten $h_{1j}, h_{2j}, \dots, h_{kj}$ der Merkmale X und Y , bei denen bezüglich Y die Bedingung $Y = b_j$ zutrifft. Am Ende des betonten Bereichs steht die durch Aufsummieren der k genannten Häufigkeiten errechnete Randhäufigkeit $h_{\cdot j}$ von Y .

Bedingte
Häufigkeits-
verteilung
für X

		Ausprägung von Y							Randverteilung von X
		b_1	b_2	\dots	b_j	\dots	b_m		
Ausprägung von X	a_1	h_{11}	h_{12}	\dots	h_{1j}	\dots	h_{1m}	$h_{1\cdot}$	
	a_2	h_{21}	h_{22}	\dots	h_{2j}	\dots	h_{2m}	$h_{2\cdot}$	
	\vdots	\vdots	\ddots				\vdots	\vdots	
	a_i	h_{i1}	h_{i2}	\dots	h_{ij}	\dots	h_{im}	$h_{i\cdot}$	
	\vdots	\vdots				\ddots	\vdots	\vdots	
	a_k	h_{k1}	h_{k2}	\dots	h_{kj}	\dots	h_{km}	$h_{k\cdot}$	
		$h_{\cdot 1}$	$h_{\cdot 2}$	\dots	$h_{\cdot j}$	\dots	$h_{\cdot m}$	n	
Randverteilung von Y									

Tab. 9.7: Absolute Häufigkeiten für die Ausprägungen von X unter der Bedingung $Y = b_j$

Teilt man jedes der k Elemente $h_{1j}, h_{2j}, \dots, h_{kj}$ durch die Randhäufigkeit $h_{\cdot j}$, so erhält man die relativen Häufigkeiten für das Auftreten der Ausprägungen a_1, a_2, \dots, a_k unter der Bedingung $Y = b_j$. Es resultieren **bedingte relative Häufigkeiten** für X unter der Bedingung $Y = b_j$. Kürzt man diese mit $f_X(a_i|b_j)$ ab, hat man

$$f_X(a_i|b_j) := \frac{h_{ij}}{h_{\cdot j}} \quad i = 1, 2, \dots, k. \quad (9.8)$$

Diese k bedingten relativen Häufigkeiten konstituieren die **bedingte Häufigkeitsverteilung** für X unter der Bedingung $Y = b_j$.

	b_1	\dots	b_j	\dots	b_m	
a_1	h_{11}	\dots	h_{1j}	\dots	h_{1m}	
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	
a_k	h_{k1}	\dots	h_{kj}	\dots	h_{km}	
	$h_{\cdot 1}$	\dots	$h_{\cdot j}$	\dots	$h_{\cdot m}$	

Randverteilung von Y

Division aller
Elemente der j -ten
Spalte durch $h_{\cdot j}$

Bedingte Häufigkeits-
verteilung von X un-
ter der Bedingung
 $Y = b_j$ ($j = 1, \dots, m$)

Abb. 9.5: Bestimmung der bedingten Häufigkeitsverteilung für X

Beispiel 9.3: Bedingte Häufigkeiten beim ZDF-Politbarometer

Die Bestimmung und Interpretation bedingter relativer Häufigkeiten sei anhand der Daten zum ZDF-Politbarometer vom 8. Dezember 2017 illustriert.

		Ausprägungen von Y		Randverteilung von X
		σ b_1	φ b_2	
Ausprägungen von X	 a ₁	235	226	461
	 a ₂	219	134	353
	 a ₃	86	105	191
	 a ₄	86	75	161
	 a ₅	88	45	133
	 a ₆	93	30	123
	Sonstige a ₇	17	12	29
		824	627	1 451
Randverteilung von Y				
Ausprägungen von X	 a ₁	235	226	461
	 a ₂	219	134	353
	 a ₃	86	105	191
	 a ₄	86	75	161
	 a ₅	88	45	133
	 a ₆	93	30	123
	Sonstige a ₇	17	12	29
		824	627	1 451
Randverteilung von Y				

Tab. 9.8: Berechnung bedingter Häufigkeiten (ZDF-Politbarometer)

Bei diesem Datensatz ist z. B. die Frage von Interesse, ob zwischen der Parteidräferenz X und dem Geschlecht Y der befragten Personen in der Stichprobe ein Zusammenhang besteht. Tabelle 9.8 zeigt zweifach die Kontingenztafel für

absolute Häufigkeiten aus Beispiel 9.1. In der oberen Fassung der Tabelle sind die Häufigkeiten hervorgehoben, die sich auf die Wahlpräferenzen der Männer in der Stichprobe beziehen (Hervorhebung der ersten Spalte), während die untere Fassung diejenigen Häufigkeiten betont, die sich auf die Ausprägung $X = a_2$ beziehen (Hervorhebung der zweiten Zeile der Kontingenztafel). Die obere Version von Tabelle 9.8 betont, dass von den $n = 1\,451$ Personen der Stichprobe insgesamt 824 Befragte männlich waren und dass innerhalb dieser Teilstichprobe 235 Männer die CDU/CSU, 219 die SPD, 86 die Grünen, 86 die Linken, 88 die FDP, 93 die AfD und 17 sonstige Parteien favorisiert hatten.

Die bedingten relativen Häufigkeiten $f_X(a_1|b_1)$, $f_X(a_2|b_1)$, \dots , $f_X(a_6|b_1)$, die die bedingte Häufigkeitsverteilung für X unter der Bedingung $Y = b_1$ definieren, bestimmen sich nach (9.8) als

$$\begin{aligned} f_X(a_1|b_1) &= \frac{h_{11}}{h_{\cdot 1}} = \frac{235}{824} \approx 0,285 & f_X(a_2|b_1) &= \frac{h_{21}}{h_{\cdot 1}} = \frac{219}{824} \approx 0,266 \\ f_X(a_3|b_1) &= \frac{h_{31}}{h_{\cdot 1}} = \frac{86}{824} \approx 0,104 & f_X(a_4|b_1) &= \frac{h_{41}}{h_{\cdot 1}} = \frac{86}{824} \approx 0,104 \\ f_X(a_5|b_1) &= \frac{h_{51}}{h_{\cdot 1}} = \frac{88}{824} \approx 0,107 & f_X(a_6|b_1) &= \frac{h_{61}}{h_{\cdot 1}} = \frac{93}{824} \approx 0,113 \\ f_X(a_7|b_1) &= \frac{h_{71}}{h_{\cdot 1}} = \frac{17}{824} \approx 0,021. \end{aligned}$$



Der Wert 0,285 sagt z. B. aus, dass von den Personen in der Stichprobe,

Aufgabe 9.2

die männlichen Geschlechts waren, ca. 28,5% bei der „Sonntagsfrage“ vom 8.

Dezember 2017 die CDU/CSU favorisiert hatten.

Im unteren Teil von Tabelle 9.8 ist hervorgehoben, dass von den $n = 1\,451$ Personen der Stichprobe 353 Personen die SPD favorisierten, nämlich 219 Männer ($Y = b_1$) und 134 Frauen ($Y = b_2$). Die bedingten relativen Häufigkeiten $f_Y(b_1|a_2)$ und $f_Y(b_2|a_2)$, die die bedingte Häufigkeitsverteilung für das Merkmal Y unter der Bedingung $X = a_2$ repräsentieren, errechnen sich gemäß (9.7) als

$$f_Y(b_1|a_2) = \frac{h_{21}}{h_{\cdot 2}} = \frac{219}{353} \approx 0,620 \quad f_Y(b_2|a_2) = \frac{h_{22}}{h_{\cdot 2}} = \frac{134}{353} \approx 0,380.$$

Das Ergebnis 0,380 beinhaltet, dass von den Personen in der Stichprobe, die sich für die SPD entschieden hatten, 38,0% weiblich waren.

Wann liegt kein Zusammenhang vor?

Anhand der bedingten Häufigkeitsverteilungen lässt sich konkretisieren, wann man von einem *fehlenden* Zusammenhang zweier Merkmale X und Y spricht, d. h. von Unabhängigkeit der Merkmale. Intuitiv wird man **Unabhängigkeit** von X und Y als gegeben ansehen, wenn die Ausprägung eines Merkmals keinen Einfluss auf die Ausprägung des anderen Merkmals hat. Dies aber bedeutet, dass eine bedingte Häufigkeitsverteilung für ein Merkmal nicht davon abhängt, welche Merkmalsausprägung für das andere Merkmal vorausgesetzt wird. So dürfte die bedingte Häufigkeitsverteilung für X unter der Bedingung $Y = b_j$ nicht davon abhängen, welche der m Ausprägungen b_1, b_2, \dots, b_m als Bedingung gewählt wird,

d. h. die m bedingten Häufigkeitsverteilungen $f_X(a_1|b_j)$, $f_X(a_2|b_j)$, \dots , $f_X(a_k|b_j)$ müssten übereinstimmen ($j = 1, 2, \dots, m$). Insbesondere müssten die i -ten Elemente dieser m bedingten Verteilungen identisch sein, d. h. es würde bei Unabhängigkeit gelten

$$f_X(a_i|b_1) = f_X(a_i|b_2) = \dots = f_X(a_i|b_m).$$

Äquivalent ist wegen (9.8) die Darstellung

$$\frac{h_{i1}}{h_{.1}} = \frac{h_{i2}}{h_{.2}} = \dots = \frac{h_{im}}{h_{.m}}.$$

Wenn in der letzten Gleichung die m Brüche alle identisch sind, muss auch der Quotient aus der Summe aller m Zähler und der Summe aller m Nenner übereinstimmen. Die erstgenannte Summe ist offenbar die Randhäufigkeit $h_{i.}$, während die zweite Summe mit dem Stichprobenumfang n übereinstimmt (vgl. Tabelle 9.6). Es gilt also bei Unabhängigkeit von X und Y für jede der gemeinsamen Häufigkeiten h_{ij} der Kontingenztafel

$$\frac{h_{ij}}{h_{.j}} = \frac{h_{i.}}{n}.$$

Löst man nach h_{ij} auf, folgt, dass h_{ij} bei Unabhängigkeit der Merkmale mit $\frac{h_{i.} \cdot h_{.j}}{n}$ übereinstimmt. Für die bei empirischer Unabhängigkeit zu erwartenden Werte für die gemeinsamen Häufigkeiten von X und Y wird im Folgenden die Abkürzung

$$\tilde{h}_{ij} := \frac{h_{i.} \cdot h_{.j}}{n} \quad (9.9)$$

(lies: *h-Schlange-i-j*) verwendet. Empirische Unabhängigkeit bzw. Abhängigkeit von X und Y bedeutet dann, dass für die Häufigkeiten h_{ij} der $(k \times m)$ -Kontingenztafel

$$h_{ij} \begin{cases} = \tilde{h}_{ij} & \text{bei empirischer Unabhängigkeit der Merkmale} \\ \neq \tilde{h}_{ij} & \text{bei empirischer Abhängigkeit der Merkmale} \end{cases} \quad (9.10)$$

Formale Definition
der Unabhängigkeit
zweier Merkmale

gilt. Zwei Merkmale X und Y , deren gemeinsame Häufigkeitsverteilung durch Tabelle 9.2 gegeben ist, sind also genau dann unabhängig, wenn für jedes der $k \cdot m$ Elemente h_{ij} der Kontingenztafel $h_{ij} = \tilde{h}_{ij}$ ist mit \tilde{h}_{ij} aus (9.9). Da sich eine solche Unabhängigkeitsaussage aus Daten und nicht aus Wahrscheinlichkeitsmodellen ableitet, spricht man auch präziser von **empirischer Unabhängigkeit** der betreffenden Merkmale. Die bei Unabhängigkeit zu erwartenden Werte \tilde{h}_{ij} für die gemeinsamen Häufigkeiten sind nicht notwendigerweise ganzzahlig.

Die Aussage (9.10) impliziert, dass bei Unabhängigkeit zweier Merkmale X und Y die gesamte Information über die gemeinsame Häufigkeitsverteilung bereits in den Randverteilungen steckt. Wenn zwischen den Merkmalen hingegen ein Zusammenhang besteht, gilt dies nicht und es gibt dann von Null verschiedene Differenzen $h_{ij} - \tilde{h}_{ij}$. Diese sind der Ausgangspunkt für die Konstruktion von Zusammenhangsmaßen für nominalskalierte Merkmale (s. Abschnitt 10.1).

Beispiel 9.4: Parteipräferenz und Geschlecht

Es werde erneut der obere Teil von Tabelle 9.3 betrachtet. Dieser zeigte die Ergebnisse des Politbarometers vom 8. Dezember 2017 in Form der gemeinsamen absoluten Häufigkeiten für die Merkmale „Parteipräferenz X “ und „Geschlecht Y “. Interessant ist hier die Frage, ob sich das Wählerverhalten von Frauen und Männern unterscheidet. Um eine Aussage über einen möglichen Zusammenhang zwischen den beiden nominalskalierten Merkmalen X und Y zu gewinnen, hat man die in der Kontingenztabelle ausgewiesenen Häufigkeiten h_{ij} mit den nach (9.9) zu errechnenden Werten zu vergleichen, die bei empirischer Unabhängigkeit gelten müssten. Die Häufigkeiten $h_{11} = 235$ und $h_{12} = 226$ aus dem oberen Teil von Tabelle 9.3 sind also z. B. zu vergleichen mit

$$\tilde{h}_{11} = \frac{h_{1\cdot} \cdot h_{\cdot 1}}{n} = \frac{461 \cdot 824}{1451} \approx 261,8 \quad \tilde{h}_{12} = \frac{h_{1\cdot} \cdot h_{\cdot 2}}{n} = \frac{461 \cdot 627}{1451} \approx 199,2.$$

Die anderen 12 Werte \tilde{h}_{ij} sind analog zu bestimmen. In Tabelle 9.9 sind alle 14 beobachteten absoluten Häufigkeiten h_{ij} (auf grauem Raster) und die bei Unabhängigkeit zu erwartenden fiktiven Werte \tilde{h}_{ij} nebeneinander gestellt.

		Ausprägungen von Y		Ausprägungen von Y	
		σ b_1	φ b_2	σ b_1	φ b_2
Ausprägungen von X	 a ₁	235	226	261,8	199,2
	 a ₂	219	134	200,5	152,5
	 a ₃	86	105	69,6	82,5
	 a ₄	86	75	91,4	108,5
	 a ₅	88	45	75,5	57,5
	 a ₆	93	30	69,8	53,2
	Sonstige a ₇	17	12	16,5	12,5

Tab. 9.9: Absolute Häufigkeiten (Politbarometer) – beobachtete Werte h_{ij} (gerasterter Teil) und Werte \tilde{h}_{ij} bei empirischer Unabhängigkeit

Man erkennt, dass die sich entsprechenden Werte h_{ij} und \tilde{h}_{ij} in nicht vernachlässigbarem Umfang differieren – es ist z. B. $h_{32} = 105$ und $\tilde{h}_{32} = 82,5$. Die Daten sprechen *nicht* für eine empirische Unabhängigkeit der beiden Merkmale „Parteipräferenz X“ und „Geschlecht Y“.

Dass sich die Parteipräferenzen von Männer und Frauen unterscheiden, sieht man besonders gut anhand der nachstehenden Abbildung 9.6, die zwei Mosaik-Plots zeigt. Der linke Plot deckt sich mit Abbildung 4.6 und veranschaulicht die reale, im gerasterten Teil von Tabelle 9.9 wiedergegebenen Häufigkeitsverteilungen für die Parteipräferenzen von Männern und Frauen. Der rechte Plot basiert hingegen auf den Daten im linken Teil von Tabelle 9.9, die sich bei Unabhängigkeit der Parteipräferenzen vom Geschlecht ergeben würden.

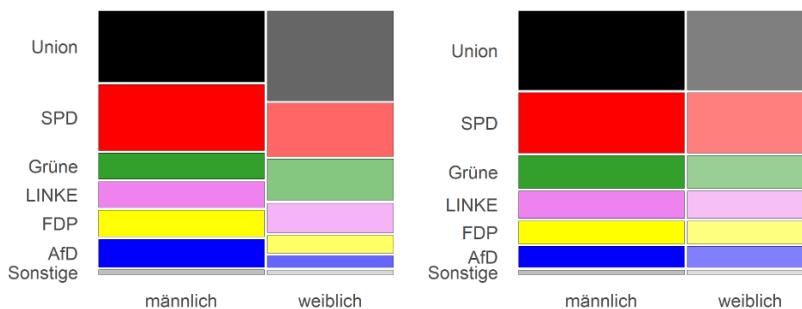


Abb. 9.6: Mosaik-Plots für geschlechtsspezifische Parteipräferenzen gemäß Politbarometer-Daten und bei empirischer Unabhängigkeit

Bivariate empirische Verteilungen für diskrete oder gruppierte stetige Merkmale lassen sich mit gestapelten Säulen- oder Balkendiagrammen visualisieren. Eine andere Möglichkeit besteht in der Verwendung neben- oder hintereinander gestellter Säulen. Letzteres führt zu einem **Doppel-Säulendiagramm** bzw. zu einem **3D-Säulendiagramm**. Abbildung 9.7 visualisiert auf der Basis von Eurostat-Daten für 2018 anhand eines Doppel-Säulendiagramms eine (4×2) -Kontingenztafel für relative Häufigkeiten – ausgewiesen in Prozent. Die Kontingenztafel bezieht sich auf das stetige Merkmal „Alter X“ (zu 4 Altersklassen gruppiert) und das diskrete Merkmal „Land Y“.

Visualisierung
empirischer
Verteilungen zweier
diskreter Merkmale

Das erstgenannte Merkmal ist die durch Bildung von vier Klassen diskretisierte demografische Schlüsselvariable „Alter X“ (Bildung der Altersgruppen „0 – 14 Jahre“, „15 – 24 Jahre“, „25 – 64 Jahre“ und „65 und mehr Jahre“), während für das Merkmal „Land Y“ hier nur zwei Ausprägungen herangezogen werden, nämlich „Deutschland“ und „Irland“. Man erkennt deutliche Unterschiede hinsichtlich der Bevölkerungsstrukturen beider Länder. So entnimmt man der neben der Grafik platzierten Kontingenztabelle z. B., dass der Anteil der unter 15-jährigen mit 13,4%

in Deutschland viel niedriger (Irland: 20,8%) und der Anteil der über 64-jährigen mit 21,7% viel höher lag (Irland: 13,9%).

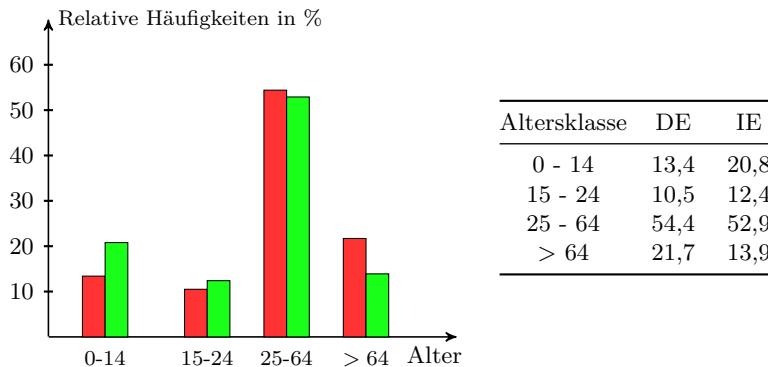


Abb. 9.7: Bevölkerungsstrukturen 2018 in Deutschland und Irland
(Doppel-Säulendiagramm; rote Teilbalken: Deutschland)

Bei allen oben genannten Varianten für die grafische Darstellung bivariater empirischer Verteilungen für diskrete oder gruppierte stetige Merkmale lassen sich die absoluten oder relativen Häufigkeiten, welche die Länge der einzelnen Säulen oder Säulenabschnitte definieren, bei Bedarf auch direkt in der Grafik ausweisen. Dies kann sinnvoll sein, wenn sich mehrere Säulen oder Säulenabschnitte hinsichtlich ihrer Länge kaum unterscheiden und die numerischen Werte nicht zusätzlich tabellarisch ausgewiesen sind.

9.3 Empirische Verteilungen stetiger Merkmale

Wenn man an n Untersuchungseinheiten die Ausprägungen zweier *stetiger* Merkmale X und Y ermittelt, wird man bei der resultierenden **bivariate Urliste** $(x_1, y_1), \dots, (x_n, y_n)$ selten beobachten, dass Merkmalspaare (x_i, y_i) mehrfach auftreten, d. h. die Häufigkeit beträgt für jedes Merkmalspaar meist 1. Grundsätzlich kann man natürlich die Merkmale durch Gruppierung diskretisieren und dann die in Abschnitt 8.1 behandelten Ansätze heranziehen. Gruppierung stetiger Merkmale ist aber mit einem Informationsverlust verbunden. Dieser kann bei sehr großen Datensätzen vertretbar sein, wenn die Aggregation von Information zu mehr Übersichtlichkeit führt.

- | | |
|---|---|
| Visualisierung
empirischer
Verteilungen bei
stetigen Merkmalen | Ein Datensatz für zwei stetige Merkmale wird üblicherweise in einem zweidimensionalen Koordinatensystem dargestellt. In diesem Koordinatensystem werden die Merkmalspaare $(x_1, y_1), \dots, (x_n, y_n)$ durch Punkte repräsentiert. Dabei resultiert ein Streudiagramm . Abbildung 9.8 zeigt zwei solche Streudiagramme. Der linke Teil der Abbildung legt einen Zusammenhang zwischen den Merkmalen X und Y nahe, während das |
|---|---|

rechte Streudiagramm diesen Eindruck nicht vermittelt. Ein Streudiagramm liefert also einen visuellen Anhaltspunkt für das Bestehen oder Fehlen eines empirischen Zusammenhangs zwischen zwei stetigen Merkmalen. Zur Quantifizierung des visuellen Eindrucks benötigt man ein Zusammenhangsmaß. Ein solches wird in Abschnitt 10.2 abgeleitet.

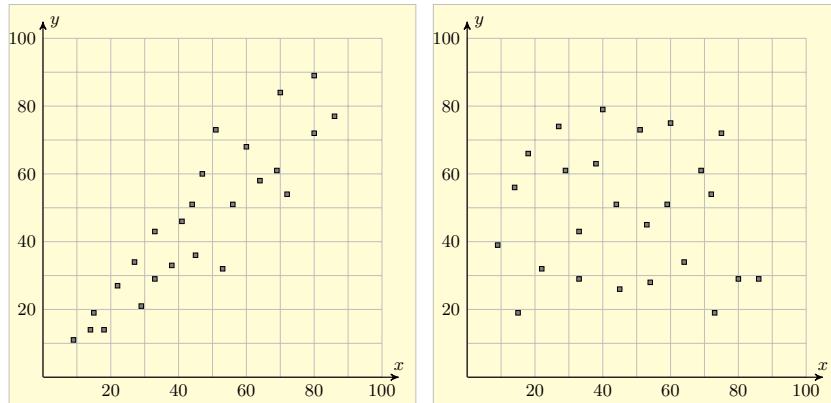


Abb. 9.8: Zwei Streudiagramme

Wenn in dem betrachteten Datensatz $(x_1, y_1), \dots, (x_n, y_n)$ Punkte mehrfach auftreten, wird dies in einem klassischen Streudiagramm nicht sichtbar. Man kann aber die einzelnen Punkte durch Kreise darstellen, deren Flächeninhalt davon abhängt, wie oft der jeweilige Punkt im Datensatz auftritt. Die resultierende Grafik wird als **Blasendiagramm** (engl.: *bubble chart*) bezeichnet. Blasendiagramme lassen sich aber auch für bivariate Datensätze $(x_1, y_1), \dots, (x_n, y_n)$ heranziehen, bei denen kein Wert mehrfach auftritt. Der Flächeninhalt der einzelnen Blasen kann dabei durch ein drittes stetiges Merkmal Z repräsentiert sein.

Blasendiagramme lassen sich ebenso zur Veranschaulichung univariater Häufigkeitsverteilungen verwenden. Der Flächeninhalt der Blasen ist hier durch die Werte des Datensatzes definiert. Die Position der Kreismittelpunkte ist in dem Falle nicht exakt festgelegt. Das Stockholmer Institut für Friedensforschung und Rüstungskontrolle **SIPRI** visualisiert z. B. univariate Daten zum Ländervergleich von Militärausgaben anhand eines Blasendiagramms und bindet dabei eine Weltkarte ein.

Wenn man ein Blasendiagramm für mehrere aufeinanderfolgende Jahre t_0, t_1, \dots erstellt und diese in einer Animation hintereinander schaltet, wird die Zeit t als vierte Variable in die Visualisierung einbezogen. Ein derartiges dynamisches Blasendiagramm ist über das nebenstehende Icon aktivierbar. Das Diagramm zeigt die Lebenserwartung Y in verschiedenen Ländern in Abhängigkeit von Pro-Kopf-Einkommen X (BIP / Kopf). Die Bevölkerungsstärke Z der Länder ist durch die Flächeninhalte der Blasen repräsentiert. Die Variable „Zeit t “ variiert hier von $t = 1800$ bis



Dynamisches
Blasendiagramm
„BIP pro Kopf und
Lebenserwartung“
(Gapminder)

$t = 2013$. Zur Erstellung dynamischer Blasendiagramme kann man z. B. *JMP* oder *R* heranziehen.

Beispiel 9.5: Portfolio-Analyse anhand eines Blasendiagramms

Zur Verwendung von Blasendiagrammen sei ein Beispiel aus dem Bereich des Marketings angeführt. Es bezeichne X den Marktanteil von vier konkurrierenden 10-Zoll-Tablets A, B, C und D eines Herstellers im Jahr 2020. Die vier Produkte unterscheiden sich bezüglich des Verkaufspreises und der Ausstattungsmerkmale, weisen insbesondere unterschiedliche Speicherkapazitäten auf. Ferner sei Y die Veränderungsrate der Marktanteile gegenüber dem Vorjahr und Z der mit den vier Produkten in 2020 erzielte Umsatz (in Millionen Euro).

	A	B	C	D
X	3,1	5,4	9,8	16,8
Y	6,9	3,2	6,8	1,9
Z	30,5	52,5	86,0	145,4

Tab. 9.10: Marktanteil, Veränderungsrate und Umsatz für vier Tablets

Stellt man die Werte $(x_1, y_1), \dots, (x_4, y_4)$ als Punkte dar und zeichnet um die Punkte Kreise, deren Fläche zum Wert des Merkmals „Umsatz Z “ proportional ist, resultiert ein Blasendiagramm. Das in Abbildung 9.9 wiedergegebene Diagramm kann für die Sortimentsplanung des kommenden Jahres herangezogen werden. Man sieht z. B., dass das Tablet C im Jahr 2020 hinsichtlich aller drei Merkmale besser als Tablet B abschnitt. Tablet C ist auch erfolgreicher als A, weil es sich bei etwa gleichen Werten für Y bezüglich der Merkmale X und Y besser behauptete. Zwischen den Produkten A und B gibt es hingegen keine eindeutige Rangordnung – A ist bezüglich Marktanteil X und Umsatz Z schlechter, weist aber eine höhere Veränderungsrate Y auf. Eine analoge Aussage gilt für den Vergleich von C und D.

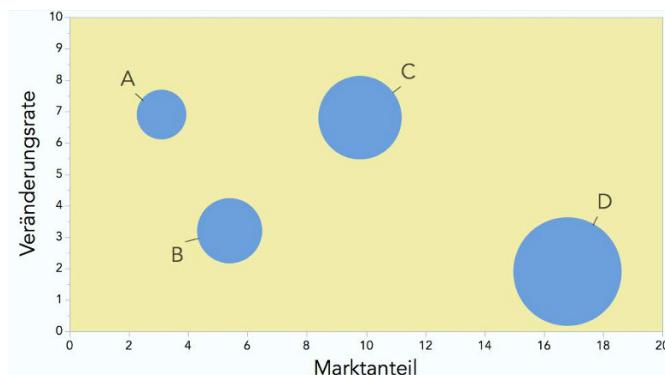


Abb. 9.9: Analyse eines Produktportfolios anhand eines Blasendiagramms



10 Zusammenhangsmaße



Vorschau auf
das Kapitel

Als ein Zusammenhangsmaß für zwei diskrete Merkmale X und Y mit k bzw. m Ausprägungen wird zunächst der χ^2 -Koeffizient vorgestellt, dessen obere Schranke vom Umfang n des Datensatzes und auch von der Anzahl k und m der Zeilen bzw. Spalten einer Kontingenztabelle abhängt. Aus diesem Maß wird der sog. Phi-Koeffizient abgeleitet, dessen obere Schranke nur noch von k und m abhängt. Ein auch nicht mehr von der Dimension der Kontingenztafel abhängendes normiertes Zusammenhangsmaß ist Cramèr's V , dessen Berechnung anhand von Daten des ZDF-Politbarometers illustriert wird.

Zur Messung des Zusammenhangs zwischen zwei metrisch skalierten Merkmalen werden die empirische Kovarianz als nicht-normiertes und der Korrelationskoeffizient r nach Bravais-Pearson als normiertes Maß vorgestellt. Die Formel für r lässt sich auch auf die Ränge von ordinalskalierten Daten beziehen – dies führt zum Rangkorrelationskoeffizienten nach Spearman.

10.1 Nominalskalierte Merkmale

In Abschnitt 9.2 wurde mit (9.10) formalisiert, was unter einem fehlenden Zusammenhang für zwei nominalskalierte Merkmale X und Y zu verstehen ist, also unter **empirischer Unabhängigkeit** dieser Merkmale. Sie wurde als gegeben angenommen, wenn beim Vergleich der in einer $(k \times m)$ -Kontingenztabelle zusammengefassten gemeinsamen Häufigkeiten h_{ij} für diese Merkmale mit den bei Unabhängigkeit zu erwartenden Häufigkeiten \tilde{h}_{ij} aus (9.9) eine durchgehende Übereinstimmung festgestellt wird. Wenn keine Übereinstimmung festgestellt wird, also ein empirischer Zusammenhang vorliegt, will man diesen anhand eines geeigneten Zusammenhangsmaßes quantifizieren. Es liegt nahe, die $k \times m$ Differenzen $h_{ij} - \tilde{h}_{ij}$ für die Konstruktion eines Maßes heranzuziehen. Da diese Differenzen sowohl positiv als auch negativ sein können, sich also bei Aufsummierung ganz oder teilweise zu neutralisieren vermögen, verwendet man die Summe der *quadrierten* Differenzen. Diese werden auf \tilde{h}_{ij} bezogen, d.h. man bildet die Summe der $k \times m$ Terme $\frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$.

Wenn man diese Terme analog zu Tabelle 9.1 (innerer Bereich) in einer Tabelle mit k Zeilen und m Spalten anordnet, kann man die genannte Summe errechnen, indem man z. B. zuerst die Terme $\frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$ in jeder der k Zeilen addiert und dann die k Zeilensummen aufsummiert.

Ein nicht-normiertes
Zusammenhangsmaß

Die Summe der normierten Differenzterme in der i -ten Zeile (i fest) ist gegeben durch

$$\sum_{j=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = \frac{(h_{i1} - \tilde{h}_{i1})^2}{\tilde{h}_{i1}} + \frac{(h_{i2} - \tilde{h}_{i2})^2}{\tilde{h}_{i2}} + \dots + \frac{(h_{im} - \tilde{h}_{im})^2}{\tilde{h}_{im}}.$$

Summiert man nun noch die k Zeilensummen auf, erhält man einen Term mit zwei Summenzeichen (Doppelsumme), der mit χ^2 (lies: *Chi-Quadrat*) abgekürzt und **χ^2 -Koeffizient** genannt wird:¹

$$\chi^2 := \sum_{i=1}^k \sum_{j=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}. \quad (10.1)$$

Der χ^2 -Koeffizient ist ein **Zusammenhangsmaß** für zwei nominalskaulierte Merkmale, das nach Konstruktion Null ist, wenn die Merkmale empirisch unabhängig sind. Bei einem schwachen Merkmalszusammenhang nimmt (10.1) kleine und bei starkem Zusammenhang große Werte an. Das Maß χ^2 kann aber nicht beliebig groß werden, d. h. es ist nach oben beschränkt. Die obere Schranke χ^2_{\max} hängt sowohl vom Umfang n des Datensatzes ab wie auch vom kleineren der beiden Werte k und m , die die Dimension der Kontingenztabelle festlegen. Bezeichnet man das Minimum der beiden Werte k und m mit M , so kann man zeigen (vgl. etwa TOUTENBURG / HEUMANN (2009, Abschnitt 4.2)), dass

$$0 \leq \chi^2 \leq \chi^2_{\max} = n \cdot (M - 1) \quad M := \min(k; m). \quad (10.2)$$

Wenn der χ^2 -Koeffizient den Wert χ^2_{\max} annimmt, spricht man von *vollständiger Abhängigkeit* der beiden Merkmale.

Herleitung eines normierten Zusammenhangsmaßes Wenn man zwei Kontingenztafeln gleicher Dimension hat, so erlaubt der χ^2 -Koeffizient nur dann den Vergleich der Stärke der Merkmalszusammenhänge in beiden Tabellen, wenn auch der Umfang n der in die Kontingenztafeln eingehenden Häufigkeiten übereinstimmt. Der χ^2 -Koeffizient ist daher für die Praxis noch nicht sonderlich geeignet. Ein aus (10.1) abgeleitetes Zusammenhangsmaß, dessen Wert nicht mehr von n abhängt, ist der durch

$$\Phi := \sqrt{\frac{\chi^2}{n}} \quad (10.3)$$

definierte **Phi-Koeffizient**. Auch dieses Maß ist nicht-negativ und nimmt bei einem schwachen Merkmalszusammenhang kleine Werte an. Bei einem starken Zusammenhang ist der Φ -Koeffizient offenbar durch $\sqrt{M - 1}$ nach



Harald CRAMÉR

¹Das Zusammenhangsmaß (10.1) wird in der induktiven Statistik u. a. verwendet, um Hypothesen über Merkmalszusammenhänge zu testen (sog. χ^2 -Unabhängigkeitstest; vgl. Abschnitt 16.8).

oben beschränkt, d. h. es gilt mit M aus (10.2):

$$0 \leq \Phi \leq \Phi_{\max} := \sqrt{M - 1}. \quad (10.4)$$

Der maximale Wert Φ_{\max} , den der Phi-Koeffizient bei vollständiger Abhängigkeit der beiden Merkmale annimmt, hängt zwar nicht mehr von n ab, wohl aber immer noch von M , also von der Dimension der Kontingenztabelle. Auch mit dem Phi-Koeffizienten kann man also die Stärke von Merkmalszusammenhängen bei Kontingenztabellen unterschiedlicher Dimension noch nicht direkt vergleichen. Diesen Nachteil vermeidet der auf den schwedischen Mathematiker und Statistiker Harald CRAMÉR (1893 – 1985) zurückgehende Kontingenzkoeffizient

$$V := \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n \cdot (M - 1)}}. \quad (10.5)$$

Das Zusammenhangsmaß von CRAMÉR, häufig kurz als **Cramér's V** angesprochen, nimmt stets Werte zwischen 0 und 1 an, ist also ein normiertes Zusammenhangsmaß:

$$0 \leq V \leq 1. \quad (10.6)$$



Interaktives Objekt
„Cramér's V“

Mit (10.6) lässt sich die Stärke von Merkmalszusammenhängen bei Kontingenztabellen beliebiger Dimension direkt vergleichen. Aussagen über die Richtung eines Zusammenhangs sind allerdings bei allen hier vorgestellten Zusammenhangsmaßen nicht möglich.

Beispiel 10.1: Parteipräferenz und Geschlecht

Auf der Basis der (7×2) -Kontingenztabelle mit den Daten des ZDF-Politbarometers vom 8. Dezember 2017 wurde in Beispiel 9.4 festgestellt (vgl. Tabelle 9.9), dass man von einem Zusammenhang zwischen den beiden nominalskalierten Merkmalen „Parteipräferenz X “ und „Geschlecht Y “ ausgehen muss. Die Stärke des Zusammenhangs wurde aber dort noch nicht quantifiziert.

Zur Quantifizierung der Zusammenhangsstärke lassen sich nun die Zusammenhangsmaße (10.1), (10.3) und (10.5) heranziehen. Die Berechnung des χ^2 -Koeffizienten (10.1) besteht hier aus der Bestimmung von $7 \cdot 2 = 14$ Termen $\frac{(h_{ij} - \tilde{h}_{ij})^2}{h_{ij}}$. Für den ersten Term errechnet man z. B.

$$\frac{(h_{11} - \tilde{h}_{11})^2}{\tilde{h}_{11}} = \frac{(235 - 261,8)^2}{261,8} \approx 2,74.$$

Analog ermittelt man unter Rückgriff auf die in Beispiel 9.3 bestimmten Werte \tilde{h}_{ij} die übrigen 13 Terme. Man erhält bei Rundung auf 2 Dezimalstellen

$$\begin{aligned}\chi^2 &\approx 2,74 + 1,71 + 4,65 + 0,32 + 2,06 + 7,67 + 0,02 \\ &\quad + 3,60 + 2,25 + 6,12 + 0,42 + 2,71 + 10,08 + 0,02 = 44,39.\end{aligned}$$

Da hier $n = 1\,451$ sowie $k = 7$, $m = 2$ und damit $M = \min(7; 2) = 2$ ist, folgt für die kleinste obere Schranke χ_{\max}^2 des χ^2 -Koeffizienten nach (10.2)

$$\chi_{\max}^2 = 1451 \cdot 1 = 1451.$$

Der Wert $\chi^2 \approx 44,39$ liegt deutlich näher an der unteren Schranke 0, was für einen nur schwach ausgeprägten Merkmalszusammenhang spricht. Für den Φ -Koeffizienten (10.3) gilt

$$\Phi = \sqrt{\frac{44,39}{1451}} \approx 0,175.$$

Dieser Wert und der für das Cramérsche Zusammenhangsmaß V aus (10.5) stimmen hier überein.

Man kann die vorstehenden Berechnungen natürlich auch unter Heranziehung geeigneter Statistiksoftware durchführen, wie der folgende R-Code illustriert.

```
> library(descTools) ## Paket für Phi-Koeffizient und für Cramér's V
>
> ##### Datenaufbereitung #####
> parteien <- c("Union", "SPD", "Grüne", "LINKE", "FDP", "AfD", "Sonstige")
> geschlechter <- c("männlich", "weiblich")
>
> wahlen <-
+   matrix(c(235, 219, 86, 86, 88, 93, 17,
+           226, 134, 105, 75, 45, 30, 12),
+           nrow = length(parteien),
+           ncol = length(geschlechter),
+           dimnames = list(parteien, geschlechter))
+ )
> wahlen
      männlich weiblich
Union       235     226
SPD        219     134
Grüne      86      105
LINKE      86      75
FDP        88      45
AfD        93      30
Sonstige    17      12
>
> ##### Berechnung #####
> Phi(wahlen)
[1] 0.1749061
> CramerV(wahlen)
[1] 0.1749061
>
```

Tab. 10.1: Berechnung der Zusammenhangsmaße Φ und V mit der freien Software R (Politbarometer-Daten aus Beispiel 10.1)

Im Spezialfall der in Tabelle 9.4 wiedergegebenen **Vierfeldertafel** hat man für den χ^2 -Koeffizienten (10.1) zunächst die Doppelsumme

$$\chi^2 := \sum_{i=1}^2 \sum_{j=1}^2 \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}.$$

Aus dieser Darstellung gewinnt man nach Einsetzen von 9.9 und einigen – hier nicht wiedergegebenen – Umformungen die nachstende Formel, bei der im Nenner das Produkt der Randhäufigkeiten steht:

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_1 \cdot h_2 \cdot h_{1.} \cdot h_{.2}}. \quad (10.7)$$

Bei einer Vierfeldertafel, oder – allgemeiner – im Falle $M = 2$ stimmen der Phi-Koeffizient Φ aus (10.3) und das Kontingenzmaß (10.5) von Cramér stets überein. Es gilt dann offenbar

$$\Phi = V = \frac{|h_{11}h_{22} - h_{12}h_{21}|}{\sqrt{h_1 \cdot h_2 \cdot h_{1.} \cdot h_{.2}}}. \quad (10.8)$$

Die Betragsbildung im Zähler ist notwendig, weil die dort auftretende Differenz negativ sein kann.

Zusammenhangsmessung bei binären Merkmalen



Aufgabe 10.1

Exkurs 10.1: Weitere Zusammenhangsmaße

Es gibt noch weitere Ansätze zur Messung von Zusammenhängen bei nominal-skalierten Merkmalen, die ebenfalls Modifikationen von (10.1) darstellen. Erwähnt sei ein Zusammenhangsmaß von Karl PEARSON, das meist mit K oder mit C abgekürzt wird und sich vom Φ -Koeffizienten dadurch unterscheidet, dass in (10.3) statt n der Term $\chi^2 + n$ erscheint:

$$K := \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Auch dieses Maß ist noch dimensionsabhängig. Analog zu (10.4) gilt

$$0 \leq K \leq K_{\max} = \sqrt{\frac{M-1}{M}}.$$

Mit Division durch die kleinste obere Schranke $K_{\max} := \sqrt{\frac{M-1}{M}}$ erhält man das korrigierte Zusammenhangsmaß

$$K^* = \frac{K}{K_{\max}},$$

das wie Cramér's V nur Werte zwischen 0 und 1 annimmt. Der Ansatz (10.5), der vom χ^2 -Koeffizient in einem Schritt zu einem normierten Zusammenhangsmaß führt, ist allerdings transparenter und weniger umständlich.

10.2 Metrische Merkmale

Bei Merkmalen mit metrischer Skalierung sind, anders als bei nominal-skalierten Merkmalen, die Abstände zwischen den Merkmalsausprägungen interpretierbar (vgl. erneut Tabelle 2.1). Sie können daher bei der Konstruktion von Zusammenhangsmaßen verwendet werden. Ein erstes Maß für den Zusammenhang zwischen zwei metrischen Merkmalen X und Y ist die analog zu (5.6) definierte **Kovarianz**

$$\begin{aligned}s_{xy} &:= \frac{1}{n} \cdot [(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})] \\ &= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),\end{aligned}\quad (10.9)$$

Ein nicht-normiertes Zusammenhangsmaß: die präziser auch **empirische Kovarianz** genannt wird. Wenn man die Kovarianz ohne Rechner bestimmt, kann die nachstehende Zerlegungsformel nützlich sein, bei der \bar{xy} das arithmetische Mittel aus den Produkttermen $x_1 \cdot y_1, \dots, x_n \cdot y_n$ bezeichnet:

$$s_{xy} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \cdot y_i - \bar{x} \cdot \bar{y} = \bar{xy} - \bar{x} \cdot \bar{y}.$$

Diese Formel verifiziert man, ähnlich wie (5.7), wenn man den in (10.9) hinter dem Summenzeichen stehenden Produktterm ausmultipliziert und dann die Summierung gliedweise vornimmt.

Was die Kovarianz inhaltlich bezeichnet, wird verständlich, wenn man die Datenpaare $(x_1, y_1), \dots, (x_n, y_n)$ für X und Y in einem Streudiagramm präsentiert, in das man – parallel zum ersten – noch ein zweites Koordinatensystem einzeichnet, dessen Ursprung im Punkt (\bar{x}, \bar{y}) liegt. Durch das zweite Bezugssystem sind, wie in Abbildung 10.1 dargestellt, vier Quadranten definiert. Jeder Punkt (x_i, y_i) definiert zusammen mit den auf den Achsen des zweiten Koordinatensystems liegenden Punkten (x_i, \bar{y}) und (\bar{x}, y_i) sowie dem neuen Ursprung (\bar{x}, \bar{y}) (lies: *x-quer-y-quer*) ein Rechteck mit Flächeninhalt A_i .

Verwendet man abkürzend für das Produkt der Mittelwertabweichungen $x_i - \bar{x}$ und $y_i - \bar{y}$ die Notation

$$p_i := (x_i - \bar{x})(y_i - \bar{y}) \quad i = 1, \dots, n,$$

so gilt offenbar $A_i = p_i$, wenn der Produktterm p_i positiv ist, und $A_i = -p_i$, wenn p_i negative Werte annimmt. Der erste Fall tritt genau dann ein, wenn die in p_i eingehenden Terme $(x_i - \bar{x})$ und $(y_i - \bar{y})$ entweder beide positiv oder beide negativ sind. Diese Bedingungen sind erfüllt, wenn der Punkt (x_i, y_i) im ersten oder im dritten Quadranten des neuen

Bezugssystems liegt. Der zweite Fall ist genau dann gegeben, wenn einer der beiden genannten Differenzterme positiv und der andere negativ ist. Dies wiederum trifft zu, wenn (x_i, y_i) im zweiten oder vierten Quadranten des zweiten Koordinatensystems liegt. Abbildung 10.1 veranschaulicht

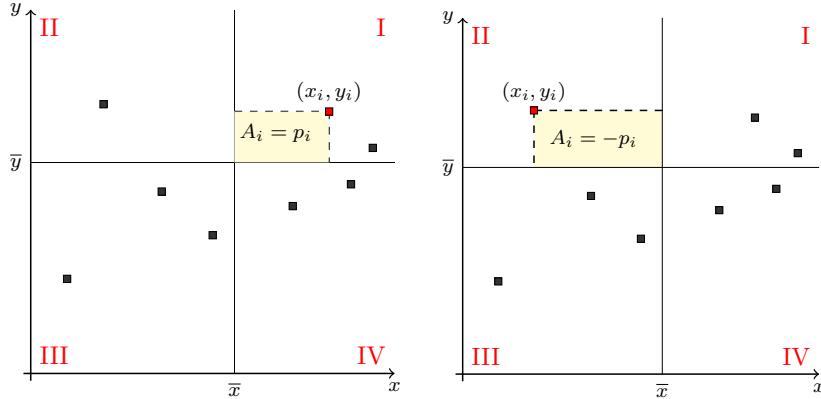


Abb. 10.1: Interpretation der Summanden in der Kovarianzformel

die beiden denkbaren Fälle. Im linken Teil der Abbildung ist ein im ersten Quadranten liegender Punkt (x_i, y_i) eingezeichnet ($p_i > 0, A_i = p_i$), im rechten Abbildungsteil ein Punkt im zweiten Quadranten ($p_i < 0, A_i = -p_i$). Datenpunkte (x_i, y_i) , die im ersten oder dritten Quadranten des mit (\bar{x}, \bar{y}) als Bezugspunkt arbeitenden Koordinatensystems liegen, liefern also einen *positiven*, Punkte im zweiten oder vierten Quadranten hingegen einen *negativen* Beitrag zur Kovarianz. Je mehr Datenpunkte im ersten und dritten Quadranten liegen, desto größer wird die Kovarianz.

Wenn alle Punkte auf einer steigenden Geraden durch (\bar{x}, \bar{y}) liegen, liefert jeder Punkt einen nicht-negativen Beitrag. Entsprechend gilt, dass die Kovarianz um so kleiner wird, je mehr Datenpunkte im zweiten und vierten Quadranten liegen. Wenn alle Punkte auf einer fallenden Geraden durch (\bar{x}, \bar{y}) liegen, liefert kein Punkt einen positiven Beitrag zur Kovarianz. Eine positive Kovarianz bedeutet also, dass die Ausprägungen der Merkmale X und Y eine gleichgerichtete Tendenz haben – kleinere bzw. größere Werte des einen Merkmals gehen tendenziell mit kleineren resp. größeren Werten des anderen Merkmals einher. Umgekehrt gibt es bei negativer Kovarianz eine gegenläufige Tendenz.

Wie der Median, der Mittelwert und die Standardabweichung ist auch die Kovarianz maßstabsabhängig. Sie kann durch Maßstabsänderung beliebig vergrößert oder verkleinert werden. Außerdem ist sie nicht dimensionslos. Ein maßstabsunabhängiges und dimensionsloses Zusammenhangsmaß erhält man, wenn man die empirische Kovarianz s_{xy} zweier metrischer Merkmale X und Y durch das Produkt ihrer Standardabweichungen s_x

Ein normiertes Zusammenhangsmaß

resp. s_y dividiert. Das resultierende Zusammenhangsmaß



Karl PEARSON

wird **Korrelationskoeffizient** genannt. Da der Ansatz (10.10) dem französischen Physiker Auguste BRAVAIS (1811 - 1863) und dem britischen Statistiker Karl PEARSON (1857 - 1936) zugeschrieben wird, spricht man auch vom **Korrelationskoeffizienten nach Bravais-Pearson**. Aus der Darstellung (10.10) ersieht man, dass die Merkmale X und Y symmetrisch eingehen. Eine Vertauschung der Merkmalsbezeichnungen ändert nichts am Wert von r .

Wenn man in (10.10) für den Zähler den Summenterm aus (10.9), im Nenner für die Standardabweichung von X den Wurzelausdruck aus (5.8) – nun mit der präziseren Schreibweise s_x anstelle von s – und für die Standardabweichung s_y ebenfalls den analog nach (5.8) erklärten Wurzelterm einsetzt, erhält man für r die ausführlichere Darstellung

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (10.11)$$

Mit (10.9) und (5.8) gewinnt man aus (10.11) noch als weitere Darstellung

$$r = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\bar{x}^2 - \bar{x}^2} \cdot \sqrt{\bar{y}^2 - \bar{y}^2}}. \quad (10.12)$$

Da die im Nenner von (10.10) auftretenden Standardabweichungen s_x und s_y positiv sind, ist das Vorzeichen von r stets mit dem Vorzeichen der Kovarianz s_{xy} identisch, d. h. der Korrelationskoeffizient r kann sowohl positive als auch negative Werte annehmen. Im ersten Fall spricht man von einer *positiven*, im zweiten Fall von einer *negativen* **Korrelation** zwischen X und Y und im Falle $r = 0$ von **Unkorreliertheit** beider Merkmale. Der Korrelationskoeffizient liegt stets zwischen -1 und $+1$:

$$-1 \leq r \leq 1. \quad (10.13)$$

Die obere Schranke $r = 1$ wird erreicht, wenn alle Datenpunkte auf einer *steigenden*, die untere Schranke $r = -1$ hingegen, wenn sich alle Datenpunkte auf einer *fallenden* Geraden liegen. In beiden Fällen, also für $|r| = 1$ (lies: r -Betrag = 1), besteht lineare Abhängigkeit zwischen den Merkmalen und die Gerade verläuft durch den Punkt (\bar{x}, \bar{y}) .

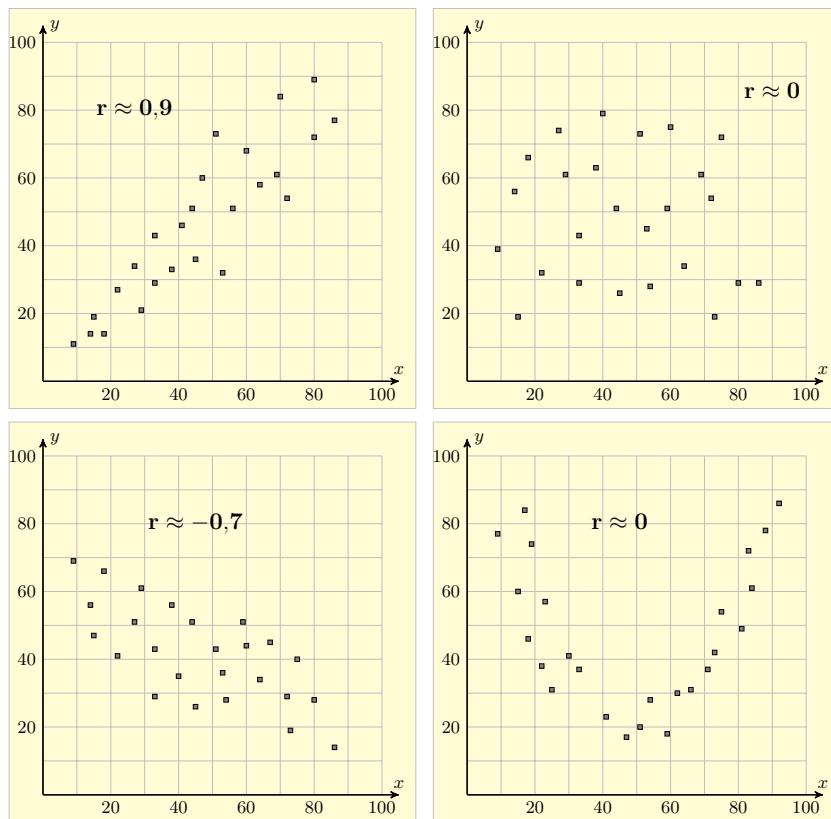


Abb. 10.2: Korrelationskoeffizienten für verschiedene Streudiagramme



Interaktives Objekt
„Korrelation“

Abbildung 10.2 veranschaulicht vier Streudiagramme, die unterschiedliche Situationen für den Zusammenhang zwischen zwei Merkmalen widerspiegeln. Die Grafiken in der oberen Hälfte zeigen erneut die Streudiagramme der Abbildung 9.5, nun aber mit Ausweis des Zusammenhangsmaßes r . Das erste Streudiagramm repräsentiert einen Fall starker positiver Korrelation ($r = 0,9$), während der Wert $r = 0$ im zweiten Fall Unkorreliertheit beinhaltet. Die Datenpaare $(x_1, y_1), \dots, (x_n, y_n)$ sind im letztgenannten Fall so auf die vier Quadranten des in Abbildung 10.1 wiedergegebenen Koordinatensystems mit Bezugspunkt (\bar{x}, \bar{y}) verteilt, dass sich die Beiträge p_i beim Aufsummieren gerade aufheben.

Das dritte Streudiagramm der Abbildung 10.2 zeigt mäßig ausgeprägte negative Korrelation ($r = -0,7$). Obwohl das vierte Diagramm Unkorreliertheit ausweist ($r = 0$), lässt es einen nicht-linearen Merkmalszusammenhang vermuten. Auch hier sind die Datenpaare so auf die vier Quadranten verteilt, dass sich die Kovarianzbeiträge p_i kompensieren. Der letzte Fall macht deutlich, dass der Korrelationskoeffizient r ein Maß für *linearen* Zusammenhang darstellt. Korrelation bedeutet, dass ein *linearer* Merkmalszusammenhang gegeben ist. Wenn $r = 0$ ist, kann durchaus ein nicht-linearer Zusammenhang vorliegen. Ein Wert $r \neq 0$ lässt nur auf

Korrelationskoeffizient r :
Maß für linearen Zusammenhang

das Vorliegen eines linearen Merkmalszusammenhangs schließen. Im Falle $|r| = 1$ spricht man *vollständiger* Korrelation (lineare Abhängigkeit), im Falle $0 < |r| < 0,5$ häufig von *schwacher*, für $0,5 \leq |r| < 0,8$ von *mäßiger* und bei Werten $0,8 \leq |r| < 1$ von *starker* Korrelation.

Beispiel 10.2: Wie gut waren die Prognosen der Sachverständigen?

Der Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung („Die 5 Weisen“) legt alljährlich eine Prognose zur wirtschaftlichen Entwicklung in Deutschland für das nächste Jahr vor. Prognostiziert wird insbesondere die Wachstumsrate für das Bruttoinlandsprodukt. Interessant ist es, für eine zurückliegende Periode zu vergleichen, wie weit sich die prognostizierten Werte von den hinterher tatsächlich beobachteten Werten unterschieden haben. Als Gütemaß kann der Korrelationskoeffizient r herangezogen werden. Bei perfekter Vorhersage würden die Ausprägungen der Merkmale „Prognose X“ und „realer Wert Y“ übereinstimmen. Die Datenpaare $(x_1, y_1), \dots, (x_n, y_n)$ lägen dann auf einer steigenden Geraden ($r = 1$). Tabelle 9.2 weist für 15 Perioden i (Jahre 1983 = 1, ..., 1997 = 15) die jeweils im Herbst des Vorjahres abgegebenen Prognosen x_i des Sachverständigenrats für die Periode i und die hinterher realisierten Werte y_i aus.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	1,0	2,5	3,0	3,0	2,0	1,5	2,5	3,0	3,5	2,5	0,0	0,0	3,0	2,0	2,5
y_i	1,2	2,6	2,5	2,5	1,7	3,4	4,0	4,6	3,4	1,5	-1,9	2,3	1,9	1,4	2,2

Tab. 10.2: Wachstumsprognosen der „Fünf Weisen“ und wahre Werte

Zur Bestimmung des Bravais-Pearsonschen Korrelationskoeffizienten r wird man eine Statistiksoftware heranziehen, etwa SPSS oder die freie Statistiksoftware R. Der nachstehende R-Output umfasst – anders als der ebenfalls wiedergegebene SPSS-Screenshot – auch die Dateneingabe.

	x	y
x	Korrelation nach Pearson Signifikanz (2-seitig) N	1 .018 15
y	Korrelation nach Pearson Signifikanz (2-seitig) N	.602* .018 15
		> x <- c(1, 2.5, 3, 3, 2, 1.5, 2.5, 3, 3.5, 2.5, 0, 0, 3, 2, 2.5) > y <- c(1.2, 2.6, 2.5, 2.5, 1.7, 3.4, 4, 4.6, 3.4, 1.5, -1.9, 2.3, 1.9, 1.4, 2.2) > cor(x, y) [1] 0.6018267

Tab. 10.3: Computerausdruck zur Berechnung des Korrelationskoeffizienten r mit SPSS (links) und mit R (rechter Teil)

Das Ergebnis $r \approx 0,60$ ist nicht überraschend – die prognostizierten und die beobachteten realen Wachstumsraten sind positiv korreliert. Der Idealwert $r = 1$ ist natürlich in der Realität kaum erreichbar, weil stets nach Abgabe einer Vorhersage noch unvorhersehbare Einflüsse und Turbulenzen auftreten können, die die tatsächlichen wirtschaftlichen Entwicklungen verändern – man denke

z. B. an die Finanzkrise von 2008 oder an die Corona-Krise 2020. Prognosen werden daher während des Prognosezeitraums bei Bedarf noch korrigiert.

Natürlich kann man r aus den Daten der Tabelle 10.2 auch anhand von (10.11) oder (10.12) manuell unter Verwendung einer Arbeitstabelle errechnen (vgl. hierzu die Lösung zur nebenstehenden Aufgabe 10.2). Man erhielte so mit (10.12) etwas mühsam

$$r = \frac{5,643 - 2,133 \cdot 2,22}{\sqrt{5,633 - 2,133^2} \cdot \sqrt{7,029 - 2,22^2}} \approx \frac{0,90774}{1,50851} \approx 0,602.$$



Aufgabe 10.2

Der Korrelationskoeffizient r kann Aufschluss darüber geben, ob es einen mehr oder weniger ausgeprägten *linearen* empirischer Zusammenhang zwischen zwei metrischen Merkmalen gibt. Ein hoher Absolutbetrag $|r|$ besagt lediglich, dass die Daten für die in Rede stehenden Merkmale entweder eine gleichgerichtete Tendenz aufweisen (im Falle $r > 0$) oder eine gegenläufige Tendenz (im Falle $r < 0$). Ein anhand eines großen Werts $|r|$ festgestellter empirischer Zusammenhang muss nicht zwingend bedeuten, dass zwischen den Merkmalen ein Kausalzusammenhang besteht.

Korrelation impliziert nicht zwingend einen sachlogischen Zusammenhang

Wenn zwei Variablen X und Y korreliert sind, ohne in einem direkten inhaltlichen Zusammenhang zu stehen, spricht man von **Scheinkorrelation**. Weniger missverständlich wäre der Begriff „Scheinkausalität“. Auf der Website [Spurious Correlations](#) des US-Amerikaners Tyler Vigen findet man eine Sammlung kurioser Beispiele mit Zeitreihendaten für jeweils zwei Merkmale X und Y , bei denen eine Korrelation gemessen wird, ohne dass eine sachlogische Verbindung vorhanden ist.

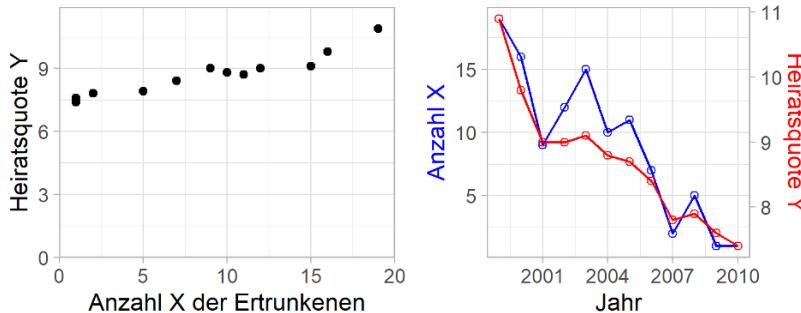


Abb. 10.3: Beispiel einer Scheinkorrelation. Oben: Streudiagramm. Unten: Zeitreihengraphen

Ein in Abbildung 10.3 veranschauliches Beispiel bezieht sich auf Daten aus der Periode 1999 – 2010, wobei X die Anzahl der US-Bürger bezeichnet, die im genannten Zeitraum aus einem Fischerboot gefallen und ertrunken sind, und Y die Anzahl der in dieser Periode registrierten Heiratsfälle in Kentucky pro 1 000 Einwohner. Der linke Teil der Abbildung

gibt die Daten $(x; y)$ als Streudiagramm wieder, d. h. unter Ausblendung der zeitlichen Abfolge. Der rechte Teil zeigt beide Zeitreihengraphen in einer einzigen Grafik mit unterschiedlichen Achsen. Der Wertebereich der roten Achse, die sich auf den roten Polygonzug bezieht, ist hier so gewählt, dass die beiden Graphen den gleichen Anfangs- und Endpunkt aufweisen. Die beiden Merkmale X und Y weisen eine gleichgerichtete Tendenz auf (starke positive Korrelation), was sich im Wert $r \approx 0,95$ für den Korrelationskoeffizienten widerspiegelt. Aus dem empirischen Befund $r \approx 0,95$ könnte man fälschlich schließen, dass zwischen der Häufigkeit der tragischen Fischerbootunfälle in den USA und der Heiratsquote in Kentucky eine inhaltliche Verbindung besteht. Es ist offensichtlich, dass diese nicht vorliegt.

Man spricht auch von Scheinkorrelation, wenn zwischen zwei Merkmalen X und Y nur ein indirekter Zusammenhang besteht – in dem Sinne, dass ein drittes Merkmal Z im Spiel ist, das mit den beiden anderen Merkmalen korreliert ist. Abbildung 10.4 veranschaulicht diesen Fall.

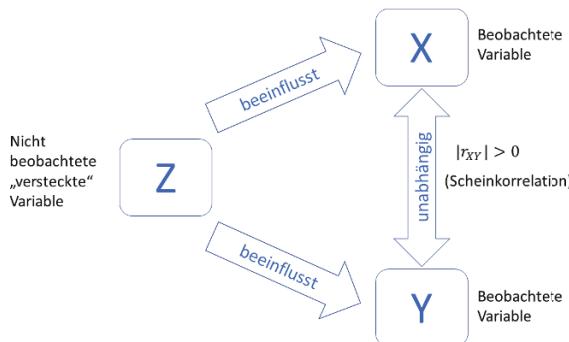


Abb. 10.4: Schematische Darstellung eines indirekten Zusammenhangs zwischen zwei Merkmalen

Ein bei HANDEL / KUHLENKASPER (2018, Abschnitt 4.3.4) angeführtes Beispiel bezieht sich auf einen bei einer Gruppe von Studenten anhand des Wertes $r \approx 0,71$ festgestellten empirischen Zusammenhang zwischen Körpergewicht X und Schuhgröße Y . Als vermittelnde Drittvariable kommt hier die Körpergröße Z in Betracht.

Exkurs 10.2: Bereinigung von Drittvariableneinflüssen

Der Korrelationskoeffizient nach Bravais-Pearson quantifiziert die Stärke eines linearen Zusammenhangs zwischen zwei Merkmalen X und Y . Bei einem vermuteten Einfluss einer Drittvariablen Z ist man daran interessiert, den Einfluss von Z „herauszurechnen“. Hierfür wird der sog. **partielle Korrelationskoeffizient** verwendet, der mit $r_{xy.z}$ abgekürzt sei. Bezeichnet r_{xy} den

Korrelationskoeffizienten für die Merkmale X und Y und r_{xz} bzw. r_{yz} den für X und Z resp. Y und Z , so ist $r_{xy.z}$ gegeben durch

$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2} \cdot \sqrt{1 - r_{yz}^2}}.$$

Das Zusammenhangsmaß $r_{xy.z}$ gibt also an, wie stark die Korrelation zwischen X und Y ausgeprägt wäre, wenn der Einfluss von Z ausgebendet würde. Die von Drittvariableneinflüssen bereinigte Korrelation heißt auch **partielle Korrelation** oder **Partialkorrelation**. Durch eine solche Bereinigung lassen sich auch Scheinkorrelationen aufdecken. Bei dem zur Illustration von Abbildung 10.4 erwähnten Beispiel eines empirisch ermittelten Zusammenhangs zwischen Körpergewicht und Schuhgröße ergab sich nach Bereinigung um den Effekt der Körpergröße der Wert $r_{xy.z} \approx 0,04$.

In der *Psychologie* nennt man eine Drittvariable, die den Zusammenhang zwischen einer unabhängigen Variablen X und einer abhängigen Variablen Y beeinflusst, **Moderatorvariable**.

Das Phänomen „Scheinkorrelation“ stellt keinesfalls das Konzept von empirischer Korrelation und statistischer Inferenz infrage. Man darf nur aus empirischen Zusammenhangsmaßen keine voreiligen Schlussfolgerungen ziehen. Die korrekte Interpretation einer empirisch festgestellten Korrelation erfordert zumeist erheblichen Aufwand (Recherche, Fachwissen, etc.). Hierin liegt ein entscheidender Unterschied zwischen Mensch und Maschine: Mit leistungsfähiger Software lassen sich zwar Formeln und Modelle extrem schnell durchrechnen, doch die Interpretation und die Beurteilung der Frage, ob die Ergebnisse zur Generierung von Wissen und zur Entscheidungsfindung beitragen, bleiben dem Menschen überlassen.

In der statistischen Praxis interessiert man sich manchmal für Merkmale, die entweder grundsätzlich nicht direkt beobachtbar sind oder für die Daten aus technischen oder rechtlichen Gründen nicht erhoben werden können. Der erste Fall liegt vor, wenn die in Rede stehenden Merkmale hypothetische Konstrukte sind (sog. latente Variablen wie z. B. Intelligenz, Leistungsmotivation, Erfolg, Lebensqualität). Der zweite Fall ist etwa gegeben, wenn die Datenerfassung für Merkmale zu teuer oder zu zeitaufwändig ist oder Datenschutzbestimmungen eine Datenerhebung nicht zulassen. In solchen Situationen bietet es sich an, anstelle eines nicht beobachtbaren Merkmals hilfsweise ein anderes Merkmal zu erfassen, das mit dem ersten Merkmal stark korreliert ist und mit diesem in einem engen Kausalzusammenhang steht. Das hilfsweise verwendete Merkmal nennt man **Proxyvariable**. So kann das Merkmal „beruflicher Erfolg“ anhand der Proxyvariablen „Bruttoeinkommen“ gemessen werden und der Lebensstandard in einem Land anhand des jeweiligen Bruttoinlandsprodukts.

Beispiel 10.3: Umsatzschätzung mit Proxyvariablen

Flughäfen erwirtschaften heute einen großen Teil ihres Umsatzes durch die Vermietung von Ladengeschäften. Die Mieten sind häufig an den Umsatz gekoppelt, so dass es im unmittelbaren Interesse eines Flughafenbetreibers liegt, wenn Passagiere viel einkaufen.

An einem großen Flughafen wie demjenigen in Frankfurt variieren die Einkaufsmöglichkeiten an den verschiedenen Terminals beträchtlich. Neben Speisen und Getränken sind Modeartikel, Kosmetik, Elektronikartikeln, Uhren und Spirituosen im Angebot. Gleichermassen variieren die Präferenzen und Bedürfnisse der Kunden. Ihre Kaufentscheidungen werden von Persönlichkeitsmerkmalen, dem Wegverlauf zum Gate und den dabei passierten Geschäften, Warte- und Gehzeiten oder dem Reiseanlass beeinflusst.

Will man ein Prognosemodell erstellen, das für den Flughafen wichtige Faktoren und die Stärke ihres Einflusses auf die Ladenumsätze identifiziert, ist es wichtig zu verstehen, welche Faktoren von Flughafenbetreibern kontrollierbar sind und wie diese Faktoren ohne Störung des Betriebsflusses mit dem Ziel einer Umsatzmaximierung adjustiert werden können. Einige umsatzbeeinflussende Faktoren sind Merkmale der Passagiere. Letztere können in Deutschland aus Datenschutzgründen nicht oder nur eingeschränkt erhoben werden (z. B. Geschlecht oder Nationalität), Daten zu anderen Merkmalen (etwa Kaufkraft oder Reiseanlass) sind schlicht nicht verfügbar.

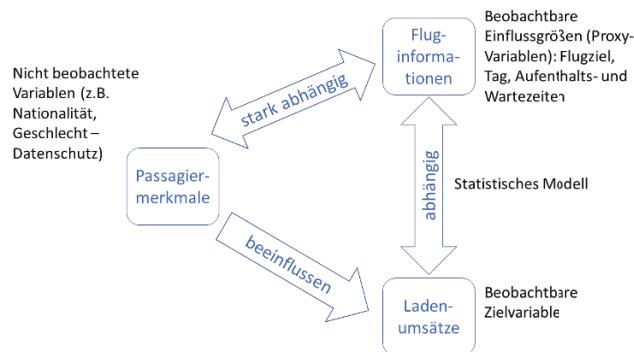


Abb. 10.5: Einflussgrößen für den Umsatz in Flughafenläden

Ein Umsatzprognosemodell kann nicht direkt beobachtbare, aber für die Fragestellung relevante Merkmale durch Proxyvariablen substituieren, und z. B. anstelle des nicht erfassten Passagiermerkmals „Nationalität“ hilfreiche die Fluginformation „Flugziel“ verwenden.

10.3 Ordinalskalierte Merkmale

Für ordinalskalierte Merkmale ist der Korrelationskoeffizient r nach Bravais-Pearson nicht anwendbar, weil in dessen Berechnung Differenzen eingehen, die bei ordinaler Skalierung nicht erklärt sind (vgl. Tabelle 2.1). Ein auf Charles SPEARMAN (1863 - 1945) zurückgehender Ansatz sieht vor, bei ordinalskalierten Merkmalen X und Y zunächst für jeden Wert x_i und unabhängig davon auch für jeden Wert y_i die Rangposition $rg(x_i)$ bzw. $rg(y_i)$ zu bestimmen und dann die Formel (10.11) für r so zu modifizieren, dass sie sich nicht mehr auf die originären Datenpaare (x_i, y_i) , sondern auf $(rg(x_i), rg(y_i))$ bezieht. Dazu werden in (10.11) x_i und y_i durch $rg(x_i)$ bzw. $rg(y_i)$ sowie \bar{x} und \bar{y} durch die Mittelwerte \bar{rg}_x resp. \bar{rg}_y der Rangplätze ersetzt. Man erhält so den mit r_{SP} (lies: r - s - p) abgekürzten **Rangkorrelationskoeffizienten nach Spearman**:

$$r_{SP} = \frac{\sum_{i=1}^n (rg(x_i) - \bar{rg}_x)(rg(y_i) - \bar{rg}_y)}{\sqrt{\sum_{i=1}^n (rg(x_i) - \bar{rg}_x)^2} \cdot \sqrt{\sum_{i=1}^n (rg(y_i) - \bar{rg}_y)^2}}. \quad (10.14)$$

Da r_{SP} sich als Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf Paare $(rg(x_i), rg(y_i))$ von Rangpositionen interpretieren lässt, gelten die Schranken aus (10.13) auch für den Rangkorrelationskoeffizienten, d. h. es gilt

$$-1 \leq r_{SP} \leq 1. \quad (10.15)$$

Während r ein Maß für einen linearen Zusammenhang zwischen den Beobachtungswerten für zwei Merkmale darstellt, misst r_{SP} nur einen linearen Zusammenhang zwischen den Rangplätzen der Merkmalswerte. Bezogen auf die originären Merkmalswerte selbst misst r_{SP} lediglich, ob ein gleich- oder gegensinniger *monotoner* Zusammenhang vorliegt. Bei gleichsinnigem Zusammenhang ist $r_{SP} > 0$, bei gegensinnigem Zusammenhang gilt $r_{SP} < 0$ und bei fehlendem Zusammenhang $r_{SP} = 0$. Das Zusammenhangsmaß r_{SP} ist grundsätzlich auch für metrische Merkmale anwendbar und hat hier den Vorteil einer geringeren Empfindlichkeit gegenüber extremen Merkmalswerten (höhere Robustheit gegenüber Ausreißern). Der Vorteil wird aber mehr als aufgehoben durch den Nachteil, dass r_{SP} nur die Rangpositionen der einzelnen Merkmalswerte verarbeitet und damit die in metrisch skalierten Daten enthaltene Information nur sehr eingeschränkt ausschöpft.

Wenn man voraussetzt, dass kein Rangplatz mehrfach besetzt ist, vereinfacht sich die Darstellung (10.14). Die Mittelwerte \bar{rg}_x resp. \bar{rg}_y der Rangplätze sind dann jeweils identisch mit dem Mittelwert aus den ersten n natürlichen Zahlen, also der Zahlen $1, 2, \dots, n$. Man kann zeigen, dass die Summe der Zahlen $1, 2, \dots, n$ durch $\frac{n(n+1)}{2}$ gegeben ist, ihr Mittel-



wert also durch $\frac{n+1}{2}$. Einsetzen in (10.14) liefert bei Verwendung der Abkürzung d_i für die Differenz der Rangpositionen $rg(x_i)$ und $rg(y_i)$ nach elementaren Umformungen

$$r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)} \quad d_i := rg(x_i) - rg(y_i). \quad (10.16)$$

Beispiel 10.4: Berechnung von r_{SP}

Die Berechnung des Spearmanschen Rangkorrelationskoeffizienten sei anhand eines fiktiven Datensatzes für zwei ordinalskalierte Merkmale illustriert. Es sei angenommen, dass zwei unabhängige Kreditsachbearbeiter die Kreditwürdigkeit von fünf Sparkassenkunden anhand einer 10-stufigen Ratingskala bewerten, bei der die Punktzahl 1 sehr schlechte und die Punktzahl 10 sehr gute Bonität bezeichne. Die Ergebnisse der Bewertungen sind in der zweiten und vierten Spalte von 10.4 ausgewiesen.

Kunden-Nr. i	Sachbearbeiter A		Sachbearbeiter B		
	Bewertung x_i	$rg(x_i)$	Bewertung y_i	$rg(y_i)$	d_i
1	5	4	6	3	1
2	8	2	9	1	1
3	9	1	7	2	-1
4	2	5	4	5	0
5	6	3	5	4	-1

Tab. 10.4: Bonitätsbewertung von Sparkassenkunden

Ausgangspunkt für die Berechnung von r_{SP} ist die Rechenformel (10.16). Um die Formel anwenden zu können, müssen die Rangplätze der Ausgangsdaten x_1, x_2, \dots, x_5 und y_1, y_2, \dots, y_5 bestimmt werden. Der erste Sachbearbeiter hat den dritten Kunden am besten beurteilt; der Punktzahl $x_3 = 9$ wird daher der Rangplatz 1 zugewiesen. Beim zweiten Sachbearbeiter hat der zweite Kunde die beste Bewertung und infolgedessen erhält hier $y_2 = 9$ den Rangplatz 1. Entsprechend ermittelt man die übrigen acht Rangplätze, die in der dritten und letzten Spalte der Tabelle wiedergegeben sind.

Zur Berechnung des Spearmanschen Korrelationskoeffizienten r_{SP} benötigt man neben der Länge n des bivariaten Datensatzes nur die Rangdifferenzen. Setzt man die Differenzen d_i und $n = 5$ in (10.16) ein, resultiert der Wert

$$r_{SP} = 1 - \frac{6 \cdot [1^2 + 1^2 + (-1)^2 + 0^2 + (-1)^2]}{5 \cdot (25 - 1)} = 0,8.$$

Zwischen den Beurteilungen der beiden Sachbearbeiter gibt es also einen ausgeprägten gleichsinnig monotonen Zusammenhang.

Teil II

Wahrscheinlichkeits- rechnung und schließende Statistik



Lernziele zu Teil II

Nach Bearbeitung des zweiten Teils dieses Manuskripts sollten Sie

- mit Grundbegriffen der Wahrscheinlichkeitsrechnung und der Kombinatorik vertraut sein;
- wissen, dass es diskrete und stetige Zufallsvariablen gibt, deren Verhalten anhand von Verteilungsmodellen charakterisiert wird;
- die Binomialverteilung einschließlich des Spezialfalls der Bernoulli-Verteilung sowie die hypergeometrische Verteilung als Vertreter diskreter Verteilungen kennen;
- die genannten diskreten Verteilungen anhand ihrer Wahrscheinlichkeits- und Verteilungsfunktion und anhand von Lage- und Streuungsparametern charakterisieren können;
- die Normalverteilung als wichtigste stetige Verteilung einschließlich des Spezialfalls der Standardnormalverteilung kennen und anhand ihrer Dichte- und Verteilungsfunktion sowie anhand von Lage- und Streuungsparametern charakterisieren können;
- wissen, dass die Chi-Quadrat-, die t- und die F-Verteilung weitere stetige Verteilungen sind, die sich aus der Normalverteilung ableiten;
- Maße zur Beschreibung des Zusammenhangs zwischen zwei Zufallsvariablen kennen;
- in der Lage sein, einige Stichprobenfunktionen zu benennen und zur Schätzung von Kenngrößen für Verteilungsmodelle (z. B. Erwartungswert) heranzuziehen;
- neben der Punktschätzung von Modellparametern auch das Konzept der Intervallschätzung verstanden haben;
- mit Grundbegriffen des Testens von Hypothesen vertraut sein und verschiedene Arten von Tests benennen können;
- mit den beim Testen möglichen Fehlern vertraut sein und wissen, dass sich die Leistungsfähigkeit von Tests anhand der Gütfunktion bewerten lässt;
- zu einer Punktwolke anhand der Kleinst-Quadrat-Methode eine Regressionsgerade bestimmen und deren Anpassungsgüte quantifizieren können;
- die Grundidee und Zielsetzung der Varianzanalyse sowie den Zusammenhang zwischen Regressions- und Varianzanalyse erläutern können.



11 Zufall und Wahrscheinlichkeit



Vorschau auf
das Kapitel

In diesem Kapitel werden u. a. die Begriffe „Zufallsprozess“, „Ereignis“ und „Ergebnismenge“ eingeführt. Dabei werden Venn-Diagramme zur Veranschaulichung herangezogen. Geklärt wird auch der Wahrscheinlichkeitsbegriff, insbesondere der an bestimmte Voraussetzungen gebundene Ansatz zur Berechnung von Wahrscheinlichkeiten nach Laplace. Anschließend erfolgt eine Vorstellung des Urnenmodells. Es werden vier Fälle unterschieden, die beim Ziehen von n Elementen aus einer Urne mit N Elementen auftreten können (Ziehen mit und ohne Zurücklegen, Ziehen mit und ohne Berücksichtigung der Reihenfolge).

In Analogie zu den bedingten relativen Häufigkeiten der beschreibenden Statistik wird noch definiert, was unter den Begriffen „bedingte Wahrscheinlichkeit“ und „Unabhängigkeit von Ereignissen“ zu verstehen ist. Näher eingegangen wird auf bedingte Wahrscheinlichkeiten, die bei medizinischen Test- und Diagnoseverfahren eine wichtige Rolle spielen.

11.1 Grundbegriffe der Wahrscheinlichkeitsrechnung

Aus dem Alltagsleben ist jedem von uns bekannt, dass es Vorgänge gibt, deren Ergebnis vom Zufall abhängt. Man denkt vielleicht zunächst an Glücksspiele (Roulette, Würfelspiele, Ziehung der Lottozahlen), an die Entwicklung von Börsenkursen oder an Wahlergebnisse, die z. B. vom Wetter am Wahltag beeinflusst werden können. Versicherungen sind an der Abschätzung von Schadensverläufen oder der Lebenserwartung von Neugeborenen interessiert, Politikverantwortliche wollen demografische Entwicklungen prognostizieren können und Unternehmen benötigen statistische Informationen zur Quantifizierung von Marktrisiken. Die Wahrscheinlichkeitsrechnung stellt Modelle bereit, die es erlauben, den Verlauf zufallsabhängiger Prozesse abzuschätzen und von Stichproben auf Grundgesamtheiten zu schließen. Die bisher thematisierte beschreibende Statistik charakterisiert gegebene Datensätze ohne einen Rückschluss auf Eigenschaften umfassenderer Grundgesamtheiten zu vermitteln.

Zufallsvorgänge im
Alltagsleben

Ein **Zufallsvorgang** ist ein Prozess, der zu einem von mehreren, sich gegenseitig ausschließenden Ergebnissen ω (lies: *Klein-Omega*) führt. Welches Ergebnis eintritt, ist vorab nicht bekannt. Die möglichen Ergebnisse ω heißen **Elementarereignisse** und werden in einer mit Ω (lies: *Groß-Omega*) bezeichneten Menge

$$\Omega = \{\omega : \omega \text{ ist Elementarereignis}\} \quad (11.1)$$

Darstellung der Ergebnisse von Zufallsvorgängen durch Mengen

zusammengefasst. Die Menge Ω heißt **Ergebnismenge**. Sie kann endlich oder auch unendlich viele Elemente enthalten. Eine Teilmenge A von Ω heißt **Ereignis**. Elementarereignisse sind somit Ereignisse, die nicht weiter zerlegbar sind, also einelementige Teilmengen von Ω darstellen.

Ist A eine Teilmenge von Ω , abgekürzt $A \subset \Omega$ (lies: A ist *Teilmenge* von Ω) und ω das Ergebnis des Zufallsprozesses, so sagt man, dass das Ereignis A eingetreten ist, wenn ω ein Element von A ist, kurz, wenn $\omega \in A$ gilt (lies: ω ist *Element* von A). Das mit \bar{A} (lies: *Komplementärmenge* zu A) bezeichnete **Komplementärereignis** zu A ist das Ereignis, das genau dann eintritt, wenn A nicht eintritt. Die Menge \bar{A} umfasst alle Elementarereignisse, die zu Ω , nicht aber zu A gehören. Man schreibt hierfür auch $\bar{A} = \Omega \setminus A$ (lies: \bar{A} ist *Differenzmenge* von Ω und A). Da auf jeden Fall eines der Elemente der Menge Ω als Ergebnis des Zufallsvorgangs realisiert wird, ist durch Ω ein **sicheres Ereignis** definiert. Das Komplementärereignis $\bar{\Omega}$ zum sicheren Ereignis Ω ist das **unmögliche Ereignis**, das durch die leere Menge \emptyset dargestellt wird

Aus Ereignissen, also Teilmengen einer Ergebnismenge Ω , lassen sich durch logische Verknüpfung der sie repräsentierenden Mengen neue Ereignisse bilden. So ist durch die **Schnittmenge** $A \cap B$ der Ereignisse A und B ein Ereignis definiert, das genau dann eintritt, wenn sowohl A als auch B eintritt. Zwei Ereignisse A und B , deren Schnittmenge die leere Menge \emptyset ist, schließen sich aus. Man spricht auch von **disjunkten Ereignissen**. Die Vereinigungsmenge $A \cup B$ beschreibt ein Ereignis, das dann realisiert wird, wenn mindestens eines der beiden Ereignisse A oder B eintritt. Zur Veranschaulichung solcher zusammengesetzter Ereignisse werden häufig sog. **Venn-Diagramme** verwendet. Diese bestehen aus einem Rechteck, in dem die Ausgangsereignisse (Mengen A, B, \dots) als Kreise oder Ellipsen dargestellt sind. Das Rechteck repräsentiert die Ergebnismenge Ω , von dem die eingezeichneten Mengen Teilmengen sind.

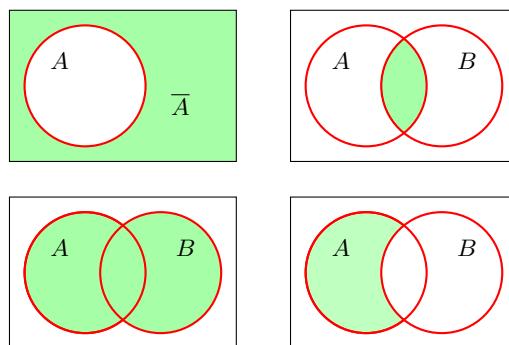


Abb. 11.1: Venn-Diagramme für \bar{A} , $A \cap B$, $A \cup B$ und $A \setminus B$

Abbildung 11.1 zeigt vier Venn-Diagramme. Die oberen beiden Teilgrafiken veranschaulichen das Komplementärereignis $\bar{A} = \Omega \setminus A$ und die Schnittmenge $A \cap B$. Die unteren zwei Teilgrafiken zeigen die Vereinigungs menge $A \cup B$ resp. die Differenzmenge $A \setminus B = A \cap \bar{B}$. Die dargestellten Ereignisse bzw. Mengen sind innerhalb des Ω symbolisierenden Rechtecks jeweils durch dunklere Färbung ausgewiesen.



Aufgabe 11.1

Beispiel 11.1: Ergebnismenge (Münzwurf und Würfeln)

Beim *einfachen Münzwurf* besteht die Ergebnismenge Ω aus nur zwei Elementen, nämlich den beiden möglichen Ausgängen {Zahl, Kopf}. Da immer entweder „Zahl“ oder „Kopf“ auftritt, ist $\Omega = \{\text{Zahl, Kopf}\}$ ein sicheres Ereignis. Das Ereignis, dass beim Münzwurf weder „Zahl“ noch „Kopf“ erscheint, ist ein unmögliches Ereignis.

Beim *zweifachen Münzwurf* ist die Ergebnismenge Ω durch die vier Paare

$$\Omega = \{(Z, Z), (Z, K), (K, Z), (K, K)\}$$



gegeben, wenn man die Abkürzungen „Z“ (Zahl) und „K“ (Kopf) verwendet.

Beim *Würfeln mit einem Würfel* ist die Ergebnismenge Ω durch die Menge $\{1, 2, 3, 4, 5, 6\}$ der ersten sechs natürlichen Zahlen gegeben. Die möglichen Augenzahlen sind hier die Elementarereignisse. Ein Beispiel für ein aus mehreren Elementarereignissen zusammengesetztes Ereignis A ist beim Würfeln mit einem Würfel das Ereignis $A = \{1, 3, 5\}$ (Augenzahl ist ungerade). Das Komplementärereignis $\bar{A} = \Omega \setminus A$ zu A ist hier $\bar{A} = \{2, 4, 6\}$ (Augenzahl ist gerade). Definiert man noch das Ereignis $B = \{5, 6\}$ (Augenzahl ist größer als 4), so gilt für die Schnittmenge der beiden Ereignisse A und B

$$A \cap B = \{5\} = \{\text{Augenzahl ist ungerade und größer als 4}\}.$$

Aufgabe 11.2

Beim *Würfeln mit zwei Würfeln* umfasst Ω schon 36 Elementarereignisse, nämlich die zu Paaren

$$\Omega = \{(1; 1), (1; 2), \dots, (1; 6), (2; 1), (2; 2), \dots, (2; 6), \dots, (6; 1), (6; 2), \dots, (6; 6)\}$$

zusammengefassten Augenzahlen des ersten und des zweiten Würfels. Durch

$$A = \{(1; 1), (1; 2), (2; 1)\} = \{\text{Augensumme beider Würfel beträgt höchstens 3}\}$$

ist hier ein aus mehreren Elementarereignissen zusammengesetztes Ereignis definiert. Das Komplementärereignis $\bar{A} = \{\text{Augensumme ist größer als 3}\}$ umfasst dann die 33 Paare der Menge Ω , die nicht zu A gehören.

Die obigen Beispiele bezogen sich auf Zufallsvorgänge mit nur *endlicher* Anzahl von Elementarereignissen. Würde jeder einmal pro Woche Lotto spielen bis das Traumergebnis „Sechs Richtige und Zusatzzahl“ erreicht wird, so könnte die Anzahl der erforderlichen Spiele von

Zufallsvorgänge mit unendlicher Ergebnismenge 1 bis ∞ variieren, d. h. die Ergebnismenge wäre hier durch die Menge $\Omega = \{1, 2, 3, \dots\} = \mathbb{N}$ der natürlichen Zahlen gegeben. Eine Ergebnismenge Ω mit *nicht-endlicher* Anzahl von Elementen resultiert ebenfalls, wenn man ein Aktienpaket besitzt und dieses so viele Tage halten will, wie der Verkaufswert eine bestimmte Schranke nicht überschritten hat. Die Überschreitung der kritischen Schranke kann hier schon am ersten Tag, nach einiger Zeit oder nie eintreten.

Zufallsvorgänge können unter *kontrollierten* oder *nicht-kontrollierten* Bedingungen ablaufen. Im erstgenannten Fall spricht man von einem **Zufallsexperiment**. Die Ziehung der Lottozahlen ist unter gleichbleibenden Bedingungen wiederholbar und daher ein Beispiel für ein kontrolliertes Zufallsexperiment. Die Durchschnittstemperatur im Monat Juli an einem bestimmten Ort ist hingegen das Ergebnis eines Zufallsprozesses, das unter nicht-kontrollierten Bedingungen zustande kommt.

Unabhängig davon, ob ein Zufallsprozess unter kontrollierten Bedingungen abläuft oder nicht, ist man i. d. R. daran interessiert, die Chance für das Eintreten von Ereignissen A anhand einer Maßzahl $P(A)$ zu bewerten, die nicht von subjektiven Einschätzungen abhängt und im folgenden als **Wahrscheinlichkeit** für das Eintreten eines Ereignisses A angesprochen wird.¹ In der Alltagssprache wird der Begriff „Wahrscheinlichkeit“ häufig mit subjektiven Einschätzungen für das Eintreten von Ereignissen verbunden, etwa bei der Prognose des morgigen Wetters. In der Statistik wird der Wahrscheinlichkeitsbegriff hingegen objektiv quantifiziert. Dabei stützt man sich, wie inzwischen jeder Teilbereich der modernen Mathematik, auf ein System von Grundannahmen – sogenannte *Axiome*.

Der heute gängige Wahrscheinlichkeitsbegriff der Statistik geht auf den russischen Mathematiker Andrej KOLMOGOROFF (1903 - 1987) zurück. Die Bewertung der Chance für das Eintreten eines Ereignisses (Teilmenge der Ergebnismenge Ω einschließlich des unmöglichen Ereignisses \emptyset und des sicheren Ereignisses Ω) erfolgt anhand einer Funktion P , die jedem Ereignis A eine als Wahrscheinlichkeit des Ereignisses A bezeichnete Zahl $P(A)$ zuordnet, welche folgenden Bedingungen genügt:²



Andrej
KOLMOGOROFF

K1: $P(A) \geq 0$ (Nicht-Negativitätsbedingung)

K2: $P(\Omega) = 1$ (Normierung)

K3: $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \emptyset$
(Additivität bei disjunkten Ereignissen).

¹Der Buchstabe „P“ steht für „probability“, das englische Wort für „Wahrscheinlichkeit“. Man findet anstelle von $P(..)$ in der Literatur auch $Pr(..)$ oder $W(..)$.

²Das dritte Axiom ist hier für den Fall formuliert, dass die Ergebnismenge Ω nur endlich viele Elemente enthält. Bei Zufallsvorgängen mit nicht-endlicher Ergebnismenge ist K3 etwas allgemeiner zu fassen und schließt hier auch den Fall der Vereinigung abzählbar unendlich vieler und paarweise disjunkter Ereignisse ein.

Diese als **Axiome von Kolmogoroff** bezeichneten Bedingungen weisen eine auffallende Analogie mit den Eigenschaften relativer Häufigkeiten auf. Auch relative Häufigkeiten sind nicht-negativ und durch 0 nach unten und 1 nach oben begrenzt. Ferner addieren sich bei einem Merkmal, dessen Ausprägungen durch eine Menge $M = \{a_1, a_2, \dots, a_k\}$ beschrieben sind, die relativen Häufigkeiten für je zwei disjunkte Teilmengen von M und die Summe aller relativen Häufigkeiten ist stets 1.

Aus dem Axiomensystem von Kolmogoroff lassen sich einige elementare Rechenregeln für Wahrscheinlichkeiten ableiten. Unter Heranziehung der Venn-Diagramme aus Abbildung 11.1 verifiziert man die Gleichungen

$$P(\bar{A}) = 1 - P(A) \quad (11.2)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (11.3)$$

$$P(A \setminus B) = P(A) - P(A \cap B). \quad (11.4)$$

Rechenregeln für
Wahrscheinlichkeiten

Gleichung (11.3) wird z. B. anhand des dritten Venn-Diagramms aus Abbildung 11.1 verständlich (Diagramm für $A \cup B$). Da A und B hier nicht disjunkt sind, muss man bei der Berechnung von $P(A \cup B)$ die Summe aus $P(A)$ und $P(B)$ um $P(A \cap B)$ vermindern, weil andernfalls die Wahrscheinlichkeit für den Überschneidungsbereich doppelt zählte.

Das Axiomensystem von Kolmogoroff legt also Eigenschaften fest, die für Wahrscheinlichkeiten gelten müssen, und liefert den Ausgangspunkt für die Herleitung von Rechenregeln für Wahrscheinlichkeiten. Es macht vor allem den Wahrscheinlichkeitsbegriff von persönlichen Einschätzungen unabhängig. Allerdings liefert das System noch keinen Ansatzpunkt zur Berechnung von Wahrscheinlichkeiten für Ereignisse. Um Wahrscheinlichkeiten quantifizieren zu können, benötigt man Zusatzinformationen über den jeweiligen Zufallsvorgang. Eine solche Zusatzinformation kann z. B. darin bestehen, dass man weiß, dass die Ergebnismenge die nachstehenden Bedingungen erfüllt:

Berechnung von
Wahrscheinlichkeiten
erfordert zusätzliche
Information

L1: Die Ergebnismenge ist endlich, also $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$.

L2: Die Wahrscheinlichkeiten für die n Elementarereignisse sind gleich.

Bedingungen für
Laplace-Experimente

Ein Zufallsexperiment mit diesen beiden Eigenschaften wird nach dem französischen Mathematiker Simon Pierre LAPLACE (1749 – 1827) auch **Laplace-Experiment** genannt. Dieser berechnete unter den einschränkenden Voraussetzungen L1 und L2 die Wahrscheinlichkeit eines Ereignisses A als Quotient aus der Anzahl der für A „günstigen“ Fälle und

der Anzahl aller möglichen Ergebnisse des Zufallsexperiments.³



Simon Pierre
LAPLACE

Eine nach (11.5) berechnete Wahrscheinlichkeit erfüllt stets die Bedingungen K1 – K3. Der Ansatz von Laplace ist folglich mit dem Axiomensystem von Kolmogoroff verträglich, betrifft aber nur eine spezielle Gruppe von Zufallsvorgängen. Dass K1 und K2 bei Gültigkeit von (11.5) erfüllt sind, folgt z. B. sofort daraus, dass der Zähler in (11.5) stets einen Wert besitzt, der zwischen 0 und dem Wert des Nenners liegt.

Gleichung (11.5) liefert für viele Anwendungen – etwa bei Glücksspielen – eine leicht handhabbare und sehr nützliche Rechenformel. Eine Definition des Begriffs „Wahrscheinlichkeit“ stellt (11.5) in Verbindung mit L1 und L2 aber nicht dar, weil der zu erklärende Begriff der Wahrscheinlichkeit schon in die Annahme L2 eingeht.

Beispiel 11.2: Wahrscheinlichkeiten bei Laplace-Experimenten

Mit dem Laplace-Ansatz kann man z. B. die Wahrscheinlichkeit für Ereignisse beim Würfeln, bei Münzwürfen oder beim Roulette bestimmen. Die Ergebnismenge ist hier endlich, d. h. die Bedingung L1 ist erfüllt. Damit auch L2 erfüllt ist, sei die Gleichwahrscheinlichkeit der Elementarereignisse vorausgesetzt – bei Würfelspielen oder bei Münzwürfen spricht man auch von der Verwendung „fairer“ Würfel resp. Münzen.

Beim Würfeln mit *einem Würfel* ist dann z. B. die Wahrscheinlichkeit für

$$A = \{5, 6\} = \{\text{Augenzahl ist größer als } 4\}$$



Interaktives Objekt
„Augensummen“

(mit Modell)

durch $P(A) = \frac{2}{6} = \frac{1}{3} \approx 0,333$ gegeben, weil von den 6 möglichen Ausgängen genau 2 für A „günstig“ sind, nämlich die Augenzahlen 5 und 6. Auch die Wahrscheinlichkeit für den Eintritt des Komplementärereignisses $\bar{A} = \Omega \setminus A$ lässt sich nach (11.5) ermitteln als $P(\bar{A}) = \frac{4}{6} = \frac{2}{3}$ oder anhand von (11.2) gemäß $P(\bar{A}) = 1 - \frac{1}{3} = \frac{2}{3} \approx 0,667$. Beim Würfeln mit *zwei Würfeln* ergibt sich für die Wahrscheinlichkeit des Ereignisses

$$A = \{\text{Augensumme aus beiden Würfen ist höchstens } 3\},$$

der Wert $P(A) = \frac{3}{36} = \frac{1}{12} \approx 0,0833$, weil die Ergebnismenge Ω hier 36 Elementarereignisse umfasst, von denen 3 als „günstig“ einzustufen sind.

Beim *dreifachen Münzwurf* kann man die Wahrscheinlichkeit für

$$A = \{\text{Bei drei Münzwürfen tritt zweimal „Zahl“ auf}\}$$

³Die Bezeichnung „günstig“ ist wertfrei (neutral) zu verstehen, kann sich also sowohl auf ein willkommenes Lottoereignis als auch auf das Vorliegen einer Erkrankung beziehen, und bedeutet lediglich „ A ist eingetreten“.

ebenfalls anhand des Laplace-Ansatzes (11.5) berechnen. Die Ergebnismenge Ω ist bei erneuter Verwendung von „Z“ für „Zahl“ und „K“ für „Kopf“ durch die acht Tripel

$$\Omega = \{(Z, Z, Z), (Z, Z, K), (Z, K, Z), (K, Z, Z), \\ (K, K, Z), (K, Z, K), (Z, K, K), (K, K, K)\}$$

Aufgabe 11.3

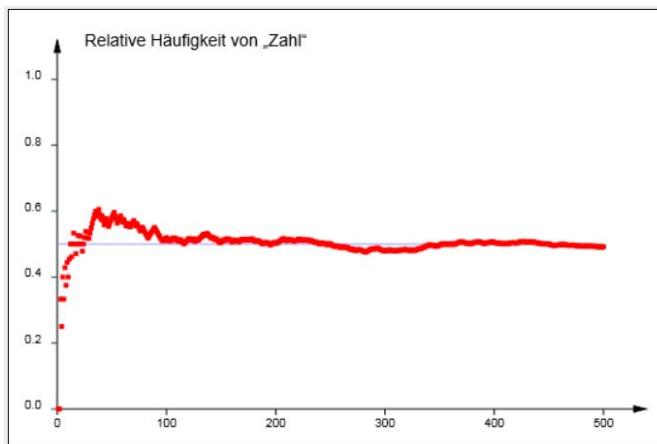


gegeben. Jedes Tripel besitzt bei der hier getroffenen Annahme einer „fairen“ Münze die gleiche Eintrittswahrscheinlichkeit. Es gilt dann $P(A) = \frac{3}{8} = 0,375$, weil bei 3 der 8 Elementarereignisse „Zahl“ zweifach auftritt.

Ein anderer Ansatz zur Berechnung der Wahrscheinlichkeit $P(A)$ für ein Ereignis A beinhaltet – unter der Voraussetzung der beliebigen Wiederholbarkeit eines Zufallsexperiments unter konstanten Bedingungen – die Bestimmung von $P(A)$ als Grenzwert der relativen Häufigkeit für das Eintreten von A . Wenn man z. B. eine „faire“ Münze n -mal wirft und die relativen Häufigkeiten $f_j(\text{Zahl})$ für das Eintreten von „Zahl“ während dieses Zufallsexperiments verfolgt ($j = 1, 2, \dots, n$), so stellt sich am Ende ein Wert $f_n(\text{Zahl})$ ein, der sich tendenziell dem Wert 0,5 nähert und zwar umso deutlicher, je größer man n wählt.

Wahrscheinlichkeiten
als Grenzwert
relativer Häufigkeiten

Abbildung 11.2 zeigt den Verlauf eines virtuellen Münzwurfexperiments mit $n = 500$. Man kann hier, anders als in der Realität, die Wahrscheinlichkeit $p := P(\text{Zahl})$ für „Zahl“ und damit auch die von $1-p = P(\text{Kopf})$ für „Kopf“ (oder für „Wappen“) bei der Programmierung festlegen und folglich als bekannt voraussetzen.



Interaktives Objekt
„Münzwurf“

Abb. 11.2: Entwicklung der relativen Häufigkeiten für „Zahl“ bei 500-fachem Wurf einer fairen Münze



Interaktives Objekt
„Ereignisse beim
Würfeln“



Da man ein Zufallsexperiment in der Praxis nicht unendlich oft, sondern nur n -mal durchführen kann, verwendet man f_n als Approximation (Schätzwert) für die interessierende Wahrscheinlichkeit, wobei die Schätzgüte sich mit wachsendem n tendenziell verbessert. Die Bestimmung von Wahrscheinlichkeiten als Grenzwert⁴

$$f_n(A) \xrightarrow{n \rightarrow \infty} P(A) \quad (11.6)$$

relativer Häufigkeiten bei Zufallsexperimenten mit endlicher Ergebnismenge ist, anders als der Laplace-Ansatz (11.5), nicht an die Bedingung L2 der Gleichwahrscheinlichkeit der Elementarereignisse gebunden. Ein Zufallsexperiment mit zwei Ausgängen A und \bar{A} und unterschiedlichen Eintrittswahrscheinlichkeiten $P(A)$ sowie $P(\bar{A})$ hat man z. B. beim Würfeln mit einem Würfel, wenn man nur zwischen den Ereignissen $A = \{\text{Augenzahl ist } 6\}$ und $\bar{A} = \{\text{Augenzahl ist kleiner als } 6\}$ differenziert. In diesem Falle gilt (11.6) mit $P(A) = \frac{1}{6}$.

11.2 Zufallsstichproben und Kombinatorik

Um die Wahrscheinlichkeit $P(A)$ von Ereignissen A bei Laplace-Experimenten nach (11.5) zu berechnen, muss man zunächst die dort im Nenner auftretende Anzahl aller Elementarereignisse bestimmen – die Anzahl im Zähler ergibt sich dann aus weiteren logischen Überlegungen. Hierzu kann man sich der Methoden der **Kombinatorik** bedienen. Diese repräsentiert ein Teilgebiet der Mathematik, das sich mit der Ermittlung der Anzahl von Möglichkeiten bei der Anordnung und Auswahl von Objekten befasst.

- Stichprobenmodelle:
- Ziehen mit und ohne Zurücklegen
 - Ein anschauliches Modell, das in der Kombinatorik zur Herleitung zentraler Ergebnisse für Zufallsvorgänge mit endlicher Ergebnismenge eingesetzt wird, ist das **Urnenmodell**. Man stelle sich ein Gefäß (Urne) mit N durchnummerierten Kugeln vor, von denen n zufällig ausgewählt werden. Die Auswahl der Kugeln ist als Ziehung einer **Zufallsstichprobe** des Umfangs n aus einer Grundgesamtheit mit N Elementen zu interpretieren. Wenn jede denkbare Stichprobe des Umfangs n mit gleicher Wahrscheinlichkeit realisiert wird, liegt eine **einfache Zufallsstichprobe** vor. Wieviele Möglichkeiten der Auswahl der n Elemente es gibt, hängt davon ab, ob jedes Element der Stichprobe einzeln gezogen und nach der Ziehung wieder zurückgelegt wird oder ob ohne Zurücklegen ausgewählt wird. Im ersten Fall spricht man von einer **Stichprobenziehung mit Zurücklegen** oder vom **Urnenmodell mit Zurücklegen**. Der zweite

⁴Die Konvergenz in (11.6) bezieht sich auf die Konvergenz eines *Zufallsprozesses*, auch *stochastischer Prozess* genannt. Man spricht in diesem Zusammenhang von *stochastischer Konvergenz*. Zur formalen Definition stochastischer Konvergenz vgl. TOUTENBURG / HEUMANN (2008, Abschnitt 5.1).

Fall charakterisiert eine **Stichprobenziehung ohne Zurücklegen** bzw. das **Urnenmodell ohne Zurücklegen**.

Ein n -facher Münzwurf lässt sich z. B. als eine Stichprobenziehung *mit Zurücklegen* interpretieren. Dazu muss man sich eine Urne mit zwei Kugeln vorstellen (je eine mit der Aufschrift „Zahl“ und „Kopf“), aus der n -mal jeweils eine Kugel gezogen und vor der nächsten Ziehung zurückgelegt wird. Die Ausgangssituation ist also bei der Entnahme eines jeden Elements der Stichprobe unverändert – stets befinden sich zwei Kugeln in der Urne. Ein Beispiel für eine Stichprobenziehung *ohne Zurücklegen* ist die Ziehung der Lottozahlen. Hier ist es ausgeschlossen, dass eine Zahl wiederholt gezogen wird. Beim Urnenmodell ohne Zurücklegen ändert sich die Ausgangssituation mit Ziehung jeder Kugel – die Anzahl der auswählbaren Kugeln nimmt mit jedem Auswahlschritt ab.

Die Anzahl der Möglichkeiten, aus einer Urne n Kugeln zu ziehen, wird aber nicht nur davon bestimmt, ob mit oder ohne Zurücklegen gezogen wird. Sie hängt auch davon ab, ob es darauf ankommt, in welcher Reihenfolge die n nummerierten Kugeln gezogen werden. Man unterscheidet hier zwischen einer **Stichprobenziehung mit Berücksichtigung der Anordnung** und einer **Stichprobenziehung ohne Berücksichtigung der Anordnung**. Man kann auch vom **Urnenmodell mit Berücksichtigung der Anordnung** bzw. vom **Urnenmodell ohne Berücksichtigung der Anordnung** sprechen. Wenn die Anordnung berücksichtigt wird, liegt eine **geordnete Auswahl** vor, andernfalls eine **ungeordnete Auswahl**.

- Ziehen mit und ohne Berücksichtigung der Anordnung

Stehen bei der Olympiade im 100-m-Endlauf der Männer 8 Läufer am Start, so kann man die Medaillenvergabe mit der Ziehung einer Stichprobe des Umfangs $n = 3$ aus einer Grundgesamtheit des Umfangs $N = 8$ vergleichen (Ziehen ohne Zurücklegen), wobei die ersten drei gezogenen Kugeln die Medaillengewinner festlegen. Die Reihenfolge ist hier also wesentlich. Bei der Ziehung der Lottozahlen spielt die Reihenfolge, in der die Zahlen gezogen werden, hingegen keine Rolle.

Anzahl der Möglichkeiten einer geordneten Auswahl von n Elementen:

Die Wahrscheinlichkeiten, die man nach (11.5) bestimmt, hängen also davon ab, welche Variante des Urnenmodells zugrunde gelegt wird. Es werde zunächst unter Verwendung des Urnenmodells die Anzahl der möglichen Zufallsstichproben des Umfangs n ermittelt, die sich ergeben, wenn die Reihenfolge der gezogenen Elemente berücksichtigt wird. Zieht man aus einer Urne mit N Kugeln eine Stichprobe des Umfangs n *ohne Zurücklegen*, so gibt es bei der Ziehung der ersten Kugel N Auswahlmöglichkeiten. Bei der zweiten Ziehung gibt es noch $N - 1$ und bei Auswahl der n -ten Kugel nur noch $N - n + 1$ Möglichkeiten. Die Anzahl der Möglichkeiten für die Ziehung einer Zufallsstichprobe des Umfangs n aus N Elementen beträgt somit $N \cdot (N - 1) \cdot \dots \cdot (N - n + 1)$. Dieser Produktterm lässt sich kompakter schreiben, wenn man auf die Kurzschreibweise $N!$ (lies:

- beim Ziehen ohne Zurücklegen

N-Fakultät) und $(N - n)!$ (lies: *N-minus-n-Fakultät*) für das Produkt der ersten N resp. $N - n$ natürlichen Zahlen zurückgreift.⁵ Man erhält dann für die gesuchte Anzahl die Darstellung

$$\begin{aligned} N \cdot (N - 1) \cdot \dots \cdot (N - n + 1) &= \frac{N \cdot (N - 1) \cdot \dots \cdot 1}{(N - n) \cdot (N - n - 1) \cdot \dots \cdot 1} \\ &= \frac{N!}{(N - n)!} \end{aligned} \quad (11.7)$$

Zieht man hingegen aus einer mit N Kugeln gefüllten Urne nacheinander n Kugeln *mit Zurücklegen*, so gibt es für die Auswahl jeder einzelnen Kugel stets N Möglichkeiten. Die Gesamtzahl der Möglichkeiten für die Ziehung einer Zufallsstichprobe des Umfangs n aus N Elementen ist nun gegeben durch

$$\underbrace{N \cdot N \cdot \dots \cdot N}_{n\text{-mal}} = N^n. \quad (11.8)$$

Es bleibt noch die Anzahl der möglichen Zufallsstichproben des Umfangs n für den Fall zu bestimmen, dass die Reihenfolge der gezogenen Elemente keine Rolle spielt. Wieder sei zuerst der Fall der Ziehung *ohne Zurücklegen* betrachtet. Wenn man n nummerierte Kugeln hat, gibt es $n!$ Möglichkeiten, diese anzutragen. Man nennt die verschiedenen Anordnungen auch **Permutationen** der n Elemente. Für kleine Werte von n kann man leicht verifizieren, dass es $n!$ Anordnungsmöglichkeiten gibt. Für beliebiges n lässt sich die Aussage durch vollständige Induktion beweisen.⁶ Der Bruchterm $\frac{N!}{(N-n)!}$ aus (11.7), der unterschiedliche Anordnungen der n Stichprobenelemente berücksichtigt, ist also durch $n!$ zu dividieren, wenn die Reihenfolge der Elemente keine Rolle spielt. Man erhält so

$$\frac{\frac{N!}{(N-n)!}}{n!} = \frac{N!}{(N-n)! \cdot n!}.$$

Der rechtsstehende Term wird **Binomialkoeffizient** genannt und mit $\binom{N}{n}$ (lies: *N über n*) abgekürzt.⁷ Es gilt also

$$\binom{N}{n} := \frac{N!}{(N-n)! \cdot n!}. \quad (11.9)$$

⁵Ist k eine natürliche Zahl, so bezeichnet $k! := 1 \cdot 2 \cdot \dots \cdot k$ das Produkt aus allen natürlichen Zahlen von 1 bis k . Für 0 ist die Fakultät durch $0! = 1$ definiert.

⁶Die vollständige Induktion ist ein elegantes Beweisverfahren der Mathematik, mit dem man Aussagen herleiten kann, die für alle natürlichen Zahlen gelten. Die Grundidee besteht darin, die Gültigkeit der betreffenden Aussage für $n = 1$ zu verifizieren und dann zu zeigen, dass aus der Annahme der Gültigkeit der Aussage für ein beliebiges n auch die Gültigkeit der Aussage für $n + 1$ folgt.

⁷Der Binomialkoeffizient $\binom{N}{n}$ gibt die Anzahl der Möglichkeiten an, aus einer Menge mit N Elementen n Elemente ohne Zurücklegen und ohne Berücksichtigung der Reihenfolge zu ziehen. Es ist $\binom{N}{0} = 1$, $\binom{N}{1} = N$ und $\binom{N}{N} = 1$.

- beim Ziehen mit Zurücklegen

Anzahl der Möglichkeiten einer ungeordneten Auswahl von n Elementen:

- beim Ziehen ohne Zurücklegen

Für den Fall der zufälligen Auswahl von n aus N Elementen mit Zurücklegen sei die Anzahl der Möglichkeiten ohne Beweis angegeben – vgl. z. B. MOSLER / SCHMID (2011, Abschnitt 1.2.3). Sie ist gegeben durch

- beim Ziehen mit Zurücklegen

$$\binom{N+n-1}{n} = \frac{(N+n-1)!}{(N-1)! \cdot n!}. \quad (11.10)$$

Tabelle 11.1 fasst die Ergebnisse für die vier betrachteten Fälle zusammen.

	Ziehen <i>ohne</i> Zurücklegen	Ziehen <i>mit</i> Zurücklegen
Ziehen <i>mit</i> Berücksichti- gung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
Ziehen <i>ohne</i> Berücksichti- gung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

Tab. 11.1: Anzahl der Möglichkeiten der Ziehung einer Stichprobe des Umfangs n aus einer Grundgesamtheit mit N Elementen

Beispiel 11.3: Varianten von Stichprobenziehungen

Für die Formeln (11.7) - (11.10) sei je ein Anwendungsbeispiel genannt und durchgerechnet. Als Beispiel für die Anwendung von (11.7) lässt sich die Bestimmung der Anzahl der Möglichkeiten für die Verteilung der Gold-, Silber- und Bronzemedaillen beim 100-m-Endlauf der Männer bei der Olympiade anführen. In der Terminologie des Urnenmodells werden $n = 3$ Kugeln aus einer Urne mit $N = 8$ nummerierten Kugeln *ohne Zurücklegen* gezogen und *mit Berücksichtigung der Anordnung*. Man erhält also

$$\frac{8!}{(8-3)!} = \frac{8 \cdot 7 \cdot \dots \cdot 1}{5 \cdot 4 \cdot \dots \cdot 1} = 8 \cdot 7 \cdot 6 = 336.$$

Zur Illustration der Anwendung von (11.8) kann das Würfeln mit zwei Würfeln herangezogen werden, etwa das simultane Werfen je eines roten und eines grünen Würfels. In Beispiel 11.1 wurde bereits die Ereignismenge Ω dieses Zufallsexperiments wiedergegeben. Die Menge Ω umfasst 36 Zahlenpaare $(i; j)$, wobei i die Augenzahl des ersten und j die des zweiten Würfels darstellt ($i = 1, 2, \dots, 6$; $j = 1, 2, \dots, 6$). Das Zufallsexperiment lässt sich als Ziehen einer Stichprobe des Umfangs $n = 2$ aus einer Grundgesamtheit des Umfangs $N = 6$ *mit Zurücklegen* und *mit Berücksichtigung der Reihenfolge* interpretieren. Die Anzahl der möglichen Ausgänge ergibt sich daher auch nach (11.8) als $6^2 = 36$.

Die Anzahl der möglichen Ausgänge beim deutschen Zahlenlotto lässt sich anhand von (11.9) ermitteln, weil es hier um eine Stichprobenziehung *ohne*

Zurücklegen und *ohne Berücksichtigung der Anordnung* geht. Es resultiert

$$\binom{49}{6} = \frac{49!}{43! \cdot 6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13\,983\,816.$$

Die Wahrscheinlichkeit dafür, 6 Richtige zu erzielen, ist also extrem gering. Man errechnet mit (11.5) den Wert

$$\frac{1}{13\,983\,816} \approx 0,0000000715 = 7,15 \cdot 10^{-8}.$$

Als Beispiel für die Anwendung von (11.10) sei die Wahl eines Unternehmensvorstands genannt, bei der 3 Bewerber $B1$, $B2$ und $B3$ zur Auswahl stehen.



Aufgabe 11.4

Die Mitglieder des Auswahlgremiums setzen bei einer geheimen Wahl auf dem Wahlzettel 2 Kreuze, wobei zwei verschiedene Kandidaten je einmal oder ein Bewerber zweimal angekreuzt werden kann (Möglichkeit der Stimmenhäufung). Es sei vorausgesetzt, dass weder Enthaltungen noch ungültige Wahlzettel auftreten. Der Wahlvorgang entspricht in der Sprache des Urnenmodells der Ziehung einer Stichprobe des Umfangs $n = 2$ aus einer Urne mit $N = 3$ Kugeln (*Ziehen mit Zurücklegen und ohne Berücksichtigung der Anordnung*). Die Anzahl der Wahlmöglichkeiten ist also durch (11.10) bestimmt und man erhält

$$\binom{3+2-1}{2} = \binom{4}{2} = \frac{4!}{2! \cdot 2!} = 6.$$

Die 6 Elemente der Ergebnismenge Ω lassen sich bei diesem einfachen Beispiel mit nur 3 Kandidaten leicht angeben. Es gilt offenbar

$$\Omega = \{(B1, B1), (B1, B2), (B2, B2), (B1, B3), (B2, B3), (B3, B3)\}.$$

11.3 Bedingte Wahrscheinlichkeiten

In Abschnitt 9.1 wurden Häufigkeiten auch auf Teilmengen einer Population bezogen. So wurde in den Beispielen 9.1 – 9.2 bei der relativen Häufigkeit für die Wahl einer Partei X in einer Grundgesamtheit von befragten Personen nach dem Geschlecht Y der Befragten differenziert. Dies führte zu bedingten Häufigkeiten, z. B. zur bedingten relativen Häufigkeit $f_X(a_1|b_2)$ dafür, dass eine Person die Partei $X = a_1$ wählte (CDU/CSU) und der Bedingung $Y = b_2$ genügte (Person ist weiblich).

Bedingte Wahrscheinlichkeiten

In ähnlicher Weise kann man bei der Berechnung von Wahrscheinlichkeiten nach (11.5) innerhalb der Ergebnismenge Ω eine Teilmenge herausgreifen, für die eine Zusatzbedingung erfüllt ist, und diese Zusatzinformation bei der Wahrscheinlichkeitsberechnung nutzen. Will man etwa bei einer unbekannten Familie mit zwei Kindern die Wahrscheinlichkeit $P(A)$ angeben, dass beide Kinder Mädchen sind, käme man bei Annahme der Gleichwahrscheinlichkeit der Geburt eines Jungen und eines Mädchens

und Fehlen von Zusatzinformation nach (11.5) auf den Wert $\frac{1}{4}$, weil es vier Elementarereignisse $(J, J), (J, M), (M, J), (M, M)$, gibt, von denen eines als „günstig“ im Sinne des Eintritts des Ereignisses A ist. Hat man aber bereits die Information B , dass auf jeden Fall eines der Kinder ein Mädchen ist, wird man den Fall (J, J) bei der Berechnung der gesuchten Wahrscheinlichkeit ausschließen, die Anzahl der möglichen Ergebnisse im Nenner von (11.5) also nur noch auf die für das Ereignis B günstigen Fälle beziehen, und so auf den Wert $\frac{1}{3}$ kommen. Die mit der Vorinformation B berechnete Wahrscheinlichkeit wird **bedingte Wahrscheinlichkeit** von A unter der Bedingung B genannt und mit $P(A|B)$ abgekürzt (lies: *Wahrscheinlichkeit von A unter der Bedingung B*). Man erhält die bedingte Wahrscheinlichkeit $P(A|B)$ als

$$P(A|B) = \frac{\text{Anzahl der für } A \cap B \text{ günstigen Ergebnisse}}{\text{Anzahl der für } B \text{ günstigen Ergebnisse}}. \quad (11.11)$$

Da die Wahrscheinlichkeit $P(A)$ für den Eintritt von A durch (11.5) erklärt ist, gilt analog für die Wahrscheinlichkeiten $P(A \cap B)$ und $P(B)$

$$P(A \cap B) = \frac{\text{Anzahl der für } A \cap B \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}$$

$$P(B) = \frac{\text{Anzahl der für } B \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}.$$

Multipliziert man den Bruchterm in der Formel für $P(A \cap B)$ mit dem Kehrwert $\frac{1}{P(B)}$ des Bruchterms der letzten Gleichung, resultiert

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (11.12)$$

Zusammenhang zwischen bedingten Wahrscheinlichkeiten

Analog gilt für die bedingte Wahrscheinlichkeit $P(B|A)$ die Darstellung

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (11.13)$$

Die Formeln (11.12) und (11.13) kann man verwenden, um $P(A \cap B)$ zu berechnen, wenn $P(A|B)$ und $P(B)$ resp. $P(B|A)$ und $P(A)$ bekannt sind. Auflösen dieser Gleichungen nach $P(A \cap B)$ liefert ja

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A). \quad (11.14)$$

Für die bedingten Wahrscheinlichkeiten $P(A|B)$ und $P(B|A)$ gilt also

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}. \quad (11.15)$$

Unabhängigkeit von Ereignissen

Diese Aussage wird nach dem englischen Mathematiker und Pfarrer THOMAS BAYES (1702 - 1761) auch **Satz von Bayes** genannt.⁸ Zwei zufällige Ereignisse A und B werden als **unabhängig** oder auch als **stochastisch unabhängig** bezeichnet, wenn das Eintreten eines Ereignisses, etwa B , keinen Einfluss auf das andere Ereignis hat. Formal bedeutet dies, dass $P(A|B)$ und $P(A|\bar{B})$ beide mit $P(A)$ identisch sind. Man kann in diesem Falle in (11.12) den Term $P(A|B)$ durch $P(A)$ ersetzen und erhält dann nach Multiplikation mit $P(B)$

$$P(A \cap B) = P(A) \cdot P(B). \quad (11.16)$$

Zwei zufällige Ereignisse A und B sind also genau dann unabhängig, wenn sie der Bedingung (11.16) genügen. Unabhängig sind z. B. die Ergebnisse zweier aufeinanderfolgender Roulettespiele oder Münzwürfe.

Zur Berechnung bedingter Wahrscheinlichkeiten kann man alternativ zur Verwendung der vorstehenden Formeln auch Baumdiagramme oder Kontingenztabellen für absolute Häufigkeiten mit Randverteilungen heranziehen (vgl. hierzu das folgende Beispiel und die Aufgaben 11.5-7).

Beispiel 11.4: Berechnung bedingter Wahrscheinlichkeiten

Es sei eine Gruppe von 60 drogenabhängigen Personen betrachtet, die stationär (Ereignis A) oder ambulant (Ereignis \bar{A}) behandelt werden.⁹ Alle Personen werden einem HIV-Test unterzogen. Bei 15 Personen fällt der Test positiv aus (Ereignis B), bei den anderen 45 negativ (Ereignis \bar{B}). Von den HIV-positiv getesteten Personen sind 80% in stationärer Behandlung, während von den HIV-negativ getesteten Personen nur 40% stationär therapiert werden.

Wählt man zufällig eine der 60 Personen aus, so sind

- $P(B) = \frac{15}{60} = 0,25$ und $P(\bar{B}) = \frac{45}{60} = 0,75$ die Wahrscheinlichkeiten dafür, dass diese Person HIV-positiv resp. HIV-negativ ist;
- $P(A|B) = 0,8$ und $P(A|\bar{B}) = 0,4$ die Wahrscheinlichkeiten dafür, dass eine HIV-positiv resp. HIV-negativ getestete Person in stationärer Behandlung ist.

Die Gleichung $P(A|B) = 0,8$ ergibt sich z. B. aus der Vorinformation, dass in der Gruppe der HIV-positiv getesteten Personen 80% in stationärer Behandlung sind. Die Wahrscheinlichkeit $P(A \cap B)$ dafür, dass die zufällig ausgewählte

⁸Der Satz von Bayes ist ein Grundpfeiler eines als **Bayes-Statistik** bezeichneten Zweiges der Statistik, der sich als Alternative zur klassischen Statistik versteht und mit einem von (11.5) abweichenden Wahrscheinlichkeitsbegriff arbeitet. Eine umfassende Einführung in die Bayes-Statistik findet man bei TSCHIRK (2018).

⁹Dieses Beispiel ist adaptiert aus CAPUTO / FAHRMEIR / KÜNSTLER / LANG / PIGEOT / TUTZ (2009, Kapitel 4).

Person stationär therapiert wird und auch HIV-positiv ist, lässt sich aus (11.14) gewinnen, wenn man dort die Werte für $P(A|B)$ und $P(B)$ einsetzt:

$$P(A \cap B) = P(A|B) \cdot P(B) = 0,8 \cdot 0,25 = 0,2.$$

Analog verifiziert man für die Wahrscheinlichkeit $P(A \cap \bar{B})$, dass die ausgewählte Person stationär therapiert wird und HIV-negativ ist, den Wert

$$P(A \cap \bar{B}) = P(A|\bar{B}) \cdot P(\bar{B}) = 0,4 \cdot 0,75 = 0,3.$$

Die Wahrscheinlichkeit $P(A)$ dafür, dass die ausgewählte Person – gleich ob HIV-positiv oder HIV-negativ getestet – stationär behandelt wird, setzt sich dann additiv aus den beiden Wahrscheinlichkeiten $P(A \cap B)$ und $P(A \cap \bar{B})$ zusammen. Dies folgt aus dem Axiom K3 von Kolmogoroff. Dieses ist anwendbar, weil – vgl. die rechte Hälfte der vierteiligen Abbildung 11.1 – die Mengen $A \cap B$ und $A \cap \bar{B}$ disjunkt sind und ihre Vereinigung A ergibt. Es gilt also

$$P(A) = P(A \cap B) + P(A \cap \bar{B}) = 0,3 + 0,2 = 0,5.$$

Die Ereignisse A und B sind abhängig, weil andernfalls (11.16) gelten müsste.

Die vorstehenden Berechnungen sind transparenter, wenn man ein Baumdiagramm oder eine Vierfeldertafel mit Randverteilungen heranzieht. Die im Vorspann dieses Beispiels vermittelte Information lässt sich z. B. bei Verwendung einer Vierfeldertafel wie folgt darstellen (vgl. auch Tabelle 9.5):

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
stationär (A)	12	18	30
ambulant (\bar{A})	3	27	30
Spaltensummen	15	45	60

Tab. 11.2: Vierfeldertafel für vier Kategorien von Suchtpatienten

Die kursiv gesetzten Zahlen sind entweder explizit im Text aufgeführt oder waren als relative Häufigkeiten vorgegeben und unter Berücksichtigung von $n = 60$ in absolute Häufigkeiten umzurechnen (80 % resp. 40 % der Grundgesamtheit).

Aus Tabelle 11.2 ergibt sich die Wahrscheinlichkeit $P(A|B)$, dass eine HIV-positive Person stationär behandelt wird, als Quotient $\frac{12}{15} = 0,8$. Analog liest man aus der Vierfeldertafel für die Wahrscheinlichkeit $P(A \cap B)$, dass eine Person sowohl stationär behandelt als auch positiv auf HIV getestet wird, unmittelbar das Ergebnis $\frac{12}{60} = 0,2$ ab.



Aufgaben 11.5-7

Exkurs 11.1: Das „Ziegenproblem“

Eine interessante Denkaufgabe ist das sog. „Ziegenproblem“, im angelsächsischen Sprachraum auch „Monty-Hall-Problem“ genannt. Es wurde in der Ausgabe vom 18. 11. 2004 der Wochenzeitung *Die Zeit* wie folgt beschrieben:

Sie sind Kandidat einer Fernsehshow und dürfen eine von drei verschlossenen Türen auswählen. Hinter einer der Türen wartet der Hauptgewinn, ein prachtvolles Auto, hinter den anderen beiden steht jeweils eine Ziege. Frohgemut zeigen Sie auf eine der Türen, sagen wir Nummer 1. Doch der Showmaster, der weiß, hinter welcher Tür sich das Auto befindet, lässt sie nicht sofort öffnen, sondern sagt geheimnisvoll: „Ich zeige Ihnen mal was!“ Er lässt eine andere Tür öffnen, sagen wir Nummer 3 - und hinter dieser steht eine Ziege. Nun fragt der Showmaster lauernd: „Bleiben Sie bei Tür Nummer 1, oder wählen Sie doch lieber Nummer 2?“ Was sollten Sie tun?

Der Showmaster interveniert also, *bevor* die vom Kandidaten gewählte Tür geöffnet wird. Es wird unterstellt, dass der Showmaster stets

- die Tür mit der zweiten Ziege öffnet, wenn sich der Kandidat bei seiner Wahl von Tür 1 für eine Tür mit einer Ziege entschieden hat;
- zufällig eine der beiden Türen auswählt, hinter denen eine Ziege steht, wenn sich der Kandidat mit der Wahl von Tür 1 auf Anhieb für die Tür mit dem Auto entschieden hat.

Es bezeichne A_i das Ereignis, dass das Auto hinter der i -ten Tür steht ($i = 1, 2, 3$) und S_2 und S_3 das Ereignis, dass der Showmaster nach Wahl von Tür 1 durch den Kandidaten die Tür 2 resp. Tür 3 öffnet. Da der Kandidat am Anfang keine Zusatzinformation hat, gilt für ihn nach (11.5)

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}.$$

Der Kandidat hat sich zunächst für Tür 1 entschieden und diese Entscheidung führt mit Wahrscheinlichkeit $P(A_1) = \frac{1}{3}$ zum Gewinn des Autos. Man könnte meinen, dass die Chancen für die richtige Entscheidung zwischen den nach der Intervention des Showmasters verbleibenden Türen gleich groß sind. Dies ist allerdings nicht korrekt, denn die Zusatzinformation ändert nichts an der Ausgangssituation $P(A_1) = \frac{1}{3}$. Es gilt weiterhin $P(A_2) + P(A_3) = \frac{2}{3}$, nur weiß man jetzt, dass einer der Summanden $P(A_2)$ und $P(A_3)$ den Wert 0 hat. Ein Wechsel der ursprünglich gewählten Tür verdoppelt somit die Gewinnchancen.

Man kann diesen Sachverhalt auch so verdeutlichen: Trifft A_1 zu, so führt ein Festhalten an der Entscheidung für Tür 1 zum Gewinn des Autos. Wäre A_2 zutreffend, hätte der Showmaster Tür 3 geöffnet und ein Wechsel wäre hier angezeigt. Gleiches gilt für den Fall A_3 . Hier würde der Showmaster Tür 2 öffnen und wiederum wäre eine Korrektur der ursprünglichen Entscheidung von Vorteil. In zwei von drei Fällen wäre also ein Umentscheiden besser.

11.4 Sensitivität und Spezifität

Bedingte Wahrscheinlichkeiten spielen u. a. in der Medizin eine wichtige Rolle. Es sei das Merkmal „Gesundheitsstatus“ betrachtet, für das nur zwischen den Ausprägungen „gesund“ und „krank“ unterschieden werde. In Abschnitt 9.1 hatten wir solche Merkmale mit genau zwei Ausprägungen **binär** oder **dichotom** genannt. Differenziert man beim Gesundheitsstatus zwischen der nicht direkt beobachtbaren Variablen „wahrer Gesundheitsstatus“ und dem beobachtbaren Merkmal „Gesundheitsstatus laut einem Test- oder Diagnosebefund“, hat man zwei dichotome Merkmale mit den Ausprägungen „tatsächlich gesund“ (Zustand A) und „tatsächlich krank“ (Zustand \bar{A}) resp. „Testergebnis negativ / für gesund befunden“ (Ereignis B) und „Testergebnis positiv / für krank befunden“ (Ereignis \bar{B}). Die Häufigkeiten für die möglichen vier Ausprägungskombinationen beider Merkmale in einer N Personen umfassenden Grundgesamtheit lassen sich in einer **Vierfeldertafel** (Spezialfall einer **Kontingenztabelle**) zusammenfassen.

Um die Vierfeldertafel für die beiden dichotomen Merkmale übersichtlich zu präsentieren, seien nachstehende Abkürzungen eingeführt:

$r_p :=$ Anzahl der als krank diagnostizierten Personen, die tatsächlich krank sind (Befund ist richtig-positiv);

$f_n :=$ Anzahl der als gesund diagnostizierten Personen, die in Wahrheit krank sind (Befund ist falsch-negativ);

$f_p :=$ Anzahl der als krank diagnostizierten Personen, die in Wahrheit gesund sind (Befund ist falsch-positiv);

$r_n :=$ Anzahl der als gesund diagnostizierten Personen, die tatsächlich gesund sind (Befund ist richtig-negativ).

Mit diesen Bezeichnungen kann man die Vierfeldertafel einschließlich der Randverteilungen wie folgt schreiben:

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächlich krank (A)	r_p	f_n	$r_p + f_n$
Tatsächlich gesund (\bar{A})	f_p	r_n	$f_p + r_n$
Spaltensummen	$r_p + f_p$	$f_n + r_n$	N

Tab. 11.3: Vierfeldertafel mit Häufigkeiten für richtige und falsche Befunde bei medizinischen Test- und Diagnoseverfahren

Die Randverteilung des Merkmals „wahrer Gesundheitsstatus“ ist durch die beiden Zeilensummen gegeben, also durch die Anzahl $r_p + f_n$ aller tatsächlich Kranken und die Anzahl $f_p + r_n$ aller tatsächlich Gesunden. Analog ist die Randverteilung des Merkmals „Gesundheitsstatus laut

Test- / Diagnosebefund“ durch die beiden Spaltensummen definiert, d. h. durch die Anzahlen $r_p + f_p$ bzw. $f_n + r_n$ aller vom Verfahren als „krank“ resp. „gesund“ bewerteten Personen.

Gütemaße Es ist wünschenswert, dass der Anteil an richtig-positiven und richtig-negativen Beobachtungen hoch und der Anteil an falsch-negativen und falsch-positiven Befunden niedrig ist. Maßzahlen für die Güte des eingesetzten Diagnoseverfahrens sind dessen Sensitivität und Spezifität. Die **Sensitivität** gibt die Wahrscheinlichkeit dafür an, dass tatsächlich Kranke auch als „krank“ diagnostiziert werden. Sie wird auch **Trefferrate** genannt. Die **Spezifität** bezeichnet hingegen die Wahrscheinlichkeit dafür, dass tatsächlich Gesunde auch als „gesund“ eingestuft werden.

$$\text{Sensitivität} := P(\text{Diagnose „krank“} \mid \text{tatsächlich krank}) = P(B|A)$$

$$\text{Spezifität} := P(\text{Diagnose „gesund“} \mid \text{tatsächlich gesund}) = P(\overline{B}|\overline{A}).$$

Sensitivität beinhaltet einen Vergleich der Anzahl r_p der korrekt als „krank“ diagnostizierten Personen mit der Anzahl $r_p + f_n$ aller tatsächlich Kranken. Spezifität vergleicht die Anzahl r_n der korrekt als „gesund“ diagnostizierten Personen mit der Anzahl $f_p + r_n$ aller tatsächlich Gesunden. Für die vorstehenden bedingten Wahrscheinlichkeiten gelten demnach die – auch mit (11.11) herleitbaren – Darstellungen

$$\text{Sensitivität} = \frac{r_p}{r_p + f_n} \quad (11.17)$$

$$\text{Spezifität} = \frac{r_n}{f_p + r_n}. \quad (11.18)$$

Eng verknüpft mit (11.18) ist die **Fehlalarmrate** eines Diagnoseverfahrens. Sie vergleicht die Anzahl f_p der fälschlich als „krank“ eingestuften Gesunden mit der Anzahl $f_p + r_n$ aller tatsächlich Gesunden:

$$\text{Fehlalarmrate} = \frac{f_p}{f_p + r_n}. \quad (11.19)$$

Offenbar ergänzen sich (11.18) und (11.19) zu 1. Es gilt folglich

$$\text{Fehlalarmrate} = 1 - \text{Spezifität}.$$

Die genannten Gütemaße werden meist als Prozentwerte ausgewiesen.

Aussagekraft der Gütemaße Es reicht bei der Bewertung eines medizinischen Test- oder Diagnoseverfahrens auf keinen Fall, sich lediglich die Sensitivität anzuschauen. Sie sagt nur etwas über die Fähigkeit des Verfahrens aus, Kranke wirklich als „krank“ zu identifizieren. Auch die Spezifität alleine hat nur eine begrenzte Aussagekraft. Sie informiert lediglich über die Fähigkeit des Verfahrens,

Gesunde tatsächlich als „gesund“ einzustufen. Das nachstehende fiktive Beispiel verdeutlicht, dass hohe Werte für Sensitivität und Spezifität nicht die Verlässlichkeit eines Diagnoseverfahrens garantieren.

Beispiel 11.5: Fehlalarme bei der Krebsdiagnose

Ein neues Krebsdiagnoseverfahren werde in einem Massenscreening auf 100 000 Personen angewandt. Das Verfahren zielt auf eine Krebsart, von der 1% der zu untersuchenden Personen betroffen sind. Die latente Variable „wahrer Gesundheitsstatus“ hat demnach bei 1 000 Personen die Ausprägung „tatsächlich erkrankt“ und bei 99 000 Personen die Realisation „tatsächlich gesund“. Das Diagnoseverfahren habe sowohl für die Sensitivität als auch für die Spezifität einen Wert von 99%.

Die hohe Sensitivität von 99% besagt, dass das Diagnoseverfahren bei 99 % der Kranken den Krebs entdeckt - es werden hier 990 der 1 000 Krebserkrankungen identifiziert. Der Wert 99% für die Spezifität beinhaltet, dass das Verfahren bei 99% der Gesunden auch eine Einstufung als „gesund“ erzielt - von den 99 000 Gesunden werden nur 990 Personen (1%) fälschlich als „krank“ eingestuft.

Tabelle 11.4 fasst die Häufigkeiten in einer Vierfeldertafel zusammen. Dabei sind die Fehldiagnosen (990 Fehlalarme und 10 fälschlich unterbliebene Alarne) durch Fettdruck betont.

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächlich krank (A)	990	10	1 000
Tatsächlich gesund (\bar{A})	990	98 010	99 000
Spaltensummen	1 980	98 020	100 000

Tab. 11.4: Vierfeldertafel mit fiktiven Häufigkeiten für richtige und falsche Befunde bei einem Verfahren zur Krebsdiagnose

Demnach hat hier nur jede zweite Person, die das Verfahren mit der schweren Diagnose „krebskrank“ konfrontierte, tatsächlich Krebs. Die Patienten würde bei dem skizzierten Diagnoseverfahren vor allem die bedingte Wahrscheinlichkeit



Aufgaben 11.6-7

$$P(\text{Patient ist krank} \mid \text{Test ist positiv}) = P(A|B) = \frac{990}{1980} = 0,5.$$

eines korrekten Alarms interessieren. Veröffentlicht wird aber zumeist der Wert für die Sensitivität, hier also

$$P(\text{Test ist positiv} \mid \text{Patient ist krank}) = P(B|A) = \frac{990}{1000} = 0,99.$$

Der Grund für die hohe Wahrscheinlichkeit $P(B|\bar{A}) = 0,5$ für das Auftreten falsch-positiver Diagnosen liegt hier am geringen Vorkommen der Krebsart.



**Kritisch
nachgefragt**

Eine Boulevard-Zeitung lobte unlängst einen von der Universität Heidelberg entwickelten Test für Brustkrebs mit einer Trefferrate von 75% als „Welt-sensation“. Dass ein Wert von 75% für die Sensitivität eines medizinischen Diagnoseverfahrens noch nicht viel besagt und kritisch zu hinterfragen ist, verdeutlicht das folgende Beispiel zu einem Diagnoseverfahren für Leberkrebs bei Männern, bei der die Sensitivität des auf 100 000 Patienten angewandten Testverfahrens sogar über 75% liegt. Tabelle 11.5 zeigt die Häufigkeiten für korrekte und inkorrekte Testbefunde in Form einer Vierfeldertafel.

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächlich krank (A)	8	2	10
Tatsächlich gesund (\bar{A})	45 995	53 995	99 990
Spaltensummen	46 003	53 997	100 000

Tab. 11.5: Vierfeldertafel mit Häufigkeiten für richtige und falsche Befunde bei einem Verfahren zur Diagnose von Leberkrebs

Für die Sensitivität $P(B|A)$ errechnet man nach (11.17) den Wert $\frac{8}{10} = 0,80$ und für die Spezifität $P(\bar{B}|\bar{A})$ nach (11.18) den Wert $\frac{53\,995}{99\,990} \approx 0,54$. Die Wahrscheinlichkeit $P(A|B)$ für einen korrekten Alarm hat hier den inakzeptablen Wert $\frac{8}{46\,003} \approx 0,000174$ (0,174%).

Dass bei der Bewertung von Prozentangaben für die Sensitivität und Spezifität medizinischer Tests „Statistical Literacy“ hilfreich ist, zeigt auch ein realitätsnahe Beispiel aus einem Artikel der Wochenzeitschrift *Die Zeit* vom 7. Mai 2020 (Rubrik „Wissen“), bei dem es um einen Antikörpertest auf eine vorausgegangene COVID-19-Erkrankung geht. Ein Schweizer Pharmakonzern bewirbt seit April 2020 einen solchen Test, der eine Sensitivität von 100% und eine Spezifität von 99,8% haben soll. Wenn dies zutrifft, wären in einer Region, in der es bisher kaum COVID-19-Infektionen gab, z. B. nur bei 0,2% der Bevölkerung, bei 1 000 Tests zwei richtig-positive und zwei falsch-positive Befunde zu erwarten. Dies entspräche einer Fehlalarmquote von 50%. Bei einer Infektionsquote von 2% läge der Fehlalarmanteil noch immerhin bei 10%.

11.5 Wahrscheinlichkeitsverteilungen

Eine Kernaufgabe der beschreibenden Statistik besteht darin, Ausprägungen von Merkmalen, also Daten, anhand von Grafiken und Kenngrößen zu charakterisieren. Wirft man z. B. einen Würfel n -mal, so kann man die relative Häufigkeit der Augenzahlen anhand eines Stab- oder Säulendiagramms darstellen. Die beobachteten Häufigkeiten sind das Ergebnis eines Zufallsexperiments. Bei einem Datensatz zu Bruttoverdiensten bietet es sich an, die Daten zu Einkommensklassen zu gruppieren und anhand eines

Interpretation von
Daten als Ergebnis
von Zufallsvorgängen

Histogramms zu präsentieren. Auch hier ist das Ergebnis zufallsabhängig, wenn die Daten anhand einer Zufallsstichprobe gewonnen wurden.

Ein Vorgang, bei dem Zufallseinflüsse ins Spiel kommen, nennt man **stochastisch** – im Gegensatz zu einem **deterministischen** Vorgang, bei dem der Ausgang exakt vorhersagbar ist. Ein Modell, das Zufallseinflüsse berücksichtigt, bezeichnet man auch als **stochastisches Modell**. Bei einem **deterministischen Modell** spielen Zufallseinflüsse hingegen keine Rolle. Bei der Prognose von Aktienkursverläufen wird man mit einem stochastischen Modell arbeiten, bei der Vorhersage des Zeitpunkts der nächsten Sonnenfinsternis mit einem deterministischen Modell. Das Teilgebiet der Statistik, das sich mit der Modellierung von Zufallsvorgängen, und der Berechnung von Wahrscheinlichkeiten anhand solcher Modelle und kombinatorischer Überlegungen befasst, nennt man **Stochastik**, zu der die **Wahrscheinlichkeitsrechnung** und die **schließende Statistik** gehören. Für die schließende Statistik ist die Verwendung *stochastischer* Modelle typisch; die Wahrscheinlichkeitsrechnung ist das Fundament.

Wenn man die Ausprägungen eines Merkmals als Ergebnis eines Zufallsvorgangs interpretiert, spricht man das Merkmal als **Zufallsvariable** an (engl: *random variable*) und die Ergebnisse des Zufallsprozesses als **Ausprägungen** oder **Realisierungen** der betreffenden Zufallsvariablen. Je nachdem, ob das Merkmal, dessen Ausprägungen durch einen Zufallsvorgang vermittelt werden, diskret oder stetig ist, hat man eine diskrete resp. eine stetige Zufallsvariable. Bei einer *diskreten* Zufallsvariablen ist demnach die Anzahl der Ausprägungen abzählbar. Ein Beispiel für eine diskrete Zufallsvariable ist die Anzahl der Richtigen beim Lotto. Bei einer *stetigen* Zufallsvariablen ist die Menge der Ausprägungen hingegen durch ein Intervall gegeben. Die Anzahl der Ausprägungen ist hier nicht mehr abzählbar. Als Beispiel kann die Wartezeit bis zum Auftreten einer bestimmten Zahl beim Roulette oder die Körpergröße einer zufällig aus einer größeren Menschengruppe ausgewählten Person angeführt werden.

Eine Zufallsvariable repräsentiert eine Abbildung, die jedem Ereignis eine Wahrscheinlichkeit zuordnet. Bei einer stetigen Zufallsvariablen geht es darum zu quantifizieren, mit welcher Wahrscheinlichkeit die Variable Realisationen ober- oder unterhalb eines Schwellenwerts oder innerhalb eines Intervalls annimmt. So ist es bei der Zufallsvariablen „Lebensdauer X eines Leuchtmittels“ interessant zu wissen, mit welcher Wahrscheinlichkeit eine Brenndauer von mindestens 1 000 Stunden erreicht wird.

Würfelt man mit einem Würfel n -mal, stellt man bei ausreichend groß gewähltem n fest, dass die relativen Häufigkeiten für die Augenzahlen in guter Näherung identisch sind. Der Befund legt es nahe, bei der Charakterisierung des Experiments mit einem Modell zu arbeiten, bei dem die Eintrittswahrscheinlichkeiten für die einzelnen Augenzahlen als gleich

Diskrete und stetige Zufallsvariablen



Video „Diskrete Zufallsvariablen“



Interaktives Objekt „Gleichverteilung (Würfelexperiment)“

groß angenommen werden. Dieses Modell, das auch als **diskrete Gleichverteilung** angesprochen wird – genauer: als Spezialfall der diskreten Gleichverteilung – ist in Kapitel 12 näher beschrieben.

Verwendung von Modellen Es gibt viele Beispiele für den sinnvollen Einsatz von Verteilungsmodellen. Erfasst man etwa die Körpergröße aller 30-jährigen Männer in Deutschland und präsentiert die Ergebnisse in Form eines Histogramms, wird man eine Gesetzmäßigkeit vermuten und eine glockenförmige Kurve heranziehen, die das Histogramm approximiert. Hinter der Kurve steht das Modell der **Normalverteilung** (vgl. Kapitel 13) – man kennt die Gaußsche Glockenkurve noch vom früheren 10-DM-Schein. Erfasst man die Körpergröße aller 30-jährigen Männer etwa in Japan, kann dasselbe Modell zum Einsatz kommen, Zentrum und Streuung der verwendeten Normalverteilung sind aber nicht notwendigerweise identisch.

Diskrete und stetige Verteilungen Allgemeiner nennt man ein Modell, welches das Verhalten einer Zufallsvariablen vollständig beschreibt, **Wahrscheinlichkeitsverteilung** oder kurz **Verteilung** (engl: *probability distribution*) der betreffenden Zufallsvariablen. Will man diese von der **empirischen Verteilung** eines Merkmals unterscheiden, bezeichnet man sie auch als **theoretische Verteilung**. In Abhängigkeit vom Status der Zufallsvariablen unterscheidet man zwischen **diskreten Verteilungen** (engl.: *discrete distributions*) und **stetigen Verteilungen** (engl.: *continuous distributions*).

Charakterisierung von Verteilungen: Zur vollständigen Beschreibung des Verhaltens einer *beliebigen* Zufallsvariablen X kann man die **Verteilungsfunktion** von X heranziehen, die jedem reellen Wert x eine Wahrscheinlichkeit

$$F(x) := P(X \leq x) \quad (11.20)$$

zuordnet. Zwecks Unterscheidung von der empirischen Verteilungsfunktion (4.5) nennt man die Funktion (11.20) auch präziser **theoretische Verteilungsfunktion**.¹⁰ Zur Charakterisierung einer *empirischen* Verteilung wurde neben der empirischen Verteilungsfunktion die relative Häufigkeitsverteilung herangezogen (vgl. die Abbildungen 4.13 und 4.14). Zur Beschreibung einer *theoretischen* Verteilung kann man außer der Verteilungsfunktion (11.20) ebenfalls eine zweite Funktion heranziehen. Der relativen Häufigkeitsverteilung in der beschreibenden Statistik entspricht bei diskreten Verteilungen die **Wahrscheinlichkeitsfunktion**. Diese verknüpft jede Realisation x einer diskreten Zufallsvariablen X mit einer Eintrittswahrscheinlichkeit $P(X = x)$. Bei stetigen Verteilungen ist das Analagon zur relativen Häufigkeitsverteilung die **Dichtefunktion**. Aus dieser lassen sich für eine stetige Zufallsvariable Aussagen des Typs $P(X \leq x)$, $P(X > x)$ oder $P(a \leq X \leq b)$ ableiten.

¹⁰Um die Notation möglichst einfach zu halten, wird für die empirische und die theoretische Verteilungsfunktion dieselbe Bezeichnung $F(x)$ verwendet.

Wie bei empirischen Verteilungen kann man auch bei theoretischen Verteilungen Kenngrößen angeben, die das Zentrum der Verteilung beschreiben oder die Variabilität der Zufallsvariablen, die dieser Verteilung folgt. Als Lageparameter der Verteilung einer Zufallsvariablen sind der **Erwartungswert** und die theoretischen **Quantile** zu nennen, als Streuungsparameter die theoretische **Standardabweichung** oder deren Quadrat, die theoretische **Varianz**. Der in der Literatur meist unterdrückte Zusatz „theoretisch“ soll hier betonen, dass es in der beschreibenden Statistik analoge Begriffe gibt, die mit dem Zusatz „empirisch“ versehen sind und sich dort auf Häufigkeitsverteilungen beziehen.

Die Variabilität einer Zufallsvariablen lässt sich auch durch Angabe eines Schwankungsintervalls charakterisieren. Ein **Schwankungsintervall** für eine Zufallsvariable X zur Sicherheit $1 - \alpha$ bezeichnet ein Intervall, in das die Ausprägungen von X mit einer vorgegebenen Wahrscheinlichkeit $1 - \alpha$ fallen. Meistens werden Schwankungsintervalle verwendet, bei denen die Wahrscheinlichkeit gleich groß ist, dass eine Ausprägung unterhalb oder oberhalb des Intervalls liegt. Man spricht hier auch von **zentralen Schwankungsintervallen**.

Zwischen empirischen Verteilungen von Merkmalen (Häufigkeitsverteilungen) und theoretischen Verteilungen von Zufallsvariablen gibt es auffällige Analogien, von denen einige in Tabelle 11.6 betont sind:

	Beschreibende Statistik	Wahrscheinlichkeitsrechnung
Bezugsrahmen	Menge aller untersuchungsrelevanten Merkmalsträger	Menge der möglichen Ausprägungen einer Zufallsvariablen
Verteilungen	Empirische Verteilung eines Merkmals, festgelegt durch - relative Häufigkeiten - empirische Verteilungsfunktion	Theoretische Verteilung einer Zufallsvariablen, festgelegt durch - Wahrscheinlichkeits- oder Dichtefunktion - theoretische Verteilungsfunktion
Kenngrößen	Mittelwert, Median, empirische Quantile, empirische Varianz	Erwartungswert, theoretische Quantile, theoretische Varianz

Tab. 11.6: Analogien zwischen empirischen und theoretischen Verteilungen

Kenngrößen
theoretischer
Verteilungen

Eine Alternative zur
Beschreibung des
Streubereichs von
Zufallsvariablen

Wichtig ist aber eine Unterscheidung von Daten- und Modellebene. Verteilungen von Zufallsvariablen sind als Modelle zu verstehen, mit denen man Strukturen und Gesetzmäßigkeiten, die großen Datenmengen zugrunde liegen können, zu approximieren versucht. Kenngrößen theoretischer

Verteilungen – von den Kenngrößen empirischer Verteilungen klar abzugegrenzen – sind in der Praxis aus den Daten zu schätzen. In der Praxis gilt es auch, Hypothesen zu testen, die sich auf Kenngrößen von Wahrscheinlichkeitsmodellen beziehen (vgl. hierzu die Kapitel 15 – 16).

Wichtige
Verteilungen
(mit Anwendungs-
beispielen)

In Kapitel 12 werden einige diskrete und in Kapitel 13 einige stetige Verteilungen ausführlich vorgestellt. Im Folgenden sind beispielhafte Anwendungsfelder der dort behandelten Verteilungsmodelle genannt:

Diskrete Verteilungen:

- Diskrete Gleichverteilung: Glücksspiele (Würfeln, Roulette);
- Binomialverteilung (Bernoulli-Verteilung als Spezialfall): Glücks Spiele (z. B. Münzwurfxperimente), Approximation der hypergeometrischen Verteilung in der Qualitätssicherung;
- Hypergeometrische Verteilung: Glücksspiele (Lotto), Qualitätssicherung (Eingangsprüfungen für Warenlose).

Stetige Verteilungen:

- Stetige Gleichverteilung: Modellierung von Wartezeiten;
- Normalverteilung: Modellierung von Messfehlern, Approximation der Verteilung der Summe unabhängiger Zufallsvariablen, Schadensabschätzung bei Versicherungen;
- χ^2 , t - und F -Verteilung: Testen von Hypothesen (Verteilungsmodell für Prüfgrößen).

Weitere
Verteilungen
(mit Anwendungen)



Neben den vorstehend aufgelisteten Modellen gibt es zahlreiche weitere Verteilungen, von den als weitere *diskrete* Verteilungen die **Poisson-Verteilung** und die **geometrische Verteilung** erwähnt seien. Die Poisson-Verteilung wird zur Modellierung seltener Ereignisse verwendet, die geometrische Verteilung u. a. als Wartezeitverteilung. Als stetige Verteilungen seien noch die **Exponentialverteilung** und die **Lognormalverteilung** genannt. Die Exponentialverteilung findet u. a. in der Technik bei der Analyse der Lebensdauer von Hardwarekomponenten Anwendung, die Lognormalverteilung zur Modellierung von Einkommen und anderer nicht-negativer Merkmale. Eine eingehendere Behandlung dieser hier nur erwähnten Verteilungen findet man bei FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Kapitel 5 – 6). Ob ein bestimmtes Verteilungsmodell, etwa das der Normalverteilung, zur Charakterisierung eines Datensatzes passt, ist Gegenstand von **Anpassungstests**. Diese sind nicht Gegenstand dieses Manuskripts. Es sei aber auf SCHLITTGEN (2012, Kapitel 18) verwiesen.



12 Diskrete Zufallsvariablen



Vorschau auf
das Kapitel

In diesem Kapitel geht es um Verteilungen *diskreter* Zufallsvariablen, also von Zufallsvariablen mit einer abzählbaren Anzahl von Ausprägungen. Die Verteilung diskreter Zufallsvariablen lässt sich anhand der Wahrscheinlichkeitsfunktion oder der theoretischen Verteilungsfunktion charakterisieren. Vorgestellt werden einige spezielle diskrete Verteilungen. Die diskrete Gleichverteilung ist das Verteilungsmodell für eine Zufallsvariable mit k Ausprägungen, die mit gleicher Wahrscheinlichkeit eintreten.

Bei der Bernoulli-Verteilung gibt es nur $k = 2$ Ausprägungen, deren Eintrittswahrscheinlichkeiten aber verschieden sein können. Führt man ein Zufallsexperiment, dessen Ausgang durch eine bernoulli-verteilte Zufallsvariable darstellbar ist (sog. Bernoulli-Experiment), n -fach durch und zählt für einen der beiden Ausgänge die Häufigkeit des Auftretens, folgt die Zählvariable einer Binomialverteilung. Die n -fache Durchführung eines Bernoulli-Experiments entspricht dem n -fachen Ziehen einer Kugel *mit Zurücklegen* aus einer Urne, die N Kugeln in zwei Farben enthält (z. B. rote und grüne Kugeln). Die Zählvariable „Anzahl der roten Kugeln“ ist dann binomialverteilt. zieht man hingegen n -mal *ohne Zurücklegen*, folgt die Zählvariable einer hypergeometrischen Verteilung. Sowohl für die Binomialverteilung als auch für die hypergeometrische Verteilung werden Anwendungen vorgestellt.

12.1 Wahrscheinlichkeits- und Verteilungsfunktion

In Kapitel 2 wurde zwischen diskreten und stetigen Merkmalen unterschieden. Ein Merkmal X wurde als *diskret* bezeichnet, wenn es nur endlich viele, höchstens aber abzählbar unendlich viele Ausprägungen annehmen kann.¹ Wenn man die Ausprägungen eines diskreten Merkmals als Ergebnis eines Zufallsvorgangs interpretiert, wird das Merkmal als diskrete Zufallsvariable angesprochen. Zählvariablen sind stets diskret.

Im Folgenden geht es um die Wahrscheinlichkeitsverteilung diskreter Zufallsvariablen – zunächst allgemein, bevor dann spezielle diskrete Verteilungsmodelle vorgestellt werden, die häufiger verwendet werden. Betrachtet sei eine diskrete Zufallsvariable X , die k Werte x_1, \dots, x_k annehmen kann. Letztere definieren die **Trägermenge** der Zufallsvariablen X . Das Verhalten von X ist vollständig definiert, wenn für jede Realisation x_i die Eintrittswahrscheinlichkeit $p_i = P(X = x_i)$ bekannt ist; $i = 1, \dots, k$. Die Funktion f , die jeder Ausprägung x_i eine Eintrittswahrscheinlichkeit p_i zuordnet, heißt **Wahrscheinlichkeitsfunktion** von X (engl: *probability*

Beschreibung
diskreter
Zufallsvariablen:

- anhand der
Wahrscheinlichkeits-
funktion

¹Zum Begriff „abzählbar unendlich“ vgl. erneut die Fußnote in Abschnitt 2.2.

density function, kurz *pdf*). Damit die Wahrscheinlichkeitsfunktion nicht nur auf der Trägermenge $\{x_1, \dots, x_k\}$, sondern für alle reellen Zahlen x erklärt ist, setzt man sie Null für alle x mit $x \neq x_i$:

$$f(x) = \begin{cases} p_i & \text{für } x = x_i; i = 1, 2, \dots, k \\ 0 & \text{für alle sonstigen } x. \end{cases} \quad (12.1)$$

Die Wahrscheinlichkeitsfunktion $f(x)$ lässt sich anhand eines Stab- oder Säulendiagramms mit k Stäben bzw. Säulen der Länge p_1, p_2, \dots, p_k darstellen. Sie kann nur nicht-negative Werte annehmen. Ferner muss die Summe der Eintrittswahrscheinlichkeiten p_1, p_2, \dots, p_k in (12.1) stets 1 sein. Hier besteht eine Analogie zu den in Kapitel 4 behandelten relativen Häufigkeitsverteilungen, denn auch relative Häufigkeiten sind nicht-negativ und summieren sich zu 1 auf.

- anhand der Verteilungsfunktion Zur Beschreibung einer diskreten Zufallsvariablen X kann man anstelle der Wahrscheinlichkeitsfunktion auch die schon in (11.17) eingeführte **Verteilungsfunktion**²

$$F(x) = P(X \leq x)$$

(engl.: *cumulative distribution function*, kurz *cdf*) von X heranziehen, die man zwecks Unterscheidung von der empirischen Verteilungsfunktion (4.5) präziser **theoretische Verteilungsfunktion** nennt. Die Verteilungsfunktion $F(x)$ einer durch (12.1) charakterisierten Zufallsvariablen X hat offenbar für $x < x_1$ den Wert Null und springt in $x = x_1$ auf den Wert $F(x_1) = p_1$. Der Funktionswert bleibt auf dem Niveau p_1 bis zur Stelle $x = x_2$, an der ein erneuter Sprung nach oben erfolgt, nun auf $F(x_2) = p_1 + p_2$, usw. Die Werte der Funktion $F(x)$ ergeben sich also dadurch, dass an den Stellen $x = x_i$ jeweils ein positiver Beitrag p_i hinzukommt, d.h. $F(x)$ ist eine monoton wachsende Treppenfunktion mit Sprungstellen in $x = x_i$. Bei der letzten Sprungstelle, also in $x = x_k$, erreicht $F(x)$ den Wert 1. Für $F(x)$ gilt demnach

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ p_1 & \text{für } x_1 \leq x < x_2 \\ \vdots & \vdots \\ p_1 + p_2 + \dots + p_{k-1} & \text{für } x_{k-1} \leq x < x_k \\ 1 & \text{für } x \geq x_k. \end{cases} \quad (12.2)$$

²Wenn man die Wahrscheinlichkeitsverteilungen zweier Zufallsvariablen unterscheiden will, kann man durch einen tiefgestellten Index deutlich machen, welche Verteilung gemeint ist. Für eine Variable X würde man z. B. präziser $f_X(x)$ und $F_X(x)$ anstelle von $f(x)$ und $F(x)$ schreiben.

Es gibt eine weitere Parallele zwischen den relativen Häufigkeitsverteilungen der beschreibenden Statistik und den Verteilungen diskreter Zufallsvariablen. Durch Aufsummieren relativer Häufigkeiten kommt man zur empirischen Verteilungsfunktion (4.5), die ebenfalls eine monoton wachsende Treppenfunktion ist, welche bis zum ersten Sprung den Wert 0 aufweist und an der letzten Sprungstelle den Wert 1 erreicht.

Besonders einfach ist der Fall einer diskreten Verteilung, bei der in (12.1) alle Ausprägungen x_i die gleiche Eintrittswahrscheinlichkeit $p = \frac{1}{k}$ besitzen, also $p_i \equiv p$ gilt (lies: p -i identisch p). Man spricht dann von einer **diskreten Gleichverteilung** oder genauer von einer diskreten Gleichverteilung mit Parameter p . Wahrscheinlichkeits- und Verteilungsfunktion einer diskreten Gleichverteilung mit k Ausprägungen x_1, x_2, \dots, x_k gehen aus (12.1) und (12.2) als Spezialfall hervor, wenn dort für alle Eintrittswahrscheinlichkeiten p_i der Wert $p = \frac{1}{k}$ eingesetzt wird. Aus (12.2) wird

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \frac{1}{k} & \text{für } x_1 \leq x < x_2 \\ \vdots & \vdots \\ \frac{k-1}{k} & \text{für } x_{k-1} \leq x < x_k \\ 1 & \text{für } x \geq x_k. \end{cases} \quad (12.3)$$

Die diskrete
Gleichverteilung

Die diskrete Gleichverteilung kommt z. B. ins Spiel, wenn man mehrfach würfelt und einen „fairen“ Würfel voraussetzt, also einen Würfel, bei dem alle Augenzahlen mit gleicher Wahrscheinlichkeit auftreten. Die Zufallsvariable „Augenzahl X “ hat hier sechs Ausprägungen $x_1 = 1, x_2 = 2, \dots, x_6 = 6$, die alle die Eintrittswahrscheinlichkeit $p = \frac{1}{6}$ aufweisen. Die Wahrscheinlichkeitsfunktion $f(x)$ der zugehörigen Gleichverteilung ist im linken Teil von Abbildung 12.1 wiedergegeben. Der rechte Teil der Abbildung zeigt die Verteilungsfunktion $F(x)$ des mit $p = \frac{1}{6}$ diskret gleichverteilten Merkmals X . Die Funktion weist für $x_1 = 1, x_2 = 2, \dots, x_6 = 6$ jeweils Sprünge der Höhe $\frac{1}{6}$ auf.

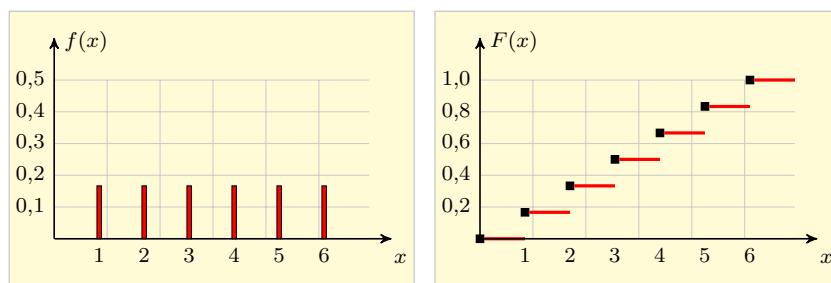
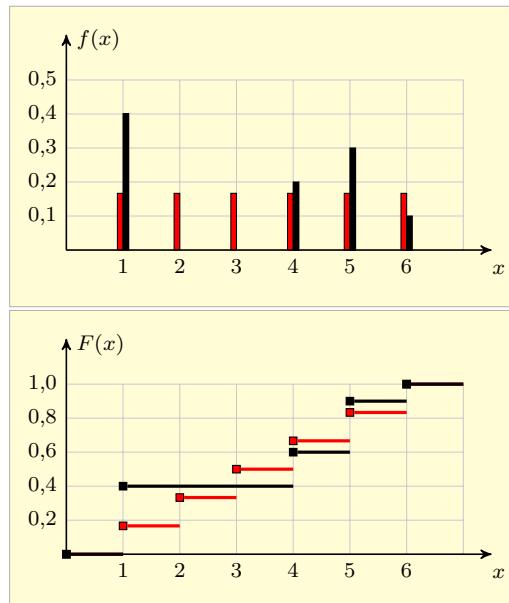


Abb. 12.1: Wahrscheinlichkeits- und Verteilungsfunktion der diskreten Gleichverteilung mit $p = \frac{1}{6}$ (Würfeln mit einem Würfel)

Beispiel 12.1: Daten- und Modellebene beim Würfelexperiment


Interaktives Objekt
„Augenzahlen“
(mit Modell)



i	f_i	F_i
1	0,4	0,4
2	0	0,4
3	0	0,4
4	0,2	0,6
5	0,3	0,9
6	0,1	1,0

Abb. 12.2: Relative Häufigkeiten für die Augenzahlen bei 10-fachem Würfeln und Modell (diskrete Gleichverteilung mit $p = \frac{1}{6}$)

Abbildung 12.2 zeigt im oberen Teil die per Simulation gewonnenen relativen Häufigkeiten in Form schwarzer Säulen für die sechs möglichen Ausprägungen bei nur 10-facher Durchführung des statistischen Experiments ($n = 10$). Im unteren Teil ist, ebenfalls in Schwarz, die hieraus resultierende empirische Verteilungsfunktion wiedergegeben. Zu Vergleichszwecken ist auch das schon in Abbildung 12.1 dargestellte Modell der diskreten Gleichverteilung mit dem Parameter $p = \frac{1}{6}$ eingezeichnet (rote Säulen). Neben der Abbildung sind in einer Tabelle die beobachteten relativen Häufigkeiten f_i für die einzelnen Augenzahlen und die mit F_i abgekürzten Werte der empirischen Verteilungsfunktion an den Stellen $x = x_i$ aufgeführt ($i = 1, 2, \dots, 6$). Die Tabelle zeigt, dass bei den 10 Würfen viermal die Augenzahl 1, zweimal die 4, dreimal die 5 und einmal die Augenzahl 6 erschien.

Abbildung 12.3 zeigt erneut die relativen Häufigkeiten und die daraus abgeleitete empirische Verteilungsfunktion, nun aber für den Fall $n = 100$. Auch hier ist zusätzlich das Modell der diskreten Gleichverteilung mit $p = \frac{1}{6}$ dargestellt. Ferner sind erneut die relativen Häufigkeiten f_i und die kumulierten Häufigkeiten F_i tabellarisch ausgewiesen. Man erkennt beim Vergleich von

Abbildung 12.3 mit Abbildung 12.2, dass das theoretische Verteilungsmodell die Simulationsergebnisse bei größerem n tendenziell besser beschreibt – die im Experiment beobachteten relativen Häufigkeiten f_i nähern sich den Werten $f(x_i) = \frac{1}{6}$ der Wahrscheinlichkeitsfunktion mit Vergrößerung von n an.

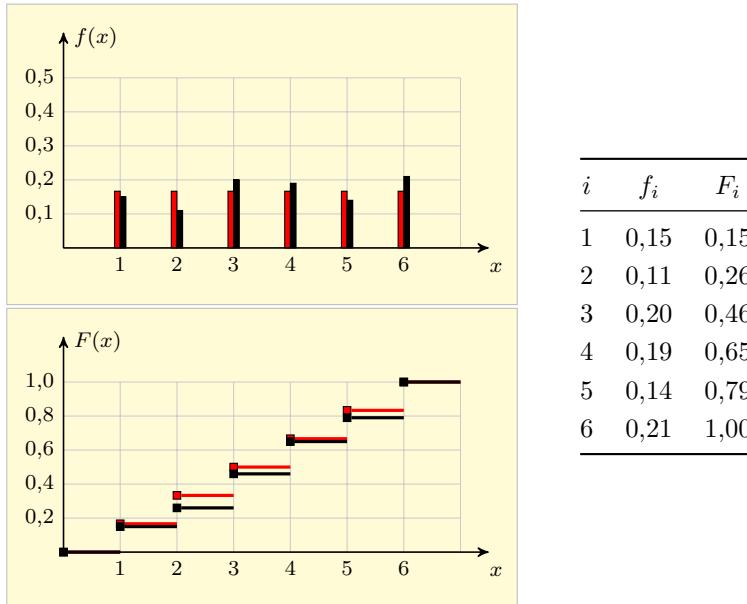


Abb. 12.3: Relative Häufigkeiten für die Augenzahlen bei 100-fachem Würfeln und Modell (diskrete Gleichverteilung mit $p = \frac{1}{6}$)

Der Wert $f_1 = 0,4$ in der Tabelle neben Abbildung 12.2 besagt z. B., dass in 40 % der Fälle, also bei 4 der $n = 10$ Würfe, die Augenzahl $x_1 = 1$ beobachtet wurde. Der entsprechende Wert $f_1 = 0,15$ in der Tabelle neben Abbildung 12.3, der sich auf $n = 100$ bezieht und hier auf die 15 % der Würfe mit Augenzahl $x_1 = 1$, liegt schon viel näher am theoretischen Wert $f(x_1) = \frac{1}{6} \approx 0,17$.

Exkurs 12.1: Wahlhäufigkeiten für vierstellige PIN-Codes

Hotelsafes kann man meist mit vierstelligen benutzerdefinierten Zahlenkombinationen schließen und öffnen. Auch von Tablets und Smartphones kennt man vierstellige PINs (*persönliche Identifikationsnummern*), die frei wählbar sind. Wenn man für die Eingabe einer der insgesamt 10 000 vierstelligen Zahlenkombinationen jeweils 4 Sekunden benötigte, könnte das sukzessive Durchprobieren bis zum Aufspüren der Geheimzahl mehr als 11 Stunden beanspruchen.

Wenn man eine vierstellige PIN wählt, sollte man wissen, dass es beliebte und selten gewählte Zahlenfolgen gibt und Hacker dies ausnutzen könnten. In einem Beitrag in der *Frankfurter Allgemeinen Sonntagszeitung* vom 3. August 2014 wertete der Facebook-Mitarbeiter Nick Berry die empirische Verteilung von etwa 3,4 Millionen bekannter frei wählbarer PIN-Codes aus. Einige

Zahlenkombinationen waren extrem häufig, andere wiederum auffällig selten vertreten. Kombinationen der Art $19xy$ (Geburtsjahre) oder besonders gut einprägsame Codes wie 1234 oder solche des Typs $xxxx$ sowie $xyxy$ waren die Spitzenreiter. Allein die relative Häufigkeit für die Kombination 1234 betrug etwa 0,107 (10,7%), die für 1111 immerhin ca. 0,06 (6,0%). Die Summe der relativen Häufigkeiten für die vier beliebtesten PINs - neben 1234 und 1111 waren dies 0000 und 1212 – lag bei 0,198 (19,8%).

Wenn die empirische Verteilung der Pins approximativ durch eine diskrete Gleichverteilung zu beschreiben wäre, müsste die relative Häufigkeit für jede der 10 000 möglichen Kombinationen näherungsweise bei 0,0001 (0,01%) liegen. Die empirische Verteilung der 3,4 Millionen benutzerdefinierten Codes unterscheidet sich deutlich von einer diskreten Gleichverteilung. Es überrascht daher nicht, dass Banken ihren Kunden die PINs von Bankkarten zuteilen. Diese werden dabei zufällig erzeugt (diskret gleichverteilte Zufallszahlen) und hängen damit nicht von Wahlpräferenzen der Kunden ab.



JACOB I.
BERNOULLI

Neben der diskreten Gleichverteilung ist noch ein weiterer einfacher Spezialfall einer diskreten Verteilung zu erwähnen, nämlich die nach dem Schweizer Mathematiker Jacob I. BERNOULLI (1655 – 1705) benannte **Bernoulli-Verteilung**, für die man auch die Bezeichnung **ZweipunktsVerteilung** findet. Diese Verteilung liegt vor, wenn eine Zufallsvariable X nur zwei Ausprägungen aufweist, etwa x_1 und x_2 oder A und \bar{A} . Die Variable X spricht man auch als **binäre Zufallsvariable** an.

Bezeichnet $p_1 = p$ die Eintrittswahrscheinlichkeit für den Fall $x = x_1$ und p_2 die für den Fall $x = x_2$, so ist offenbar $p_2 = 1 - p$. Die Wahrscheinlichkeitsfunktion (12.1) hat dann die spezielle Gestalt

$$f(x) = \begin{cases} p & \text{für } x = x_1; \\ 1 - p & \text{für } x = x_2; \\ 0 & \text{für alle sonstigen } x. \end{cases} \quad (12.4)$$

Charakterisierung der
Bernoulli-Verteilung

Durch (12.4) oder die Verteilungsfunktion

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{für } x < x_1; \\ p & \text{für } x_1 \leq x < x_2; \\ 1 & \text{für } x \geq x_2 \end{cases} \quad (12.5)$$

ist eine Bernoulli-Verteilung vollständig definiert. Ihre Gestalt hängt von p ab. Eine mit dem Parameter p bernoulli-verteilte Zufallsvariable X bezeichnet man als $Be(p)$ -verteilt und verwendet die Notation $X \sim Be(p)$ (lies: X ist *bernoulli-verteilt* mit dem Parameter p). Wenn man die Ausprägungen x_1 und x_2 zu 1 und 0 umcodiert, wird eine Bernoulli-Verteilung auch **Null-Eins-Verteilung** genannt.

Eine Grundgesamtheit, deren Elemente bezüglich eines Merkmals nur zwei Ausprägungen aufweisen, also bernoulli-verteilt sind, nennt man *dichotom*. Die Bevölkerung eines Landes ist z. B. bezüglich des Merkmals „Geschlecht“ eine dichotome Grundgesamtheit – das Merkmal „Geschlecht“ ist bernoulli-verteilt mit einem Verteilungsparameter p , der näherungsweise den Wert $p = 0,5$ hat. Auch die Ausgänge einer Serie von Münzwürfen konstituieren eine dichotome Grundgesamtheit. Jeder Wurf führt zu einer der Realisationen „Zahl“ und „Kopf“, d. h. auch hier ist eine Bernoulli-Verteilung mit $p = 0,5$ im Spiel, wenn man eine „faire“ Münze voraussetzt. Ein statistisches Experiment, dessen Ausgang durch ein bernoulli-verteiltes Merkmal beschrieben wird, heißt **Bernoulli-Experiment**.

Abbildung 11.2 zeigte für ein Münzwurfexperiment, bei dem eine „faire“ Münze 500-mal geworfen wurde, wie sich die relativen Häufigkeiten f_j für das Auftreten von „Zahl“ entwickelten ($j = 1, 2, \dots, 500$). Der in der Abbildung dargestellte Entwicklungspfad für die relativen Häufigkeiten f_j ist das Ergebnis einer Folge von Bernoulli-Experimenten, die jeweils voneinander unabhängig sind. Eine solche Folge wird auch **Bernoulli-Kette** genannt. In Abbildung 12.4 sind die relativen Häufigkeiten für zwei Münzwurfserien mit je 500 Würfen wiedergegeben. Es resultieren *zwei* Entwicklungspfade, also zwei Bernoulli-Ketten. Der Endstand $f_{500}(\text{Zahl})$ der beobachteten relativen Häufigkeit liegt bei beiden Wurfserien sehr dicht am Wert $p = 0,5$ der Eintrittswahrscheinlichkeit für „Zahl“, gegen den die Bernoulli-Ketten für $n \rightarrow \infty$ stochastisch konvergieren.³



Interaktives Objekt
„Münzwurf“

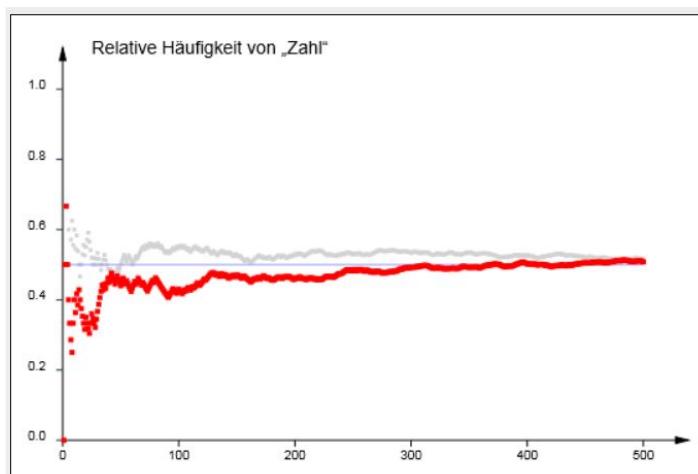


Abb. 12.4: Relative Häufigkeit für „Zahl“ bei zweifach durchgeföhrtem 500-fachen Münzwurf mit fairer Münze und Vergleich mit dem Parameter $p = 0,5$ der Bernoulli-Verteilung

³Zum Begriff der stochastischen Konvergenz vgl. erneut die Fußnote zu (11.6).

12.2 Kenngrößen diskreter Verteilungen

In Kapitel 5 wurden empirische Verteilungen durch wenige Kenngrößen charakterisiert. Zu nennen sind hier insbesondere die Lageparameter Mittelwert und Median, mit denen der Schwerpunkt einer Verteilung beschrieben wurde, sowie die Streuungsparameter Spannweite, Standardabweichung und Varianz, mit denen die Variabilität eines Datensatzes ausgedrückt werden kann. Auch theoretische Verteilungen werden durch Lage- und Streuungsmaße charakterisiert. Die Analogien zwischen empirischen und theoretischen Verteilungen sind bei den diskreten Zufallsvariablen besonders augenfällig.

Das arithmetische Mittel \bar{x} eines Datensatzes x_1, x_2, \dots, x_n , der sich auf ein diskretes Merkmal X mit k Ausprägungen a_1, a_2, \dots, a_k bezieht, lässt sich gemäß (5.4) als Summe der mit den relativen Häufigkeiten gewichteten Merkmalsausprägungen darstellen, also durch $a_1 f_1 + a_2 f_2 + \dots + a_k f_k$. In ähnlicher Weise lässt sich auch der Schwerpunkt der Verteilung der diskreten Zufallsvariablen (12.1) charakterisieren. Man bildet hier die Summe $x_1 p_1 + x_2 p_2 + \dots + x_k p_k$ der mit den Eintrittswahrscheinlichkeiten p_1, p_2, \dots, p_k gewichteten Realisationen. Diese Summe wird als **Erwartungswert** bezeichnet und mit $E(X)$ oder kürzer mit μ bezeichnet. Der Erwartungswert $E(X)$ (lies: *Erwartungswert von X*) einer nach (12.1) definierten diskreten Zufallsvariablen ist also gegeben durch



Erwartungswert
und Varianz
einer diskreten
Zufallsvariablen

$$\mu := E(X) = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum_{i=1}^k x_i \cdot p_i. \quad (12.6)$$

Die Merkmalsausprägungen a_1, a_2, \dots, a_k und die relativen Häufigkeiten f_1, f_2, \dots, f_k aus Kapitel 5 werden also hier, bei der Charakterisierung theoretischer Verteilungsmodelle, durch die Realisationen x_1, x_2, \dots, x_k einer diskreten Zufallsvariablen und deren Eintrittswahrscheinlichkeiten p_1, p_2, \dots, p_k ersetzt. Die gleichen Ersetzungen kann man auch in den Formeln (5.8) und (5.6) für die empirische Standardabweichung bzw. die empirische Varianz vornehmen. Man erhält so für die mit $V(X)$ oder σ^2 (lies: *Varianz von X* resp. *sigma-Quadrat*) abgekürzte **Varianz** der diskreten Zufallsvariablen (12.1) mit $\mu = E(X)$ die Darstellung

$$\sigma^2 := V(X) = \sum_{i=1}^k (x_i - \mu)^2 \cdot p_i. \quad (12.7)$$

Die Darstellung (12.6) geht in (12.7) über, wenn man in (12.6) anstelle von X den Term $(X - \mu)^2$ einsetzt. Es gilt also

$$\sigma^2 = E [(X - \mu)^2]. \quad (12.8)$$

Für die Varianz ist manchmal die Darstellung

$$\sigma^2 = E(X^2) - \mu^2 \quad (12.9)$$

nützlich, die sich aus (12.8) ergibt, wenn man dort den Term in der eckigen Klammer ausmultipliziert und dann den Erwartungswert gliedweise bestimmt.⁴ Der Varianzdarstellung (12.9) entspricht auf der empirischen Ebene die Zerlegungsformel (5.7).

Die **Standardabweichung** σ (lies: *sigma*) von X ist definiert durch

$$\sigma := \sqrt{V(X)}. \quad (12.10)$$

Zwischen den Kenngrößen empirischer und theoretischer Verteilungen wird in der Lehrbuchliteratur oft nicht klar unterschieden. Der Mittelwert bezieht sich auf eine empirische, der Erwartungswert immer auf eine theoretische Verteilung. Wenn von der Varianz die Rede ist, kann man durch die Verwendung der präziseren Bezeichnungen „empirische Varianz“ bzw. „theoretische Varianz“ deutlich machen, ob die Varianz eines Datensatzes (empirische Ebene) oder die einer Zufallsvariablen (Modellebene) gemeint ist. Eine analoge Aussage gilt für die Standardabweichung.

In der Praxis unterzieht man eine Zufallsvariable X mit Erwartungswert oft einer Lineartransformation $Y = aX + b$. Die Addition von b entspricht einer Verschiebung des Nullpunkts, während die Multiplikation von X mit einem von Null verschiedenen Wert a eine Streckung oder Stauchung der zur Messung verwendeten Skala beinhaltet (im Fall $a < 0$ kommt noch ein Vorzeichenwechsel hinzu). Lineartransformationen sind z. B. relevant, wenn man eine andere Skala bei der Messung verwendet (etwa Temperaturmessung in Kelvin statt in Celsius) oder wenn man X in eine Zufallsvariable Y mit Erwartungswert $E(Y) = 0$ und Varianz $V(Y) = 1$ überführen will (**Standardisierung**).

Lineartransformationen bei Zufallsvariablen

Unterzieht man eine Zufallsvariable X mit Erwartungswert $\mu = E(X)$ einer Lineartransformation $Y = aX + b$, so gilt

$$E(Y) = E(aX + b) = a \cdot E(X) + b \quad (12.11)$$

$$V(Y) = V(aX + b) = a^2 \cdot V(X). \quad (12.12)$$

Für den Erwartungswert zweier Zufallsvariablen X und Y gilt allgemein

$$E(X + Y) = E(X) + E(Y), \quad (12.13)$$

während die entsprechende Formel

$$V(X + Y) = V(X) + V(Y) \quad (12.14)$$

⁴Hierbei werden die noch folgenden Formeln (12.11) und (12.13) herangezogen.

nur gilt, wenn man Unabhängigkeit oder zumindest fehlende lineare Abhängigkeit von X und Y voraussetzt.⁵ Die Gleichungen (12.13) und (12.14) gelten entsprechen auch für die Summen von n unabhängigen Zufallsvariablen ($n \geq 2$).

- Kenngrößen der Null-Eins-Verteilung Erwartungswert und Varianz der Null-Eins-Verteilung ergeben sich unmittelbar aus den allgemeineren Formeln (12.6) und (12.7) für den Erwartungswert bzw. die Varianz diskreter Zufallsvariablen, wenn man dort $k = 2$ sowie $x_1 = 1$, $p_1 = p$, $x_2 = 0$ und $p_2 = 1 - p$ einsetzt und bei der Varianzberechnung auf (12.9) zurückgreift:

$$\mu = 1 \cdot p + 0 \cdot (1 - p) = p \quad (12.15)$$

$$\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1 - p). \quad (12.16)$$

Beispiel 12.2: Kenngrößen des Merkmals „Augenzahl“

In Abbildung 12.1 wurde die Wahrscheinlichkeitsverteilung der Zufallsvariablen „Augenzahl X beim Würfeln“ (Gleichverteilung mit Parameter $p = \frac{1}{6}$) anhand ihrer Wahrscheinlichkeitsfunktion $f(x)$ und ihrer Verteilungsfunktion $F(x)$ veranschaulicht. Da die Ausprägungen $x_i = i$ die Eintrittswahrscheinlichkeiten $p_i = p = \frac{1}{6}$ besitzen ($i = 1, 2, \dots, 6$), erhält man für den Erwartungswert $\mu = E(X)$ und die Varianz $\sigma^2 = V(X)$ aus (12.6) und (12.7)

$$\begin{aligned} \mu &= \sum_{i=1}^6 x_i \cdot p_i = \frac{1}{6} \cdot \sum_{i=1}^6 i = \frac{21}{6} = 3,5 \\ \sigma^2 &= \sum_{i=1}^6 (x_i - \mu)^2 \cdot p_i = \frac{1}{6} \cdot \sum_{i=1}^6 (i - 3,5)^2 = \frac{17,5}{6} \approx 2,92. \end{aligned}$$

Quantile als weitere Kenngrößen

Wie bei empirischen Verteilungen kann man auch bei theoretischen Verteilungen **Quantile** zur Charakterisierung heranziehen. Das **p-Quantil** einer Verteilung ist durch

$$F(x_p) = p \quad (0 < p < 1) \quad (12.17)$$

definiert, also durch den Wert x_p der Verteilungsfunktion $F(x)$, an dem $F(x)$ den Wert p annimmt. Der **Median** $\tilde{x} = x_{0,5}$ sowie das **untere Quartil** $x_{0,25}$ und das **obere Quartil** $x_{0,75}$ einer theoretischen Verteilung sind wieder spezielle Quantile, die sich bei Wahl von $p = 0,5$ resp. von $p = 0,25$ und $p = 0,75$ ergeben.

⁵Die Begriffe Unabhängigkeit sowie Unkorreliertheit von Zufallsvariablen werden noch in Kapitel 14 definiert.

Bei diskreten Verteilungen sind die Quantile durch (12.17) noch nicht eindeutig festgelegt. Bei der im rechten Teil von Abbildung 12.1 wiedergegebenen Verteilungsfunktion einer speziellen diskreten Gleichverteilung gilt z. B. $F(x) = 0,5$ für jeden Wert x aus dem Intervall $3 \leq x < 4$. Man benötigt daher hier wie bei den empirischen Quantilen noch eine Zusatzbedingung. Man kann z. B. den linken Randpunkt des Intervalls wählen, d. h. das p-Quantil x_p so festlegen, dass $F(x_p) \geq p$ gilt und gleichzeitig $F(x) < p$ für $x < x_p$. Für die diskrete Gleichverteilung in Abbildung 12.1 erhält man so für den Median $\tilde{x} = x_{0,5}$ den Wert $\tilde{x} = 3$.

12.3 Die Binomialverteilung

Es fällt nicht schwer, in verschiedenen Lebensbereichen Beispiele für Merkmale X zu finden, die nur zwei mögliche Ausprägungen haben, also den Charakter von Binärvariablen haben. Das Ergebnis eines Münzwurfs ist ein Beispiel – es kann nur „Zahl“ und „Kopf“ auftreten. Die n Ausgänge einer Serie von n Münzwürfen konstituieren eine dichotome Grundgesamtheit. Praxisrelevantere Beispiele, die sich auf dichotome Grundgesamtheiten beziehen, sind etwa die Geschlechterverteilung bei Geburten,⁶ die Verteilung eines Gendefekts in einer Population (nicht betroffene / betroffene Individuen), der beim Mikrozensus erfragte Erwerbsstatus einer Person (erwerbstätig / nicht erwerbstätig) oder der Qualitätsstatus von Produkten bei Serienfertigungen (spezifikationskonform / nicht-spezifikationskonform). Aber auch Merkmale mit mehr als zwei Ausprägungen können stets auf Binärvariablen zurückgeführt werden, wenn man sich nur dafür interessiert, ob eine bestimmte Realisation eintritt. Das Würfeln mit einem Würfel lässt sich z. B. als Bernoulli-Experiment interpretieren, wenn man sich darauf beschränkt, nur zwischen den Ereignissen „Augenzahl ist 6 / nicht 6“ oder „Augenzahl ist größer als 2 / nicht größer als 2“ zu unterscheiden.

Hat man ein Bernoulli-Experiment mit den möglichen Ausgängen $x_1 = A$ und $x_2 = \bar{A}$ und den zugehörigen Eintrittswahrscheinlichkeiten $P(A) = p$ bzw. $P(\bar{A}) = 1 - p$ mehrfach und unabhängig voneinander durchgeführt, so interessiert man sich oft dafür, wie häufig eine der beiden Realisationen auftritt, etwa A . Beim Münzwurfexperiment könnte dies z. B. die Anzahl der Ausgänge mit „Zahl“ sein. Ist n die Anzahl der unabhängig durchgeföhrten Bernoulli-Experimente und bezeichnet X die Anzahl der Ausgänge A , so ist die Zählvariable X eine diskrete Zufallsvariable mit den Ausprägungen $0, 1, \dots, n$. Wenn man den Ausgang jedes der n



Video „Binomialverteilung 1“

⁶Seit Dezember 2018 kann in Deutschland noch „divers“ als dritte Geschlechtskategorie amtlich registriert werden.

Bernoulli-Experimente anhand einer Indikatorvariablen

$$X_i = \begin{cases} 1 & \text{bei Eintritt von } x_1 = A \\ 0 & \text{bei Eintritt von } x_2 = \bar{A} \end{cases} \quad (12.18)$$

beschreibt, so lässt sich X als Summe

$$X = \sum_{i=1}^n X_i \quad (12.19)$$

der n voneinander unabhängigen null-eins-verteilten Zufallsvariablen schreiben. Die Verteilung der Zählvariablen X heißt **Binomialverteilung**. Diese ist für die statistische Praxis von großer Bedeutung. Die Null-Eins-Verteilung ist ein Spezialfall der Binomialverteilung ($n = 1$).

Kenngrößen der
Binomialverteilung

Aus (12.19) kann man leicht den Erwartungswert $E(X)$ und die Varianz $V(X)$ der binomialverteilten Variablen X ableiten. Die in (12.19) eingeschlossenen n Indikatorvariablen X_i sind voneinander unabhängig und folgen alle einer Null-Eins-Verteilung, besitzen demnach wegen (12.15) und (12.16) den Erwartungswert $E(X_i) = p$ und die Varianz $V(X_i) = p(1-p)$. Mit den Formeln (12.13) und (12.14) folgt hieraus für die Kenngrößen $\mu = E(X)$ und $\sigma^2 = V(X)$ einer Binomialverteilung

$$\mu = n \cdot p \quad (12.20)$$

$$\sigma^2 = n \cdot p \cdot (1 - p). \quad (12.21)$$

Charakterisierung der
Binomialverteilung

Da eine diskrete Zufallvariable noch nicht durch Erwartungswert und Varianz alleine, sondern erst durch die Wahrscheinlichkeitsfunktion (12.1) oder – alternativ – durch die Verteilungsfunktion (12.2) vollständig beschrieben ist, sei noch die Wahrscheinlichkeitsfunktion der Binomialverteilung abgeleitet. Hierzu werde zunächst die Wahrscheinlichkeit dafür betrachtet, dass bei dem Bernoulli-Experiment am Anfang genau x -mal der Ausgang A und danach $(n-x)$ -mal der Ausgang \bar{A} beobachtet wird, die Bernoulli-Kette also die spezielle Gestalt $A, A, \dots, A, \bar{A}, \dots, \bar{A}$ hat mit zwei homogenen Teilketten der Längen x bzw. $n-x$. Die Wahrscheinlichkeit für den Eintritt dieser speziellen Ergebnisfolge, die für die Zählvariable X zum Wert x führt, ist wegen der Unabhängigkeit der einzelnen Bernoulli-Experimente $p^x(1-p)^{n-x}$. Nun gibt es aber nicht nur eine Ergebnisfolge, sondern nach Tabelle 11.1 insgesamt $\binom{n}{x}$ mögliche Ausprägungen einer Bernoulli-Kette der Länge n , bei der insgesamt x -mal der Ausgang A auftritt. Die Reihenfolge des Auftretens der Ausgänge A innerhalb einer Ergebnisfolge hat keinen Effekt auf den Wert der Zählvariablen X . Die Wahrscheinlichkeit $P(X = x)$ dafür, dass die Anzahl der Ausgänge A innerhalb der Bernoulli-Kette einen bestimmten Wert x annimmt, ist damit gegeben durch das $\binom{n}{x}$ -fache von



Video

„Galtonbrett und
Binomialverteilung“

$p^x(1-p)^{n-x}$. Für die **Wahrscheinlichkeitsfunktion** $f(x) = P(X = x)$ der Binomialverteilung gilt also

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für } x = 0, 1, \dots, n \\ 0 & \text{für alle sonstigen } x. \end{cases} \quad (12.22)$$

Die **Verteilungsfunktion** $F(x) = P(X \leq x)$ ist auf der Trägermenge $\{0, 1, \dots, n\}$ definiert durch

$$F(x) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k} \quad x = 0, 1, \dots, n. \quad (12.23)$$

Zwischen zwei benachbarten Elementen der Trägermenge bleibt $F(x)$ auf dem Niveau des kleineren Elements, um dann an der Stelle $x = n$ den Endwert 1 zu erreichen. Eine mit Parametern n und p binomialverteilte Zufallsvariable X bezeichnet man auch als $B(n; p)$ -verteilt und schreibt dafür $X \sim B(n; p)$ (liest: X ist *binomialverteilt* mit den Parametern n und p). Die Aussagen $X \sim B(1; p)$ und $X \sim Be(p)$ sind identisch, weil die Bernoulli-Verteilung eine Binomialverteilung mit $n = 1$ ist.

Abbildung 12.5 zeigt Wahrscheinlichkeits- und Verteilungsfunktion einer $B(10; 0,5)$ -verteilten Zufallsvariablen. Der Tabelle neben der Grafik entnimmt man, dass die Verteilungsfunktion an der Stelle $x = 3$ den Wert $F(3) = 0,1719$ annimmt. Dieser Wert ist wegen $F(3) = P(X \leq 3)$ die Summe der Werte $f(0), f(1), f(2)$ und $f(3)$ der Wahrscheinlichkeitsfunktion (12.22). Durch Aufsummieren von Werten der Wahrscheinlichkeitsfunktion ergeben sich die Werte der Verteilungsfunktion.



Video „Binomialverteilung 2“



Interaktives Objekt
„Binomialverteilung“

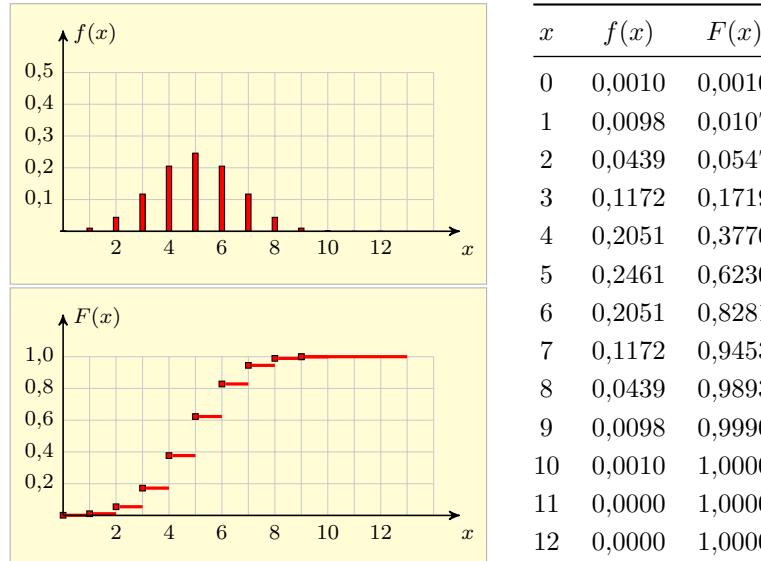


Abb. 12.5: Binomialverteilung mit $n = 10$ und $p = 0,50$

Umgekehrt kann man aus $F(x)$ durch Differenzenbildung Werte der Wahrscheinlichkeitsfunktion $f(x)$ gewinnen. Der neben Abbildung 12.5 tabellierte Wert $f(3) = P(X = 3) = 0,1172$ ergibt sich etwa als Differenz von $F(3) = P(X \leq 3) = 0,1719$ und $F(2) = P(X \leq 2) = 0,0547$. Es genügt demnach, eine der Funktionen $f(x)$ und $F(x)$ zu tabellieren.

Die Wahrscheinlichkeitsfunktion (12.22) ist für $p = 0,5$ symmetrisch bezüglich des Erwartungswerts. Für $p \neq 0,5$ gilt dies nicht mehr, wie Abbildung 12.6 beispielhaft illustriert. Die Wahrscheinlichkeitsfunktion ist hier links vom Erwartungswert $\mu = 2,5$ steiler.

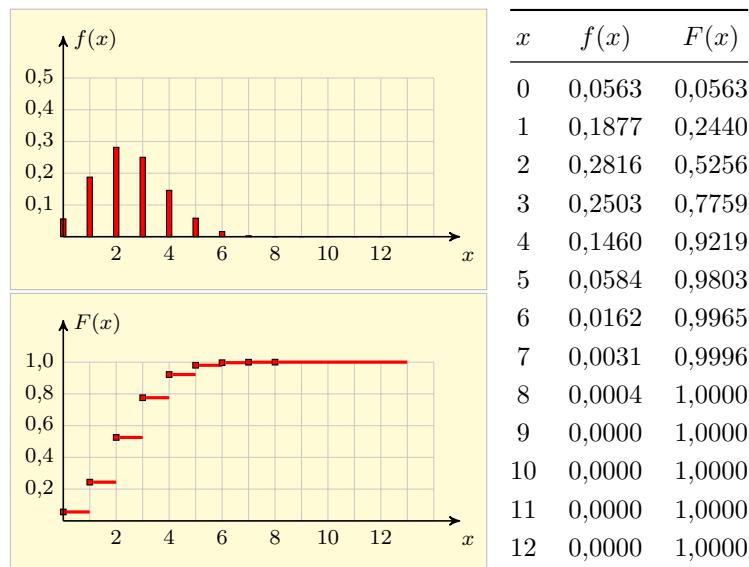


Abb. 12.6: Binomialverteilung mit $n = 10$ und $p = 0,25$

In Tabelle 19.1 des Anhangs sind Verteilungsfunktionen von Binomialverteilungen für ausgewählte Werte von n und p tabelliert. Werte der Verteilungs- und auch der Wahrscheinlichkeitsfunktion für andere Kombinationen $(n; p)$ lassen sich mit jedem Statistiksoftwarepaket, z. B. SPSS oder JMP, sowie mit EXCEL oder der kostenfreien Statistiksoftware *R* berechnen.⁷ Bezuglich der Verwendung von *R* sei auf die Einführungen von WOLLSCHLÄGER (2020) und LIGGES (2020) verwiesen.

⁷Mit *R* lassen sich Werte der Wahrscheinlichkeitsfunktion und der Verteilungsfunktion der Binomialverteilung und auch anderer diskreter Verteilungen leicht bestimmen – der Wert der Verteilungsfunktion der Binomialverteilung mit $n = 10$ und $p = 0,025$ an der Stelle $x = 3$ z. B. mit `pnbinom(3, size = 10, prob = 0.025)`.

Beispiel 12.3: Anwendung der Binomialverteilung

Wenn man eine Münze n -mal wirft, so ist die Anzahl X der Ereignisse „Zahl“ eine $B(n; p)$ -verteilte Zufallsvariable. Der Erwartungswert ist hier durch $\mu = np = \frac{n}{2}$ und die Varianz durch $V(X) = np(1-p) = \frac{n}{4}$ gegeben. Bei Verwendung einer „fairen“ Münze, also einer Münze mit gleichen Eintrittswahrscheinlichkeiten für „Zahl (Z)“ und „Kopf (K)“, gilt $p = 0,5$. Die Wahrscheinlichkeit $P(X \leq 2)$ dafür, bei 3 Würfen höchstens 2-mal den Ausgang „Zahl“ zu erhalten, ist dann durch den Wert $F(2)$ der Verteilungsfunktion der Binomialverteilung mit $n = 3$ und $p = 0,5$ gegeben, nach Tabelle 19.1 also durch $F(2) = 0,875$. Die Wahrscheinlichkeit $f(2) = P(X = 2)$ dafür, bei den drei Würfen genau zweimal „Zahl“ zu erzielen, errechnet sich als Differenz der Funktionswerte $F(2) = P(X \leq 2) = 0,875$ und $F(1) = P(X \leq 1) = 0,500$, also als 0,375. Der letztgenannte Wert wurde auch schon in Beispiel 11.2 elementar unter Verwendung des Laplace-Ansatzes (11.5) über die Kombinatorik abgeleitet.

Bei größeren Werten n werden aber kombinatorische Überlegungen aufwändig, insbesondere, wenn ein Wert $p \neq 0,5$ ins Spiel kommt. zieht man etwa aus einer Lostrommel, in der ein Anteil $p = 0,05$ Gewinne und der Rest Nieten sind, nacheinander n Lose und legt nach jeder Einzelziehung das Los in die Trommel zurück, so ist die Wahrscheinlichkeit nach 20 Ziehungen genau 4 Gewinne gezogen zu haben, errechenbar als Differenz $F(4) - F(3)$ zweier Werte der Verteilungsfunktion einer $B(20; 0,05)$ -verteilten Zufallsvariablen. Man erhält mit Tabelle 19.1 den Wert $0,9974 - 0,9841 = 0,0133$. Für die Wahrscheinlichkeit dafür, im Falle $n = 20$ und $p = 0,05$ mindestens 4 Gewinne zu ziehen, ermittelt man den Wert $1 - F(3) = 0,0159$.



Interaktives Objekt
„Rechnen mit der
Binomialverteilung“



Aufgabe 12.2

Exkurs 12.2: Fiasko beim Zentralabitur 2008 in NRW

In Nordrhein-Westfalen gab es beim Zentralabitur 2008 heftige Kritik an einer Aufgabe zur Wahrscheinlichkeitsrechnung für Mathematik-Leistungskurse. Der erste Teil der umstrittenen Aufgabe lautete wie folgt:

Der deutsche Basketball-Profi Dirk Nowitzki spielte in der amerikanischen Profiliiga beim Club Dallas Mavericks. In der Saison 2006/07 erzielte er bei Freiwürfen eine Trefferquote von 90,4 Prozent. Berechnen Sie die Wahrscheinlichkeit dafür, dass er

- (1) genau 8 Treffer bei 10 Versuchen erzielt,
- (2) höchstens 8 Treffer bei 10 Versuchen erzielt,
- (3) höchstens viermal nacheinander bei Freiwürfen erfolgreich ist.



Basketballer Nowitzki
(Quelle: dpa)

Zu bemängeln ist zunächst, dass nicht wirklich klar ist, ob sich die erwähnten 10 Würfe auf die Saison 2006/07 beziehen. Man hätte explizit betonen müssen, dass dies so gemeint ist und bei der Lösung der Aufgabenteile (1) und (2) eine konstante Trefferquote von $p = 0,904$ vorausgesetzt werden soll.

Geht man von dieser Annahme aus, ist die Wahrscheinlichkeit für die Erzielung von genau 8 Treffern bzw. von höchstens 8 Treffern bei 10 Wurfversuchen durch den Wert $f(8)$ der Wahrscheinlichkeitsfunktion $f(x)$ resp. den Wert $F(8)$ der Verteilungsfunktion $F(x)$ einer $B(10; 0,904)$ -verteilten Zufallsvariablen gegeben. Man errechnet dann z. B. für $f(8)$ nach (12.22) den Wert

$$f(8) = \binom{10}{8} \cdot 0,904^8 \cdot 0,096^2 = 45 \cdot 0,4460129 \cdot 0,009216 \approx 0,185.$$

Die Wahrscheinlichkeit für die Erzielung von vier Treffern in Folge lässt sich allerdings auch bei Annahme einer festen Trefferquote noch nicht beantworten, weil in Aufgabenteil (3) die Gesamtzahl n der Würfe nicht angegeben ist, von der das Ergebnis abhängt. Aufgabenteil (3) ist also eigentlich nicht lösbar. Unterstellt man, dass hier $n = 10$ gemeint war und codiert man „Treffer“ mit „1“ sowie das Komplementärereignis „kein Treffer“ mit „0“, hätte man aus den insgesamt $2^{10} = 1024$ möglichen Ergebnisfolgen diejenigen herauszusuchen, bei denen nie mehr als vier Einsen in Folge erscheinen. Mit $(1,0,1,1,1,1,0,0,1,1)$ hat man ein Beispiel für eine Ergebnisfolge, die dem Erfordernis „höchstens vier Treffer in Folge“ genügt.

12.4 Die hypergeometrische Verteilung

Die Binomialverteilung beschreibt das Zufallsverhalten der Zählvariablen X aus (12.19) bei einem n -fach durchgeführten Bernoulli-Experiment, wobei die einzelnen Experimente voneinander unabhängig sind. Die Zählvariable weist aus, wie häufig einer der beiden möglichen Ausgänge $x_1 = A$ und $x_2 = \bar{A}$ und $P(A) = p$ bzw. $P(\bar{A}) = 1 - p$ innerhalb der Bernoulli-Kette auftrat. Als Beispiele wurden Münzwurf- oder auch Würfeexperimente angeführt, wenn man bei letzteren nur zwischen zwei Ausgängen differenziert (etwa „gerade / ungerade Augenzahl“).

Varianten des Urnenmodells	<p>Die Grundsituation lässt sich anhand des Urnenmodells beschreiben. Eine Urne (Behälter) enthalte eine Menge roter und schwarzer Kugeln. Der Urne wird n-mal eine Kugel entnommen und man zählt die Anzahl X der roten Kugeln. Nach jeder Ziehung wird die entnommene Kugel in die Urne zurückgelegt. Der Quotient „Anzahl roter Kugeln / Anzahl aller Kugeln“, der die Wahrscheinlichkeit für die Entnahme einer roten Kugel bestimmt, bleibt hier von Ziehung zu Ziehung konstant. Die Binomialverteilung lässt sich also anschaulich durch das Urnenmodell mit Zurücklegen veranschaulichen. Dieses Modell ist z. B. beim wiederholten Münzwurf passend, weil die Ausgangslage sich nicht von Wurf zu Wurf verändert. Es ist so, als ob man einer Urne, die zwei Zettel mit der Aufschrift „Zahl“ bzw. „Kopf“ enthält, jeweils einen Zettel entnimmt und den gezogenen Zettel vor der nächsten Ziehung zurücklegt.</p>
----------------------------	--

In der Realität gibt es Situationen, bei denen das beschriebene Modell des Ziehens mit Zurücklegen nicht oder nur näherungsweise passt – man denke nur an die Ziehung der Lottozahlen oder an Befragungen von Personen auf der Basis zufälliger Stichproben. Auch in der Wareneingangsprüfung bei einem Unternehmen wird man bei Entnahme einer Stichprobe von n Elementen aus einem Warenlos ein entdecktes nicht-spezifikationskonformes Element vor der Entnahme eines weiteren Elements nicht zurücklegen. In solchen Fällen wird das **Urnenmodell ohne Zurücklegen** verwendet. Wenn man einer Urne mit N Kugeln, von denen M rot und die restlichen $N - M$ schwarz sind, nacheinander n Kugeln ohne Zurücklegen entnimmt, so repräsentiert die Ziehung jeder Kugel zwar weiterhin ein Bernoulli-Experiment, die Einzelexperimente sind aber nicht mehr voneinander unabhängig. Die Eintrittswahrscheinlichkeit für das interessierende Ereignis „Kugel ist rot“ wird jetzt nicht nur von M , sondern auch vom Umfang N der Grundgesamtheit beeinflusst. Die Verteilung der durch (12.19) definierten Zählvariablen X ist bei Annahme einer Stichprobentnahme ohne Zurücklegen nicht mehr durch eine Binomialverteilung gegeben, sondern durch die **hypergeometrische Verteilung**. Letztere ist durch drei Parameter beschrieben, nämlich durch N , M und n , und man schreibt hierfür $X \sim H(n; M; N)$ (lies: X ist hypergeometrisch verteilt mit den Parametern n , M und N). Erwartungswert $\mu = E(X)$ und Varianz $\sigma^2 = V(X)$ der hypergeometrischen Verteilung seien nur hier der Vollständigkeit halber und ohne Beweis angegeben.⁸

$$\mu = n \cdot \frac{M}{N} \quad (12.24)$$

$$\sigma^2 = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}. \quad (12.25)$$



Video „Hypergeometrische Verteilung“

Kenngrößen der hypergeometrischen Verteilung

Erwartungswert und Varianz einer $H(n, M, N)$ -verteilten Zufallsvariablen X stimmen nach (12.20) und (12.21) mit dem Erwartungswert bzw. der Varianz einer $B(n; p)$ -verteilten Variablen mit $p = \frac{M}{N}$ überein – mit dem einzigen Unterschied, dass bei der Varianz der Binomialverteilung der in (12.25) auftretende Bruchterm $\frac{N-n}{N-1}$ fehlt. Da dieser Term für $n > 1$ kleiner als 1 ist, hat die hypergeometrische Verteilung im Vergleich zur Binomialverteilung eine kleinere Varianz, wobei die Unterschiede mit wachsendem N vernachlässigbar werden. Dass die hypergeometrische Verteilung eine kleinere Varianz aufweist, ist einleuchtend, denn beim Ziehen ohne Zurücklegen wird die in einem gezogenen Stichprobenelement steckende Information (Kugel ist rot oder schwarz) nicht immer wieder verschenkt, d. h. es gibt weniger Unsicherheit über den verbleibenden Inhalt der Urne im Vergleich zum Ziehen mit Zurücklegen. Im Extremfall

⁸Eine Herleitung von Erwartungswert, Varianz und auch der Wahrscheinlichkeitsfunktion $f(x)$ der hypergeometrischen Verteilung findet man z. B. bei MOSLER / SCHMID (2011, Abschnitt 2.3.4).

der sukzessiven Ziehung aller in der Urne befindlichen Elemente ($n = N$) ohne Zurücklegen liegt vollständige Information über den Urneninhalt vor. Die Zählvariable X ist dann keine Zufallsvariable mehr, sondern eine deterministische Größe mit dem Wert M . Man erkennt den nicht-stochastischen Charakter von X im Falle $n = N$ auch anhand von (12.25), denn es gilt dann $\frac{N-n}{N-1} = 0$ und somit $V(X) = 0$.

Beispiel 12.4: Prüfpläne in der Abnahmeprüfung

Zur Überprüfung der Einhaltung von Qualitätsanforderungen an Lieferposten (Eingangsprüfung für Zulieferteile oder Endabnahme beim Warenausgang) wird aus dem betreffenden Warenlos mit N Elementen, die unter einheitlichen Bedingungen produziert wurden, eine Stichprobe entnommen. Wenn bei der Qualitätsprüfung bei jedem Element des Loses nur zwischen den Ausprägungen „spezifikationskonform / gebrauchstauglich“ und „nicht-spezifikationskonform / defekt“ unterschieden wird und die entdeckten nicht-spezifikationskonformen Stichprobenelemente gezählt und aussortiert werden, liegt die beim Urnenmodell *ohne* Zurücklegen beschriebene Grundsituation vor. Für die Anzahl X der nicht-spezifikationskonformen Stichprobenelemente gilt demnach $X \sim H(n; M; N)$, wenn n den Umfang der entnommenen Stichprobe bezeichnet und M die unbekannte Anzahl der zu beanstandenden Elemente im gesamten Los.

Falls die in der Stichprobe ermittelte Anzahl einen in einem **Prüfplan** festgelegten Schwellenwert c nicht überschreitet, wird das Warenlos angenommen. Wird c hingegen überschritten, bestimmt der Prüfplan, ob das Los zurückzuweisen ist oder ob – bei nur geringfügiger Überschreitung von c – eine weitere Stichprobe zu ziehen ist und welchen Umfang diese haben sollte.

Die gängigen Prüfpläne für die zählende Prüfung sind in der *ISO Norm 2859* beschrieben. Eine ausführliche Charakterisierung von Prüfplänen für die Abnahmeprüfung findet man bei RINNE / MITTAG (1995) und STORM (2007).



Charakterisierung der hypergeometrischen Verteilung

Die Angabe der **Trägermenge** einer $H(n; M; N)$ -verteilten Zufallsvariablen, also der Menge der möglichen Ausprägungen der Zählvariablen X , ist nicht trivial. Sie ist durch $T = \{x_{\min}, \dots, x_{\max}\}$ gegeben mit $x_{\min} = \max(0; n - N + M)$ als dem kleinsten und $x_{\max} = \min(n; M)$ als dem größten Element der Trägermenge (s. hierzu den Exkurs 12.3). Die **Wahrscheinlichkeitsfunktion** $f(x) = P(X = x)$ der hypergeometrischen Verteilung ist ebenfalls nicht so einfach ableitbar wie die der Binomialverteilung. Es gilt die Darstellung

$$f(x) = \begin{cases} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} & \text{für } x \in T \\ 0 & \text{für alle sonstigen } x, \end{cases} \quad (12.26)$$

deren Herleitung in Exkurs 12.3 noch skizziert wird. Für die **Verteilungsfunktion** $F(x) = P(X \leq x)$ gilt dann

$$F(x) = \sum_{k=0}^x \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \quad \text{für alle } x \in T. \quad (12.27)$$

Die Funktion bleibt zwischen zwei benachbarten Elementen der Trägermenge auf dem Niveau des kleineren Werts, um dann im Punkt $x_{\max} = \min(n; M)$ den Endwert 1 anzunehmen (Treppenfunktion).



Interaktives Objekt
„Hypergeometrische
Verteilung“

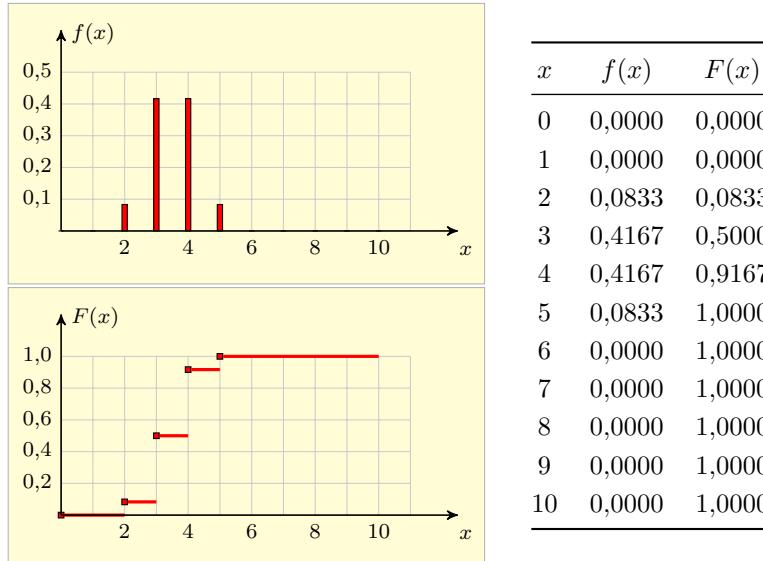


Abb. 12.7: Hypergeometrische Verteilung mit $n = 5$, $M = 7$ und $N = 10$

Abbildung 12.7 zeigt die Wahrscheinlichkeits- und die Verteilungsfunktion einer $H(5; 7; 10)$ -verteilten Zufallsvariablen. Der Erwartungswert errechnet sich hier als $\mu = 5 \cdot \frac{7}{10} = 3,5$. Die Trägermenge T der dargestellten hypergeometrischen Verteilung ist durch $T = \{x_{\min}, \dots, x_{\max}\}$ gegeben mit $x_{\min} = \max(0; 5 - 10 + 7) = 2$ und $x_{\max} = \min(5; 7) = 5$. Die Tabellierung von Werten der Funktionen (12.26) oder (12.27) ist unüblich, weil die Werte von drei Parametern abhängen. Sie lassen sich aber mit Hilfe der freien Programmiersprache R leicht bestimmen, z. B. $f(3) = 0,4167$ und $F(3) = 0,5$ (s. Tabelle neben Abbildung 12.7). Beim R-Code entsprechen die Werte in der Klammer $x, M, N - M$ und n :

```
> dhyper(3, 7, 10 - 7, 5) > phyper(3, 7, 10 - 7, 5)
[1] 0.4166667 [1] 0.5
```

Abb. 12.8: R-Code für die Bestimmung von Werten einer $H(5; 7; 10)$ -verteilten Zufallsvariablen



Aufgabe 12.3

Exkurs 12.3: Charakterisierung der hypergeometrischen Verteilung

Um die *Trägermenge* einer $H(n; M; N)$ -verteilten Zufallsvariablen zu bestimmen, sind nur die kleinst- und die größtmögliche Ausprägung der durch (12.19) erklärten Zählvariablen X im Urnenmodell ohne Zurücklegen zu ermitteln (Urne mit M roten und $N - M$ schwarzen Kugeln). Die Variable X , die sich hier als Anzahl der gezogenen roten Kugeln nach n Ziehungen interpretieren lässt, kann im Falle $n \leq M$ den Wert n offenbar nicht überschreiten. Im Falle $n > M$ ist hingegen M die Obergrenze – es können nicht mehr als M rote Kugeln gezogen werden, weil nur M rote Kugeln in der Urne vorhanden sind. Das größte Element x_{\max} der Trägermenge hat also den Wert $x_{\max} = \min(n; M)$. Ferner gilt, dass die Anzahl $n - (N - M)$ der roten Kugeln nach n Ziehungen nicht kleiner als 0 sein kann, d. h. $x_{\min} = \max(0; n - N + M)$ definiert den kleinstmöglichen Wert.

Bei der Herleitung der *Wahrscheinlichkeitsfunktion* (12.26) kann man auf Tabelle (11.1) zurückgreifen. Der Nenner von (12.26) repräsentiert die Anzahl der Möglichkeiten, aus einer Urne mit N Kugeln insgesamt n Kugeln ohne Zurücklegen zu entnehmen. Nach Tabelle (11.1) ist diese Anzahl durch $\binom{N}{n}$ gegeben, weil es auf die Reihenfolge der Ergebnisse der Ziehungen hier nicht ankommt. Der Produktterm im Zähler von (12.26) ergibt sich aus folgender Überlegung: In der Urne befinden sich vor Beginn der Ziehung M rote und $N - M$ schwarze Kugeln. Es gibt $\binom{M}{x}$ Möglichkeiten, x rote Kugeln aus M roten Kugeln auszuwählen. Damit nach n Ziehungen ohne Zurücklegen die Anzahl der gezogenen roten Kugeln genau x ist, müssen aus dem Anfangsvorrat von $N - M$ schwarzen Kugeln $n - x$ schwarze Kugeln gezogen werden. Es gibt $\binom{N-M}{n-x}$ Möglichkeiten der Auswahl dieser $n - x$ Kugeln.

Approximation der
hypergeometrischen
Verteilung

In der Praxis wendet man anstelle der hypergeometrischen Verteilung meist die einfacher handhabbare Binomialverteilung an, wenn der Umfang N der Grundgesamtheit im Vergleich zum Umfang der Stichprobe n groß ist (Faustregel: $\frac{n}{N} < 0,05$). In diesem Falle kann man für eine $H(n; M; N)$ -verteilte Zufallsvariable X in guter Näherung annehmen, dass sie $B(n; p)$ -verteilt ist mit $p = \frac{M}{N}$. Die Tragfähigkeit der Approximation liegt darin begründet, dass die Unterschiede zwischen den Situationen „Ziehen ohne / mit Zurücklegen“ mit Verkleinerung des Auswahlatzes $\frac{n}{N}$ immer weniger ins Gewicht fallen.

Die Binomialverteilung und die hypergeometrische Verteilung charakterisieren beide das Zufallsverhalten der Zählvariablen (12.19), allerdings unter verschiedenen Bedingungen. Die Variable (12.19) zählt, wie oft bei n -facher Durchführung eines Bernoulli-Experiments (n -faches Ziehen einer Kugel aus einer Urne mit roten und schwarzen Kugeln) mit den möglichen Ausgängen $x_1 = A$ (Kugel ist rot) und $x_2 = \bar{A}$ eines der beiden Ereignisse, etwa A , beobachtet wird. Beim Ziehen *mit Zurücklegen* ist

die Zählvariable binomialverteilt, beim Ziehen *ohne Zurücklegen* folgt sie einer hypergeometrischen Verteilung. Beide Verteilungen gehen im Fall $n = 1$ in die Bernoulli-Verteilung über. Beim Ziehen einer einzigen Kugel aus einer Urne mit roten und schwarzen Kugeln und der Wahrscheinlichkeit p für das Ereignis A entfällt nämlich eine Unterscheidung von Ziehen mit oder ohne Zurücklegen und die Wahrscheinlichkeitsfunktion (12.4) beschreibt den Ausgang des einmaligen Bernoulli-Experiments.

Bernoulli-Verteilung
als Spezialfall

Beispiel 12.5: Wahrscheinlichkeiten beim Lottospiel

Lotto wird in Europa nicht einheitlich gespielt. In Deutschland gibt es z. B. das Lottospiel „6 aus 49“, in der Schweiz „6 aus 45“ und in Italien „6 aus 90“. Die Wahrscheinlichkeiten für die Ereignisse „6 Richtige“, „0 Richtige“, „mindestens 4 Richtige“ o. ä. beim deutschen Lotto lassen sich anhand der hypergeometrischen Verteilung mit den Parametern $n = 6$, $M = 6$ und $N = 49$ berechnen. Dabei beinhaltet n hier die Anzahl der Kreuze auf dem Lottoschein (beim Urnenmodell die Anzahl der gezogenen Kugeln), M die maximale Anzahl der Treffer (beim Urnenmodell die Anzahl der „roten“ Kugeln in der Urne) und N die Anzahl der die Lottozahlen präsentierenden Kugeln in der Trommel (bzw. in der Urne). Der Erwartungswert für die Anzahl X der Richtigen beim Lottospiel „6 aus 49“ ist nach (12.24) durch $\mu = \frac{36}{49} \approx 0,735$ gegeben.

Für die Berechnung von Wahrscheinlichkeiten der Art „ x Richtige“ oder „mindestens x Richtige“ kann man eine Tabelle mit Werten der Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ oder der Verteilungsfunktion $F(x) = P(X \leq x)$ der hypergeometrischen Verteilung verwenden. Wenn man nicht über eine solche Tabelle verfügt, kann man die gesuchten Wahrscheinlichkeiten entweder mit R ermitteln oder direkt aus (12.26) bzw. (12.27) bestimmen.



Abb. 12.9: „Lottofee“ (ARD-Lottoziehung; Quelle: Hessischer Rundfunk)

Für das Ereignis „0 Richtige“ erhält man z. B. nach (12.26) mit Einsetzen von $n = 6$, $M = 6$ und $N = 49$ bei Beachtung von $\binom{6}{0} = 1$ die Darstellung

$$f(0) = \frac{\binom{6}{0} \binom{49-6}{6-0}}{\binom{49}{6}} = \frac{\binom{43}{6}}{\binom{49}{6}}.$$

Der Nennerterm, für den man mit (10.10) den Wert

$$\binom{49}{6} = \frac{49!}{43! \cdot 6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13\,983\,816$$

ermittelt (vgl. auch Beispiel 11.3), repräsentiert die Anzahl der möglichen Ausgänge einer Lottoziehung. Für den Zählerterm, der die Anzahl der möglichen Ausgänge mit 0 Richtigen wiedergibt, folgt

$$\binom{43}{6} = \frac{43!}{37! \cdot 6!} = \frac{43 \cdot 42 \cdot 41 \cdot 40 \cdot 39 \cdot 38}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 6\,096\,454.$$

Die Wahrscheinlichkeit $f(0) = P(X = 0)$ für das Ereignis „0 Richtige“ ist somit

$$f(0) = \frac{\binom{43}{6}}{\binom{49}{6}} = \frac{6\,096\,454}{13\,983\,816} \approx 0,436,$$

also ca. 43,6% – ein Wert, der überraschen dürfte. Die Wahrscheinlichkeit $f(6) = P(X = 6)$ für „6 Richtige“ ließe sich analog bestimmen. Da allerdings von den 13 983 816 möglichen Ausgängen einer Lottoziehung nur ein einziger Ausgang „6 Richtige“ beinhaltet, kann man $f(6) = P(X = 6)$ einfacher über

$$f(6) = \frac{1}{13\,983\,816} \approx 0,0000000715 = 7,15 \cdot 10^{-8}$$

errechnen. Von den Lottoeinnahmen werden 50% als Steuern abgeführt oder für festgelegte Zwecke verwendet und 8% an Lottospieler verteilt, die 6 Richtige haben (Gewinnklasse 2).



Aufgabe 12.4-5

Die verschwindend kleine Wahrscheinlichkeit für einen Volltreffer verringert sich beim deutschen Lotto noch um den Faktor 10 auf $\frac{1}{139\,838\,160} \approx 7,15 \cdot 10^{-9}$, wenn man das Spiel „6 aus 49 mit Superzahl“ spielt. Die „Superzahl“ ist eine Zusatzzahl, die aus der Menge $\{0, 1, \dots, 8, 9\}$ gezogen wird. Um an den legendären Jackpot zu kommen (Gewinnklasse 1 mit einer Ausschüttungsquote von 10%), muss man „6 Richtige aus 49“ haben *und* die korrekte Zusatzzahl vorweisen können. Diese Gewinnklasse ist häufig gar nicht besetzt – die vorgesehenen Gewinne werden dann auf die nächste Ziehung übertragen.

Noch geringer als ein Erreichen der Gewinnklasse 1 beim deutschen Lotto ist die Wahrscheinlichkeit eines Volltreffers beim italienischen Lotto „6 aus 90“. Sie entspricht dem Wert $f(6) = \frac{1}{622\,614\,630} \approx 1,61 \cdot 10^{-9}$ der Wahrscheinlichkeitsfunktion einer hypergeometrischen Verteilung mit den Parametern $n = 6$, $M = 6$ und $N = 90$.



**Kritisch
nachgefragt**

Wenn einem Internet-Betrüger hunderte oder gar tausende Menschen zum Opfer fallen, können die ermittelnden Behörden unmöglich jeden einzelnen davon als Zeugen befragen. Deshalb beschränkt man sich in Strafverfahren auf eine „repräsentative Auswahl“ von Geschädigten. Diese ist nach Ansicht mancher Richter bereits durch die Erfüllung recht eigenwilliger Kriterien gegeben:

„[die] vernommenen Zeugen [stellen] einen durchaus repräsentativen Querschnitt der Anleger [dar] (. . .), da die Spannbreite von der Hausfrau über einen Polizisten bis zum Universitätsprofessor für Volkswirtschaft reicht.“⁹

Als vollendetes Betrug gilt, wenn jemand getäuscht wurde und lediglich aufgrund eines Irrtums gehandelt hat. Wer z. B. nach einer harsch formulierten E-Mail 99 Euro auf das Konto einer angeblichen Anwaltskanzlei überweist, weil er glaubt, diesen Betrag schuldig zu sein (Motivlage A), ist betrogen worden. Wer aber nur zahlt, „um seine Ruhe zu haben“ (Motivlage B)¹⁰ kann nur einen Betrugsversuch geltend machen.

In einem Fall mit 53 494 Geschädigten wurden lediglich 15 Zeugen befragt. Der Bundesgerichtshof hielt das für zulässig, wollte aber aufgrund der Unschuldsvermutung „in dubio pro reo“ nur für die 15 Vernommenen, die alle angaben, allein aufgrund eines Irrtums gezahlt zu haben, von einem vollendeten Betrug ausgehen.¹¹ Um zu wissen, wie viele Geschädigte sich tatsächlich geirrt haben, müsste man die „individuelle Motivation“ einzeln untersuchen. Bei der Strafverfolgung in Massenbetrugsfällen sehen Richter offenbar bisher keinen gangbaren Weg, zu einem Urteil auf der Basis einer repräsentativen Zufallsstichprobe zu gelangen. Dabei böte sich ein statistisches Modell (hypergeometrische Verteilung) dafür an, abzuschätzen, wie viele Betroffene mit Motivlage A in der Grundgesamtheit aller N Betroffenen *mindestens* vorkommen.

Der Grundgedanke sei anhand des Urnenmodells zunächst unter Verwendung fiktiver Zahlen illustriert. Nehmen wir an, in einer Urne befinden sich $N = 100$ Kugeln, von denn M rot und $N - M$ schwarz sind mit unbekanntem M . zieht man $n = 10$ Kugeln ohne Zurücklegen, so ist die Anzahl X der roten Kugeln eine Zufallsvariable, die hypergeometrisch verteilt ist mit den Parametern $n = 10$, M und $N = 100$: $X \sim H(10; M; 100)$. Enthält die Stichprobe z. B. 9 rote Kugeln, so ist gesichert, dass für die Anzahl X der anfangs in der Urne vorhandenen roten Kugeln $M \geq 9$ gilt. Dabei ist $P(X \geq 9) = 1 - F(8)$, wobei $F(8)$ den Wert der Verteilungsfunktion (12.27) der genannten hypergeometrischen Verteilung an der Stelle $x = 8$ bezeichnet. Da der zur Berechnung von $P(X \geq 9)$ benötigte Parameter M unbekannt ist, setzen wir für M nacheinander alle theoretisch denkbaren Werte von 9 bis 100 ein und erhalten so eine Folge von Wahrscheinlichkeitswerten, die in Abbildung 12.10 als Funktion von M grafisch dargestellt ist. In der Abbildung ist der größte Wert für M gekennzeichnet, bei dem noch die Abschätzung $P(X \geq 9) < \alpha$ gilt, wobei hier $\alpha = 0,05$ gewählt wurde. Dieser in der Grafik mit M_u abgekürzte größte Wert von M beträgt

⁹BGH 5 StR 510/13 – NStZ 2014, 318

¹⁰Vgl. z.B. BGH 1 StR 314/14, NStZ 2015, 98 (mit abl. Anm. Krehl)

¹¹BGH 1 StR 263/12 – NJW 2013, 1545

hier $M_u = 61$. In der Urne befinden sich also mit einer Wahrscheinlichkeit von $1 - \alpha = 0,95$ mehr als $M_u = 61$ Kugeln. Anders formuliert: Die Untergrenze M_u wird nur mit einer Irrtumswahrscheinlichkeit von $\alpha = 0,05$ unterschritten.

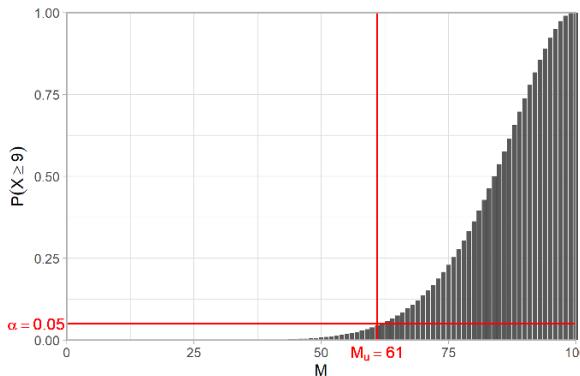


Abb. 12.10: Wahrscheinlichkeit $P(X \geq 9)$ der hypergeometrisch verteilten Zufallsvariablen X als Funktion des Verteilungsparameters M

Die vorstehenden Überlegungen lassen sich leicht auf das skizzierte Massenbetrugsverfahren übertragen. Hier war $n = 15$ die Anzahl der vernommenen Zeugen, N die Anzahl der Geschädigten und M die unbekannte Anzahl der Geschädigten mit Motivlage A. Tabelle 12.1 weist aus, wieviele Geschädigte mit Motivlage A bei der Schadensberechnung zu berücksichtigen wären, wenn man anstelle des Schadens nur der $n = 15$ vernommenen Zeugen (sichere Fälle) den Schaden für die weitaus höhere Mindestanzahl M_u der Geschädigten zugrunde legte, die mit Wahrscheinlichkeit $1 - \alpha = 0,95$ ebenfalls nur aufgrund eines Irrtums gezahlt hatten. Der Wert M_u ergibt sich aus als größter Wert M , für den $P(X \geq 15) < \alpha$ gilt, wobei $X \sim H(15; M; 53\,494)$. Für α wurden alternativ die Irrtumswahrscheinlichkeiten 0,1 %, 0,01 % und 0,001 % durchgerechnet.

$1 - \alpha$	n	M_u	Einzelschaden (Euro)	Gesamtschaden (Euro)
9,999	15	33 754	44,60	1 505 428,40
0,9999	15	28 951	44,60	1 291 214,60
0,99999	15	24 832	44,60	1 107 507,20

Tab. 12.1: Anzahl vollendeter Betrugsfälle mit Schadensangabe (modellbasierte Hochrechnung zu BGH 1 StR 263/12)

Die obige Beispielrechnung verdeutlicht, wie groß der Unterschied ist zwischen dem, was der BGH als „vollendete Betrugsfälle“ gelten ließ, und der Situation, die sich bei Anwendung eines statistischen Modells ergäbe.



13 Stetige Zufallsvariablen



Vorschau auf
das Kapitel

Auch Daten für stetige Merkmale können als Realisierungen von Zufallsvariablen aufgefasst werden. Diese lassen sich wieder durch Wahrscheinlichkeitsverteilungen beschreiben. Während die Verteilung einer diskreten Zufallsvariablen durch Wahrscheinlichkeits- und Verteilungsfunktion zu charakterisieren ist, wird bei einer *stetigen* Zufallsvariablen neben der Verteilungsfunktion die Dichtefunktion herangezogen.

Die einfachste stetige Verteilung ist die der stetigen Gleichverteilung. Sie findet bei der Modellierung von Wartezeiten Anwendung. Weitaus häufiger verwendet wird die Normalverteilung. Die Gestalt der Dichte hängt vom Erwartungswert μ und von der Standardabweichung σ bzw. der Varianz σ^2 ab. Jede Normalverteilung lässt sich in die spezielle Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ überführen (Standardnormalverteilung).

Aus der Normalverteilung werden noch drei weitere Verteilungen abgeleitet, nämlich die χ^2 -Verteilung, die t -Verteilung und die F -Verteilung. Diese Verteilungen – genauer: Quantile dieser Verteilungen – werden im Zusammenhang mit dem Testen von Hypothesen benötigt.

13.1 Dichtefunktion und Verteilungsfunktion

Die in Kapitel 12 behandelten *diskreten* Zufallsvariablen sind dadurch gekennzeichnet, dass man die Anzahl ihrer Ausprägungen abzählen kann. Sie haben also endlich viele oder zumindest abzählbar unendlich viele Ausprägungen. Diese definieren die **Trägermenge** der Variablen. Das Zufallsverhalten einer diskreten Zufallsvariablen X mit k Ausprägungen x_i ($i = 1, \dots, k$) und Eintrittswahrscheinlichkeiten $p_i = P(X = x_i)$ lässt sich vollständig durch die in (12.1) eingeführte Wahrscheinlichkeitsfunktion $f(x)$ beschreiben. Alternativ kann man auch die Verteilungsfunktion $F(x)$ aus (12.2) zur Beschreibung heranziehen, die sich durch Aufsummieren aller Werte der Wahrscheinlichkeitsfunktion bis zur Stelle x ergibt.

Bei den im Folgenden thematisierten *stetigen* Zufallsvariablen ist die Trägermenge T , also die Menge der möglichen Realisationen, ein *Intervall*. Häufig ist T die Menge \mathbb{R} aller reellen Zahlen. Das Verhalten einer stetigen Zufallsvariablen X lässt sich wie im diskreten Fall durch die **Verteilungsfunktion** (engl.: *cumulative density function*, kurz *cdf*)

Charakterisierung
stetiger Zufalls-
variablen anhand
von Dichte- und
Verteilungsfunktion

$$F(x) = P(X \leq x)$$

aus (11.17) vollständig charakterisieren. Der Ansatz (12.1), der die Eintrittswahrscheinlichkeiten bei einer diskreten Zufallsvariablen mit endlich vielen Ausprägungen zusammenfasst und hier die Wahrscheinlichkeitsfunktion definiert, ist bei einer stetigen Zufallsvariablen nicht mehr anwendbar. Man verwendet nun anstelle der Wahrscheinlichkeitsfunktion die sog. **Dichtefunktion**. Diese Funktion $f(x)$, die auch als **Wahrscheinlichkeitsdichte** oder kürzer als **Dichte** von X angesprochen wird (engl.: *probability density function*, kurz *pdf*), genügt der Nicht-Negativitätsbedingung

$$f(x) \geq 0 \quad \text{für alle reellen } x \quad (13.1)$$

und hat die Eigenschaft, dass sich jeder Wert $F(x)$ der Verteilungsfunktion durch Integration der Dichte bis zur Stelle x ergibt. Es gilt also

$$F(x) = \int_{-\infty}^x f(t)dt \quad \text{für alle reellen } x. \quad (13.2)$$

Für alle Werte x , bei denen die Dichtefunktion $f(x)$ stetig ist, stimmt sie mit der Ableitung $F'(x)$ der Verteilungsfunktion überein:

$$F'(x) = f(x). \quad (13.3)$$

Aus (13.2) folgt, dass sich bei einer stetigen Zufallsvariablen X die Wahrscheinlichkeit $P(X \leq x)$ nicht nur als Wert der Verteilungsfunktion $F(x)$ an der Stelle x , sondern auch als Fläche unter der Dichtekurve $f(x)$ bis zum Punkt x interpretieren lässt (vgl. die noch folgende Abbildung 13.4). Setzt man $x = b$ bzw. $x = a$ in (13.2) ein, erhält man Darstellungen der Werte $F(b)$ und $F(a)$ der Verteilungsfunktion und hieraus für die Differenz $F(b) - F(a)$ die Gleichung

$$F(b) - F(a) = \int_{-\infty}^b f(t)dt - \int_{-\infty}^a f(t)dt = \int_a^b f(t)dt. \quad (13.4)$$

Da die Verteilungsfunktion monoton wächst und gegen 1 strebt, besitzt die Gesamtfläche unter der Dichtekurve den Wert 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1. \quad (13.5)$$

Eine besonders einfache stetige Verteilung ist die **Rechteckverteilung**, die auch **stetige Gleichverteilung** genannt wird. Man nennt eine stetige Zufallsvariable *rechteckverteilt* oder *gleichverteilt* über dem Intervall $[a, b]$,

wenn sie die Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq x \leq b \\ 0 & \text{für alle sonstigen } x \end{cases} \quad (13.6)$$

besitzt. Die Verteilungsfunktion $F(x)$ einer über $[a, b]$ rechteckverteilten Zufallsvariablen X ergibt sich gemäß (13.2) durch Integration dieser Dichte. Die Integration liefert nur im Bereich von a bis b einen von Null verschiedenen Beitrag, d. h. es ist

$$F(x) = \begin{cases} 0 & \text{für } x < a; \\ \frac{x-a}{b-a} & \text{für } a \leq x \leq b; \\ 1 & \text{für } x > b. \end{cases} \quad (13.7)$$

Charakterisierung
der stetigen
Gleichverteilung

Die Funktion (13.6) besitzt alle Eigenschaften, die eine Dichtefunktion auszeichnen. Sie ist zum einen nicht-negativ und erfüllt außerdem die Normierungseigenschaft (13.5). Letzteres ist sofort einsichtig, wenn man sich vergegenwärtigt, dass man die Integration in (13.5) auf das Intervall $[a, b]$ beschränken kann, weil $f(x)$ außerhalb dieses Bereichs Null ist. Integriert man $f(x)$ über $[a, b]$, entspricht das Ergebnis dem Flächeninhalt $A = 1$ eines Rechtecks mit Länge $b - a$ und Höhe $\frac{1}{b-a}$.

Abbildung 13.1 zeigt die Dichtefunktion (13.6) und die Verteilungsfunktion (13.7) einer Rechteckverteilung über $[a, b]$, wobei hier beispielhaft $a = 2$ und $b = 6$ gewählt wurde. Beide Funktionen sind über die Beziehung (13.3) verknüpft, wenn man von den beiden Sprungstellen $x = a$ und $x = b$ der Dichtefunktion absieht, in denen $F(x)$ nicht differenzierbar ist. Die Dichte hat zwischen $x = 2$ und $x = 6$ den konstanten Wert $f(x) = \frac{1}{4}$ und die unter diesem Bereich liegende Fläche den Inhalt 1.

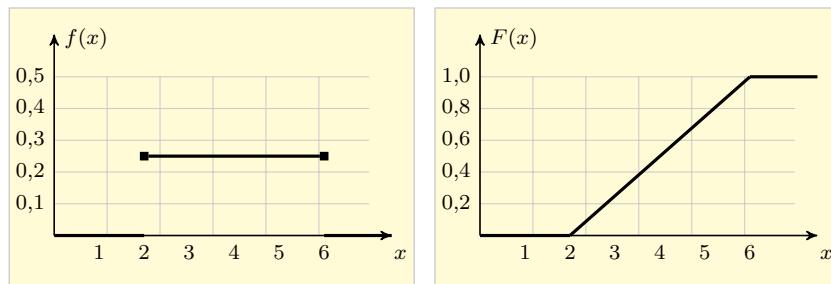


Abb. 13.1: Dichte- und Verteilungsfunktion der Rechteckverteilung über $[2,6]$

Der Wert $f(x_0)$ der Dichtefunktion einer stetigen Zufallsvariablen X an der Stelle $x = x_0$ ist *nicht* als Wahrscheinlichkeit $P(X = x_0)$ dafür zu interpretieren, dass X die Ausprägung x_0 annimmt. Man kann vielmehr

zeigen (vgl. auch den folgenden Exkurs 13.1), dass bei einer stetigen Zufallsvariablen X die Wahrscheinlichkeit $P(X = x_0)$ für jeden einzelnen Wert x_0 der Trägermenge Null ist:

$$P(X = x_0) = 0 \text{ für jeden Wert } x = x_0. \quad (13.8)$$

Die Dichtefunktion wird also nicht zur Berechnung von Wahrscheinlichkeiten für isolierte Werte herangezogen, sondern zur Berechnung von Wahrscheinlichkeiten von Ereignissen der Art „Die Realisationen von X liegen unterhalb oder oberhalb eines bestimmten Schwellenwerts“ oder „ X nimmt Realisationen x in einem Intervall $[a, b]$ an“. Im ersten Fall geht es darum, Werte $F(x)$ der Verteilungsfunktion $F(x) = P(x \leq X)$ resp. die zu 1 komplementären Werte $P(X > x) = 1 - F(x)$ zu ermitteln. Im zweiten Fall sind Differenzen $F(b) - F(a)$ von Werten der Verteilungsfunktion $F(x)$ zu bestimmen.

Beispiel 13.1: Modellierung von Wartezeiten

Die Rechteckverteilung findet u. a. Anwendung als Wartezeitverteilung. Geht man z. B. in einem Außenbezirk einer Großstadt ohne Kenntnis des Fahrplans in eine U-Bahnstation, von der alle 10 Minuten eine Bahn in Richtung Zentrum abfährt, so kann die Wartezeit X anhand einer Rechteckverteilung über $[0, 10]$ modelliert werden. Die Dichtefunktion (13.6) hat also die spezielle Gestalt

$$f(x) = \begin{cases} \frac{1}{10} & \text{für } 0 \leq x \leq 10; \\ 0 & \text{für alle sonstigen } x \end{cases}$$

und für die Verteilungsfunktion (13.7) hat man hier

$$F(x) = \begin{cases} 0 & \text{für } x < 0; \\ \frac{x}{10} & \text{für } 0 \leq x \leq 10; \\ 1 & \text{für } x > 10. \end{cases}$$

Die Wahrscheinlichkeit dafür, höchstens x Minuten zu warten ($0 \leq x \leq 10$), ist also gegeben durch $P(X \leq x) = \frac{x}{10}$.

Exkurs 13.1: Interpretation von Werten der Dichtefunktion

Anhand der Rechteckverteilung über $[a, b]$ lässt sich beispielhaft und auf indirekte Weise verdeutlichen, dass die Wahrscheinlichkeit $P(X = x_0)$ für jede Realisation x_0 einer stetigen Zufallsvariablen X Null sein muss, die Wahrscheinlichkeit $P(X = x_0)$ für das Eintreten einer bestimmten Ausprägung x_0 also nicht mit dem Wert $f(x_0)$ der Dichtefunktion verwechselt werden darf.

Bei der genannten Rechteckverteilung ist jede Realisation innerhalb des Intervalls $[a, b]$ gleichwahrscheinlich. Es sei innerhalb des Intervalls ein Wert $x = x_0$

herausgegriffen. Nimmt man nun an, dass die Wahrscheinlichkeit $P(X = x_0)$ einen von Null verschiedenen Wert hat, etwa $\frac{1}{p}$, also $P(X = x_0) = \frac{1}{p} > 0$, dann müsste diese Wahrscheinlichkeit auch für jeden weiteren Wert x in $[a, b]$ gelten. Für $p + 1$ beliebige Einzelwerte aus dem Intervall wäre dann die Summe der Wahrscheinlichkeiten $1 + \frac{1}{p}$. Dies wäre dann ein Widerspruch zu (13.5).

13.2 Kenngrößen stetiger Verteilungen

Auch bei stetigen Verteilungen ist man daran interessiert, diese durch wenige Kenngrößen zu charakterisieren. Als Lageparameter verwendet man wieder den mit μ (lies: *mü*) abgekürzten **Erwartungswert** $E(X)$ (lies: *Erwartungswert von X*). Für *diskrete* Zufallsvariablen mit endlich vielen Ausprägungen ist der Erwartungswert durch die Summe (12.6) definiert. Bei *stetigen* Zufallsvariablen sind die Ausprägungen nicht mehr abzählbar. Anstelle von (12.6) ist der Erwartungswert hier durch

$$\mu := E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad (13.9)$$

Erwartungswert und Varianz einer stetigen Zufallsvariablen

gegeben. Eine analoge, ebenfalls durch Grenzwertbetrachtungen ableitbare Formel gilt für die **Varianz** $\sigma^2 = V(X)$ (lies: *sigma-Quadrat* bzw. *Varianz von X*). Die bei einer *diskreten* Zufallsvariablen mit endlich vielen Ausprägungen gültige Summendarstellung (12.7) ist bei einer stetigen Verteilung zu ersetzen durch

$$\sigma^2 := V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx. \quad (13.10)$$

Die Varianz ist wie im diskreten Fall – vgl. (12.8) und (12.9) – nichts anderes als der Erwartungswert der quadrierten Differenz zwischen X und $\mu = E(X)$, also

$$\sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2,$$

und auch die **Standardabweichung** σ (lies: *sigma*) ist wieder durch

$$\sigma = \sqrt{V(X)}$$

erklärt. Unverändert gültig sind auch die Eigenschaften (12.11) - (12.14), die das Verhalten von Erwartungswert und Varianz bei einfachen Lineartransformationen charakterisieren. Eine besonders wichtige Lineartransformation ist die als **Standardisierung** bezeichnete Transformation einer Zufallsvariablen X in eine neue Variable $aX + b$ mit $a = \frac{1}{\sigma}$ und

$b = -\frac{\mu}{\sigma}$, die üblicherweise mit Z abgekürzt wird:

$$Z = \frac{X - \mu}{\sigma}. \quad (13.11)$$

Der Übergang von X zu Z heißt auch **z-Transformation**. Durch Einsetzen von $a = \frac{1}{\sigma}$ und $b = -\frac{\mu}{\sigma}$ in (12.11) und (12.12) verifiziert man, dass für den Erwartungswert der standardisierten Variablen $E(Z) = 0$ und für die Varianz $V(Z) = 1$ gilt.

Kenngrößen der Rechteckverteilung Für den Erwartungswert der durch (13.6) oder (13.7) definierten stetigen Gleichverteilung über $[a, b]$ sollte sich die Mitte $\frac{a+b}{2}$ des Intervalls $[a, b]$ ergeben, die das Zentrum der Verteilung markiert. Man errechnet diesen Wert in der Tat aus (13.9). Es ist nämlich

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{2 \cdot (b-a)} \cdot (b^2 - a^2).$$

Mit $(b^2 - a^2) = (b+a) \cdot (b-a)$ resultiert



$$\mu = E(X) = \frac{a+b}{2}. \quad (13.12)$$

Für die Varianz der Rechteckverteilung erhält man zunächst

Aufgabe 13.1

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_a^b \frac{x^2}{b-a} dx = \frac{1}{3 \cdot (b-a)} \cdot (b^3 - a^3)$$

und hieraus mit (12.9)

$$\sigma^2 = \frac{b^3 - a^3}{3 \cdot (b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^3}{12 \cdot (b-a)} = \frac{(b-a)^2}{12}. \quad (13.13)$$

Für die über $[0,10]$ rechteckverteilte Zufallsvariable X aus Beispiel 13.1 erhält man z. B. den Erwartungswert $\mu = 5$, die Varianz $\sigma^2 = \frac{25}{3} \approx 8,33$ bzw. die Standardabweichung $\sigma = \sqrt{\frac{25}{3}} \approx 2,89$. Der Wert $\mu = 5$ beinhaltet, dass man „im Mittel“ mit 5 Minuten Wartezeit zu rechnen hat.

Weitere Kenngrößen Neben dem Erwartungswert und der Varianz bzw. der Standardabweichung kann man noch die **Quantile** x_p heranziehen (p -Quantile), die nach (12.17) für jedes p mit $0 < p < 1$ durch $F(x_p) = p$ definiert sind. Die Quantile sind durch diese Gleichung bei stetigen Verteilungen – anders als bei diskreten Verteilungen, deren Verteilungsfunktionen ja durch Treppenfunktionen definiert sind – eindeutig erklärt, da die Verteilungsfunktion streng monoton wächst.¹ Der Median $\tilde{x} = x_{0,5}$ bezeichnet dann

¹Das p -Quantil x_p einer stetigen Zufallsvariablen mit Dichtefunktion $f(x)$ hat die Eigenschaft, denjenigen Wert auf der x -Achse zu definieren, der die Fläche zwischen x -Achse und Dichtefunktion so in zwei Teilträume zerlegt, dass die Teilträume bis zum Punkt x_p den Inhalt p haben, also $p \cdot 100\%$ der Gesamtfläche ausmachen.

den Punkt auf der x -Achse, für den $F(x) = 0,5$ ist. Von besonderer Bedeutung für das Testen von Hypothesen sind p - und $(1 - p)$ -Quantile mit kleinen Werten von p , etwa $p = 0,05$ oder $p = 0,01$. Sie haben hier die Bedeutung von Irrtumswahrscheinlichkeiten.

13.3 Normalverteilung und Standardnormalverteilung

Die Normalverteilung ist die für die Modellierung von Zufallsvorgängen weitaus wichtigste Verteilung. Sie geht auf Carl Friedrich GAUSS (1777 – 1855) zurück, der die Funktionsgleichung der glockenförmigen Dichte dieser Verteilung ableitete und erstmals auf praktische Probleme bezog. In Erinnerung an diese Pionierleistung war GAUSS mit der Dichtekurve der Normalverteilung im Hintergrund auf der Vorderseite des früheren 10-DM-Scheins abgebildet. Die Bedeutung der Normalverteilung röhrt daher, dass sie andere Verteilungen unter gewissen Voraussetzungen gut approximiert. Die Normalverteilung wird z. B. häufig zur Modellierung von Zufallsvorgängen eingesetzt, bei denen mehrere zufällige Einflussgrößen zusammenwirken. Dies gilt etwa für die industrielle Überwachung von Serienfertigungen, bei der ein stetiges Qualitätsmerkmal üblicherweise als zumindest approximativ normalverteilt angenommen wird. Aus der Normalverteilung leiten sich zudem wichtige Verteilungen ab, die beim Testen von Hypothesen als Teststatistiken verwendet werden.

Eine Zufallsvariable X folgt einer **Normalverteilung**, wenn ihre Dichte die Gestalt

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{für alle reellen } x \quad (13.14)$$



Karl Friedrich GAUSS

Dichte- und Verteilungsfunktion der Normalverteilung

besitzt.² Man entnimmt dieser Gleichung und auch den beiden folgenden Abbildungen, dass die Dichte der Normalverteilung von μ und σ^2 abhängt und bezüglich μ symmetrisch ist. Anhand der allgemeinen Formeln (13.9) und (13.10) kann man verifizieren, dass die in (13.14) eingehenden Terme μ und σ^2 nichts anderes als der Erwartungswert resp. die Varianz der Normalverteilung sind. Für eine Zufallsvariable X mit der Dichte (13.14) sagt man, dass X mit den Parametern μ und σ^2 normalverteilt sei. Hierfür wird oft die Kurznotation $X \sim N(\mu; \sigma^2)$ verwendet (lies: X ist normalverteilt mit Erwartungswert μ und Varianz σ^2).

²Die Schreibweise $\exp x$ bedeutet nichts anderes als e^x . Sie wird gerne verwendet, wenn im Exponenten Brüche stehen, weil die Brüche dann nicht hochgestellt erscheinen und damit besser lesbar sind.

Für die Verteilungsfunktion der Normalverteilung gilt mit (13.2)

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt. \quad (13.15)$$



Interaktives Objekt
„Normalverteilung“

Die Funktion ist nicht in geschlossener Form darstellbar, d. h. das Integral lässt sich nicht durch elementare Funktionen ausdrücken. Die Werte von $F(x)$ lassen sich aber unter Verwendung von Näherungsverfahren ermitteln. Dichte- und Verteilungsfunktion der Normalverteilung sind über die Beziehung (13.3) miteinander verbunden.

Im linken Teil von Abbildung 13.2 ist die Normalverteilung mit $\mu = 1$ und $\sigma^2 = 0,6^2 = 0,36$ wiedergegeben, rechts die Normalverteilung mit $\mu = -0,5$ und $\sigma^2 = 0,4^2 = 0,16$. Die Verteilungen sind bezüglich ihres Zentrums μ symmetrisch. Bei der rechts dargestellten Normalverteilung verlaufen Dichte- und Verteilungsfunktion steiler. Die in der unteren Hälfte beider Abbildungsteile eingezeichneten horizontalen gepunkteten Linien sollen verdeutlichen, dass der zum Abszissenwert μ gehörende Ordinatenwert bei jeder Normalverteilung den Wert 0,5 hat.

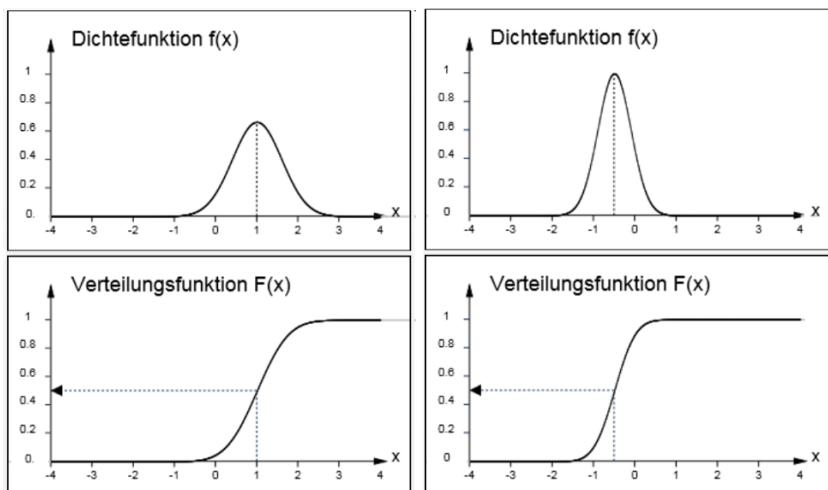


Abb. 13.2: Dichte- und Verteilungsfunktion der beiden Normalverteilungen $N(1; 0,6^2)$ und $N(-0,5; 0,4^2)$

Die Gestalt von Dichte- und Verteilungsfunktion der Normalverteilung hängt offenbar vom Erwartungswert μ und der Varianz σ^2 ab. Allgemein gilt, dass beide Funktionen einen steileren Verlauf annehmen, wenn man die Varianz σ^2 verkleinert. Für jede Normalverteilung gilt ferner, dass ihre Dichte jeweils in $x = \mu$ ihr Maximum annimmt und die Verteilungsfunktion dort einen Wendepunkt besitzt, zu dem stets der Ordinatenwert 0,5 gehört. Die Wendepunkte der Dichte jeder Normalverteilung liegen in $x = \mu - \sigma$ und $x = \mu + \sigma$.

Unterzieht man eine $N(\mu; \sigma^2)$ -verteilte Zufallsvariable X einer Lineartransformation $Y = aX + b$ mit $a \neq 0$, so ist die transformierte Variable Y ebenfalls normalverteilt, wobei sich Erwartungswert und Varianz von Y aus (12.11) und (12.12) ergeben:

$$X \sim N(\mu; \sigma^2), \quad Y = aX + b \implies Y \sim N(a\mu + b; a^2\sigma^2). \quad (13.16)$$

Für die Summe zweier unabhängiger normalverteilter Zufallsvariablen X und Y gilt ferner³

$$\begin{aligned} X &\sim N(\mu_X; \sigma_X^2), \quad Y \sim N(\mu_Y; \sigma_Y^2), \quad X \text{ und } Y \text{ unabh.} \\ \rightarrow X + Y &\sim N(\mu_X + \mu_Y; \sigma_X^2 + \sigma_Y^2). \end{aligned} \quad (13.17)$$

Man kann jede Normalverteilung auf die als **Standardnormalverteilung** bezeichnete Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ zurückführen. Hat man nämlich eine normalverteilte Zufallsvariable $X \sim N(\mu; \sigma^2)$, so kann man diese stets der Lineartransformation $Z := \frac{X-\mu}{\sigma}$ aus (13.11) unterziehen. Für die Zufallsvariable Z gilt dann $Z \sim N(0,1)$ (lies: Z ist *normalverteilt mit Erwartungswert 0 und Varianz 1* oder Z ist *standardnormalverteilt*):

$$X \sim N(\mu; \sigma^2) \xrightarrow{\text{Transformation von } X \text{ in } Z=(X-\mu)/\sigma} Z \sim N(0,1).$$

Die Dichtefunktion der Standardnormalverteilung geht aus (13.14) nach Einsetzen von $\mu = 0$ und $\sigma^2 = 1$ hervor. Für sie hat sich anstelle von $f(..)$ die spezielle Notation $\phi(..)$ (lies: *Klein-Phi*) eingebürgert:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \quad (13.18)$$

Für die Verteilungsfunktion der Standardnormalverteilung hat sich die Bezeichnung $\Phi(..)$ (lies: *Groß-Phi*) etabliert. Sie ist durch

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{t^2}{2}\right) dt \quad (13.19)$$

erklärt und wie (13.15) nicht in geschlossener Form darstellbar. Ihre Werte lassen sich anhand numerischer Verfahren bestimmen.

Lineartransformation
normalverteilter
Zufallsvariablen

Standardisierung der
Normalverteilung



Interaktives Objekt
„Standardnormalverteilung“

³Der Begriff der „Unabhängigkeit“ von Zufallsvariablen wird in Abschnitt 14.1 noch formalisiert. Eine Herleitung von (13.16) und eine Verallgemeinerung von (13.17) findet man bei FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 6.3.1).

In Abbildung 13.3 sind Dichte- und Verteilungsfunktion der Standardnormalverteilung wiedergegeben sowie ein Quantil der Verteilung. Die Quantile der Standardnormalverteilung sind durch die noch folgende Gleichung (13.24) erklärt.

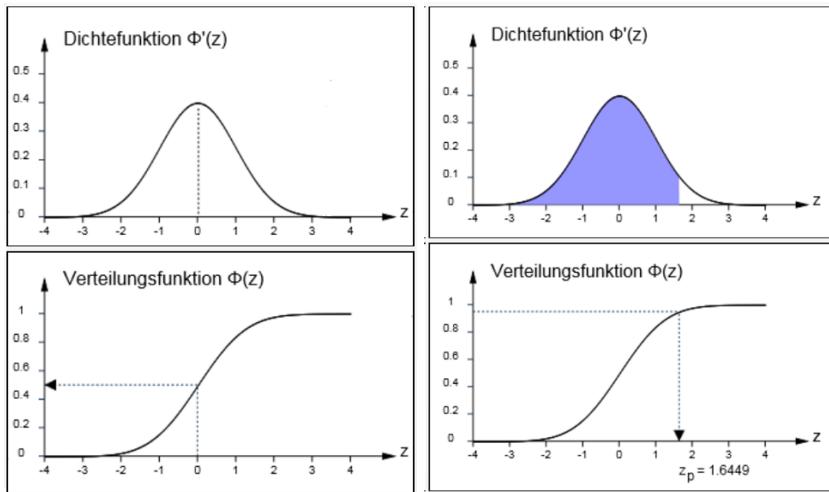


Abb. 13.3: Die Standardnormalverteilung. Links: Dichte- und Verteilungsfunktion. Rechts: 0,95-Quantil der Standardnormalverteilung

Aus der Abbildung geht hervor, dass für $\Phi(z)$ die nachstehende Symmetriebeziehung gilt:

$$\Phi(-z) = 1 - \Phi(z). \quad (13.20)$$

In Tabelle 19.2 des Anhangs sind Werte der Verteilungsfunktion $\Phi(z)$ für den Bereich $0 \leq z < 4$ tabelliert. Für negative Werte von z lassen sich die Werte der Verteilungsfunktion mit (13.20) bestimmen.



Interaktives Objekt
„Rechnen mit der
Standard-
normalverteilung“

Wegen $\Phi'(z) = \phi(z)$ ist der Wert $\Phi(z)$ an der Stelle $z = a$ als Inhalt der Fläche unter der Dichte bis zum Punkt $z = a$ interpretierbar. Der linke Teil von Abbildung 13.4 illustriert dies für den Abszissenwert $z = 1$. Die farbig markierte Fläche im oberen Teil endet hier bei $z = 1$ und hat den Inhalt $P(Z \leq 1) = \Phi(1) = 0,8413$. Dieser Wert ist im unteren Teil der linken Abbildung anhand eines auf die Ordinatenachse weisenden horizontalen Pfeils betont. Die nicht markierte Restfläche unter der Dichte hat den Inhalt $P(Z > 1) = 1 - \Phi(1) = 0,1587$.

Der rechte Teil von Abbildung 13.4 stellt die Differenz $\Phi(2) - \Phi(-1)$ als Inhalt der von $z = -1$ bis $z = 2$ gerechneten Fläche unter der Dichte dar. Die Werte $\Phi(2) = 0,9772$ und $\Phi(-1) = 0,1587$ sind auch hier durch horizontale Pfeile betont und die Differenz 0,8185 durch einen roten Doppelpfeil. Dessen Länge entspricht dem Inhalt der farbigen Fläche.

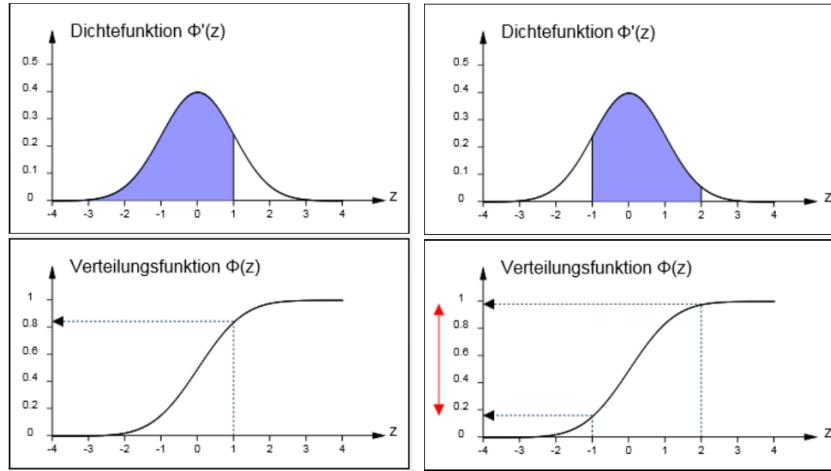


Abb. 13.4: Interpretation von Werten der Verteilungsfunktion $\Phi(z)$ als Fläche unterhalb der Dichte. Links: Veranschaulichung von $\Phi(1)$. Rechts: Visualisierung der Differenz $\Phi(2) - \Phi(-1)$

Mit den Werten $\Phi(z)$ aus Tabelle 19.2 kann man Werte $F(x)$ der Verteilungsfunktion *jeder* beliebigen Normalverteilung bestimmen. Gilt nämlich $X \sim N(\mu; \sigma^2)$, so besteht zwischen den Verteilungsfunktionen $F(x)$ von X und $\Phi(z)$ von $Z = \frac{X-\mu}{\sigma}$ die Beziehung

$$F(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Man leitet hieraus die folgenden Darstellungen ab:

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \quad (13.21)$$

$$P(X > a) = 1 - P(X \leq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \quad (13.22)$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \quad (13.23)$$

Das p -Quantil der Normalverteilung ist der Wert x_p , an dem die Verteilungsfunktion $F(x)$ den Wert p erreicht. Die **p-Quantile der Standardnormalverteilung** sind also durch

$$\Phi(z_p) = p \quad (13.24)$$

definiert. In der rechten unteren Hälfte von Abbildung 13.3 ist beispielhaft das 0,95-Quantil visualisiert.



Da die Dichte der Standardnormalverteilung symmetrisch zum Nullpunkt ist, gilt

$$z_p = -z_{1-p}. \quad (13.25)$$



Objekt „Quantile der Standardnormalverteilung“

Abbildung 13.5 veranschaulicht diese Symmetrieeigenschaft der Quantile für $p = 0,025$. Die rote Fläche an der linken Flanke der dargestellten Dichtekurve endet bei $z_{0,025}$, die an der rechten Flanke beginnt bei $z_{0,975}$. Beide markierten Flächen haben den Inhalt 0,025 und es gilt $z_{0,025} = -z_{0,975} = -1,96$.

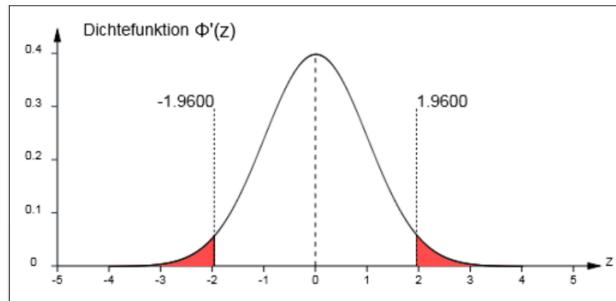


Abb. 13.5: Symmetrie der Quantile z_p und z_{1-p} der Standardnormalverteilung bezüglich des Erwartungswerts $\mu = 0$ (hier: $p = 0,025$)

Ist $X \sim N(\mu; \sigma^2)$, so sind die Quantile x_p dieser Verteilung mit denen der Standardnormalverteilung über die zu (13.11) analoge Gleichung

$$z_p = \frac{x_p - \mu}{\sigma} \quad (13.26)$$

verbunden, d. h. es ist

$$x_p = \mu + z_p \cdot \sigma.$$

Die Standardisierung einer normalverteilten Zufallsvariablen X beinhaltet demnach nur eine Reskalierung der x -Achse.

Der in Abbildung 13.5 zwischen den beiden rot markierten Flächen liegende Bereich wird als **zentrales Schwankungsintervall der Standardnormalverteilung** bezeichnet. Ist das Intervall durch $[z_{\alpha/2}; z_{1-\alpha/2}]$ definiert, spricht man präziser vom zentralen Schwankungsintervall der Standardnormalverteilung zur Sicherheit $1 - \alpha$. Eine Realisation z der Standardnormalverteilung ist mit Wahrscheinlichkeit $1 - \alpha$ in diesem Intervall enthalten.

Beispiel 13.2: Berechnung von Wahrscheinlichkeiten und Quantilen

Für eine $N(\mu; \sigma^2)$ -verteilte Zufallsvariable mit $\mu = 1$ und $\sigma^2 = 2,25$ leitet man mit (13.20) – (13.21) und Tabelle 19.2 folgende Aussagen her:

$$P(X \leq 4) = \Phi\left(\frac{4-1}{1,5}\right) = \Phi(2) = 0,9772;$$

$$P(X \leq -0,5) = \Phi\left(\frac{-0,5-1}{1,5}\right) = \Phi(-1) = 1 - \Phi(1) = 0,1587;$$

$$P(-0,5 \leq X \leq 4) = \Phi(2) - \Phi(-1) = 0,8185.$$

Die Werte 0,9772 und 0,1587 entsprechen den Flächeninhalten unter der Dichte der standardnormalverteilten Variablen Z aus (13.11) bis zum Punkt $z = 2$ bzw. $z = -1$, der Wert 0,8185 dem Inhalt der Fläche von $z = -1$ bis $z = 2$.

**Aufgabe 13.2-3**

Will man für die $N(1; 2,25)$ -verteilte Variable X die Quantile $x_{0,975}$ und $x_{0,025}$ berechnen, bestimmt man zunächst $z_{0,975}$ und $z_{0,025}$ unter Verwendung von Tabelle 19.3 und (13.25). Man erhält $z_{0,975} = 1,96$ und $z_{0,025} = -1,96$. Daraus folgt dann mit (13.26) resp. mit $x_p = \mu + \sigma \cdot z_p$ für die gesuchten Quantile

$$x_{0,975} = 1 + 1,96 \cdot 1,5 = 3,94, \quad x_{0,025} = 1 - 1,96 \cdot 1,5 = -1,94.$$

Die Wahrscheinlichkeit dafür, dass X im Intervall $[x_{0,025}; x_{0,975}] = [-1,94; 3,94]$ liegt, beträgt 0,95. Dieser Wert ist mit der Wahrscheinlichkeit identisch, dass Z Werte in $[z_{0,025}; z_{0,975}] = [-1,96; 1,96]$ annimmt (vgl. Abbildung 13.5).

Quantile der Standardnormalverteilung spielen beim Testen von Hypothesen eine wichtige Rolle. Es sind vor allem p -Quantile mit relativ kleinem oder relativ großem p , z. B. $p = 0,01$ oder $p = 0,99$. Diese häufig verwendeten Quantile sind in Tabelle 19.3 zusammengefasst. Wegen (13.25) beschränkt sich die Tabelle auf p -Quantile mit $p > 0,5$.

Beispiel 13.3: Intelligenzmessung

In der *Psychologie* misst man Intelligenz anhand von psychologischen Tests. Diese basieren auf einer möglichst repräsentativen Bevölkerungsstichprobe, die nach bestimmten Kriterien (Alter, Geschlecht) in Teilstichproben aufgegliedert wird. Ein individuelles Testergebnis kann dann zum durchschnittlichen Wert der jeweiligen Alters- und Geschlechtsgruppe in Beziehung gesetzt werden. Die Teilstichproben stellen sozusagen unterschiedliche Grundgesamtheiten dar.

Für die Aufgaben eines Intelligenztests werden Punkte vergeben und aufsummiert. Für jede Person resultiert ein Punktkrohwert oder Summenscore x , der sich als Ausprägung einer diskreten Zufallsvariablen X interpretieren lässt. Da sich die Verteilung von X i. Allg. gut durch eine Normalverteilung approximieren lässt und diese besonders einfach handhabbar ist, wird die Normalverteilung als Modell für die Verteilung der Zufallsvariablen „Summenscore X “ herangezogen.

Die Verteilungsparameter μ und σ^2 der Normalverteilung hängen von der betrachteten Grundgesamtheit ab. Man könnte nun die Summenscores standardisieren und mit der Standardnormalverteilung arbeiten. Aus historischen Gründen geht man aber in der Praxis nicht zur Standardnormalverteilung über, sondern zur Normalverteilung mit Erwartungswert $\mu = 100$ und Standardabweichung $\sigma = 15$. Man transformiert also X in eine $N(100, 15^2)$ -verteilte Variable Y . Diese Transformation kann anhand von Abbildung 13.3 verdeutlichen, wenn man dort unter die z -Achse noch eine y -Achse einzeichnet, die an der Stelle $z = 0$ den Wert $y = 100$ und für $z = -1$ bzw. $z = 1$ die Werte $y = 85$ resp. $y = 115$ annimmt. Formal lässt sich der Übergang vom Summenscore X zur transformierten Zufallsvariablen Y in zwei Schritte zerlegen. Im ersten Schritt wird X gemäß (13.11) in Z überführt, im zweiten Schritt wird Z noch in $Y = 100 + 15 \cdot Z$ transformiert. Die Realisationen von Y ergeben sich also aus den ursprünglichen individuellen Rohwerten x nach

$$y = 100 + 15 \cdot z = 100 + 15 \cdot \frac{x - \mu}{\sigma}.$$

Der errechnete y -Wert, also die individuelle Ausprägung der latenten Variablen „Intelligenz“, wird als *Intelligenzquotient* (kurz IQ) bezeichnet. Die Wahrscheinlichkeit dafür, dass eine zufällig aus der betrachteten Population ausgewählte Person einen IQ-Wert zwischen 85 und 115 hat, errechnet sich z. B. mit Tabelle 19.2 und Beachtung von $\Phi(-1) = 1 - \Phi(1)$ nach

$$P(85 \leq Y \leq 115) = P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = 2 \cdot \Phi(1) - 1 \approx 0,683,$$

also als 68,3 %. Quantile, mit 100 multipliziert, werden in der Psychologie auch als *Prozentränge* angesprochen. Der 99,5-Prozentrang der bei der Intelligenzmessung verwendeten Normalverteilung bezeichnet also z. B. denjenigen IQ-Wert $y = y_{0,995}$, der von nicht mehr als 0,5 % der betrachteten Grundgesamtheit überschritten wird. Man erhält mit Tabelle 19.3 den Wert

$$y_{0,995} = 100 + 15 \cdot z_{0,995} \approx 100 + 15 \cdot 2,5758 \approx 138,64.$$

Exkurs 13.2: Risikomaß „Value at Risk“

Im Finanzsektor spielen p -Quantile stetiger Verteilungen mit kleinen Werten p eine Rolle bei der Abschätzung potenzieller Verluste. Wenn man etwa die Tages- oder Monatsrenditen einer Aktie oder eines Wertpapier-Portfolios über einen längeren Zeitraum erfasst, so bietet es sich an, die Variable „Tagesrendite“ resp. „Monatsrendite“ anhand einer stetigen Zufallsvariablen X zu modellieren. Kann man aufgrund der beobachteten Renditen davon ausgehen, dass X den Erwartungswert 0 hat, so sind negative Realisationen Verluste. Das durch $P(X < x_{0,05}) = 0,05$ definierte (negative) 0,05-Quantil $x_{0,05}$ der Verteilung kennzeichnet dann eine Rendite, mit der man wegen $P(X \geq x_{0,05}) = 0,95$ mit Wahrscheinlichkeit 0,95 mindestens rechnen darf. Schlechtere Renditen als $x_{0,05}$, d. h. höhere Verluste, sind nur mit Wahrscheinlichkeit 0,05 zu erwarten.

Das p -Quantil – etwa mit $p = 0,05$ – der zugrunde gelegten Verteilung liefert somit einen Verlustwert, der in der betrachteten Halteperiode mit Wahrscheinlichkeit $1 - p$ nicht überschritten wird. Im Finanzsektor wird dieser mit einer Wahrscheinlichkeitsaussage verknüpfte Schwellenwert **Value at Risk** genannt. Die Wahrscheinlichkeitsaussage, die sich auf einen Value at Risk bezieht, hängt natürlich entscheidend von der zugrunde gelegten Verteilung ab. Approximativ wird oft mit der Normalverteilung gearbeitet. Hohe Verluste oder hohe Gewinne treten aber in der Realität mit höherer Wahrscheinlichkeit auf, als bei Gültigkeit des Normalverteilungsmodells zu erwarten wäre. Ein realitätsnäheres Modell müsste also im Vergleich zur Normalverteilung etwas stärker besetzte Flanken aufweisen (engl.: *fat tails*).

Aber selbst ein realitätsnahes Modell kann nur innerhalb unspektakulärer Börsenperioden hilfreich sein. Bei Eintritt folgenreicher, unerwarteter Ereignisse (Konkurs der Investmentbank Lehman Brothers und Finanzkrise 2008, Anschlag auf das World Trade Center und Fukushima-Katastrophe 2011, Ausgang des Brexit-Referendums und der US-Präsidentenwahl 2016, Corona-Pandemie 2020) zeigen sich die Grenzen von Modellen zur Risikoabschätzung von Anlagen. Nach solchen Ereignissen kann es notwendig sein, die Modelle an die neuen Gegebenheiten anzupassen.

13.4 χ^2 -, t - und F -Verteilung

Aus der Normalverteilung lassen sich einige Verteilungen ableiten, die im Zusammenhang mit der Schätzung von Modellparametern und dem Testen von Hypothesen benötigt werden. Es sind dies vor allem die χ^2 -Verteilung, die t -Verteilung und die F -Verteilung. Erstere wird u. a. zum Testen von Hypothesen über die Varianz einer Normalverteilung verwendet. Die t -Verteilung findet z. B. Verwendung beim Testen von Hypothesen zum Erwartungswert einer normalverteilten Zufallsvariablen, deren Varianz nicht bekannt ist. Die F -Verteilung spielt u. a. bei der Varianzanalyse eine zentrale Rolle als Teststatistik.

Geht man von n unabhängigen standardnormalverteilten Variablen Z_1, Z_2, \dots, Z_n aus und bildet die Summe

$$X := Z_1^2 + Z_2^2 + \dots + Z_n^2 = \sum_{i=1}^n Z_i^2 \quad (13.27)$$

Charakterisierung
der χ^2 -Verteilung

der quadrierten Variablen, so sagt man, dass die Verteilung der resultierenden Variablen X einer **χ^2 -Verteilung** mit n Freiheitsgraden folgt und verwendet die Kurznotation $X \sim \chi_n^2$ (lies: X ist χ^2 -verteilt mit n Freiheitsgraden). Man kann die χ^2 -Verteilung auch, analog zu (13.14), über die Dichtefunktion definieren. So verfahren etwa ZUCCHINI / SCHLEGEGL / NENADIC / SPERLICH (2009, Abschnitt 6.4.1). Dieser Weg ist

naheliegend – die Aussage, dass die Zufallsvariable X aus (13.27) einer χ^2 -Verteilung folgt, wäre dann nur eine abgeleitete Aussage. Die elegantere Einführung der χ^2 -Verteilung direkt über die Dichtefunktion wird hier und auch bei der noch folgenden t -Verteilung aber nicht beschritten, weil die Dichtefunktionen beider Verteilungen relativ sperrig sind.

Die Anzahl n der in (13.27) eingehenden Summanden ist ein Parameter, der die Form der Dichtefunktion der χ^2 -Verteilung determiniert. Mit der Anzahl der **Freiheitsgrade** der χ^2 -Verteilung ist also dieser Formparameter gemeint. Aus (13.9) und (13.10) sowie der Dichtefunktion der Verteilung lassen sich für den Erwartungswert und die Varianz einer χ^2_n -verteilten Variablen X die nachstehenden Gleichungen ableiten:

$$E(X) = n$$

$$V(X) = 2n.$$



Interaktives Objekt
„Quantile der
 χ^2 -Verteilung“

In Abbildung 13.6 sind Dichte- und Verteilungsfunktion der χ^2 -Verteilung für zwei ausgewählte Freiheitsgrade n grafisch dargestellt. Der linke Teil der Abbildung zeigt die χ^2 -Verteilung mit $n = 4$ Freiheitsgraden und das 0,95-Quantil $\chi_{4;0,95}^2 = 9,488$. Letzteres ist dadurch charakterisiert, dass die Verteilungsfunktion hier den Wert 0,95 hat. Der Wert 0,95 ist auch durch den Inhalt der blau markierten Fläche unter der Dichte repräsentiert. Der rechte Teil bezieht sich auf die χ^2 -Verteilung mit $n = 6$ Freiheitsgraden und deren 0,90-Quantil $\chi_{6;0,90}^2 = 10,645$. Der blau markierte Inhalt unter der Dichte hat hier den Wert 0,90.

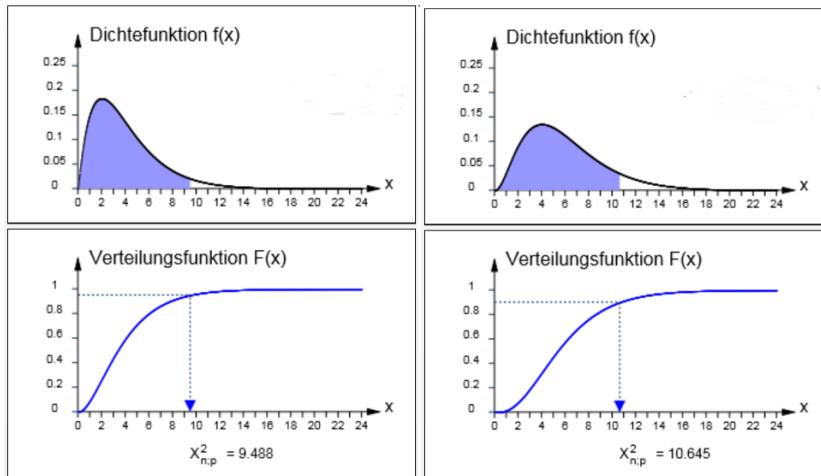


Abb. 13.6: Dichte- und Verteilungsfunktion zweier χ^2 -Verteilungen. Links:
 χ^2 -Verteilung mit $n = 4$ Freiheitsgraden und 0,95-Quantil.
Rechts: χ^2 -Verteilung mit $n = 6$ Freiheitsgraden und 0,90-Quantil

Die wiedergegebenen Dichtekurven fallen – ähnlich wie die empirischen Verteilungen für Bruttoverdienste aus Abbildung 4.8 – jeweils an der linken Flanke steiler ab. Man spricht daher von einer **linkssteilen** oder **rechtsschiefen Verteilung**. Bei einer **rechtssteilen** oder **linksschiefen Verteilung** würde die rechte Flanke steiler abfallen. In beiden Fällen liegt eine **asymmetrische Verteilung** vor.

Man sieht, dass die Gestalt der Dichtefunktion $f(x)$ und damit auch die der Verteilungsfunktion $F(x)$ einer χ^2 -Verteilung in der Tat stark von der Anzahl n der Freiheitsgrade abhängt. Gleiches gilt somit insbesondere für die durch (12.17) erklärten Quantile, die mit $\chi_{n;p}^2$ abgekürzt werden (lies: *p-Quantil der χ^2 -Verteilung mit n Freiheitsgraden*).

Dichte- und Verteilungsfunktion der χ^2 -Verteilung werden i. Allg. nur zur Berechnung von Quantilen gebraucht. Man benötigt die Quantile beim Testen, wenn die Testvariable χ^2 -verteilt ist. Da die Berechnung der Quantile rechenaufwändig ist, greift man auf Tabellen zurück, z. B. auf Tabelle 19.4. Hier sind Quantile $\chi_{n;p}^2$ für $n = 1$ bis $n = 25$ und ausgewählte Werte p zusammengestellt. Mit Tabelle 19.4 verifiziert man etwa die im rechten unteren Teil von Abbildung 13.6 veranschaulichte Aussage, dass das 0,95-Quantil der χ^2 -Verteilung mit $n = 4$ Freiheitsgraden den Wert $\chi_{4;0,95}^2 = 9,488$ und das 0,90-Quantil der χ^2 -Verteilung mit $n = 6$ Freiheitsgraden den Wert $\chi_{6;0,90}^2 = 10,645$ hat.

Aus der Standardnormalverteilung und der χ^2 -Verteilung leitet sich die **t-Verteilung** ab, die gelegentlich auch **Student-Verteilung** genannt wird. Sind X und Z unabhängige Zufallsvariablen mit $X \sim \chi_n^2$ und $Z \sim N(0; 1)$,⁴ dann folgt die Zufallsvariable

$$T := \frac{Z}{\sqrt{\frac{X}{n}}} \quad (13.28)$$

einer t -Verteilung mit n Freiheitsgraden und man schreibt $T \sim t_n$ (lies: *T ist t-verteilt mit n Freiheitsgraden*). Erstmals beschrieben wurde die t -Verteilung von William S. GOSSET (1876 – 1937). Dabei verwendete er anstelle seines Namens das Pseudonym STUDENT. Hieraus erklärt sich die Bezeichnung „Student-Verteilung“.



William S. GOSSET

Auf die direkte Einführung der t -Verteilung über ihre Dichtefunktion wird aufgrund der Komplexität der Dichteformel verzichtet und auf ZUCCHINI / SCHLEGEL / NENADIC / SPERLICH (2009, Abschnitt 6.4.3) verwiesen.

Charakterisierung
der t -Verteilung

Die in (13.28) eingehende Anzahl n der Freiheitsgrade bezeichnet wieder einen Formparameter. Für den Erwartungswert und die Varianz einer

⁴Eine Formalisierung des Begriffs „Unabhängigkeit von Zufallsvariablen“ erfolgt in Abschnitt 14.1.

t_n -verteilten Variablen T lässt sich zeigen, dass

$$\begin{aligned} E(T) &= 0 & (n > 1) \\ V(T) &= \frac{n}{n-2} & (n > 2). \end{aligned}$$

Die Dichte der t -Verteilung ist wie die der Standardnormalverteilung symmetrisch zum Nullpunkt. Analog zu (13.25) gilt daher für die durch (12.17) erklärten Quantile $t_{n;p}$ der t -Verteilung (lies: p -Quantil der t -Verteilung mit n Freiheitsgraden)

$$t_{n;p} = -t_{n;1-p}. \quad (13.29)$$

In Abbildung 13.7 sind Dichte- und Verteilungsfunktionen der t -Verteilung für zwei ausgewählte Freiheitsgrade n visualisiert. Der linke Teil zeigt die t -Verteilung mit $n = 3$ Freiheitsgraden einschließlich des 0,95-Quantils $t_{3;0,95} = 2,3534$. Der rechte Abbildungsteil bezieht sich auf die t -Verteilung mit $n = 10$ Freiheitsgraden und deren 0,975-Quantil $t_{10;0,975} = 2,2281$. Die Dichtekurve der t -Verteilung mit $n = 10$ Freiheitsgraden ist im Bereich des Erwartungswerts $\mu = 0$ höher und an den Flanken etwas schmäler.



Interaktives Objekt
„Quantile der
 t -Verteilung“

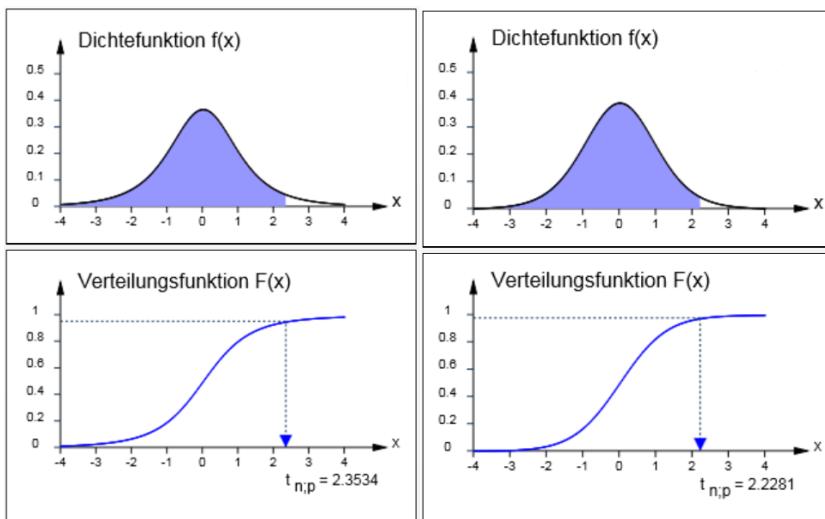


Abb. 13.7: Dichte- und Verteilungsfunktion zweier t -Verteilungen. Links: t -Verteilung mit $n = 3$ Freiheitsgraden und 0,95-Quantil. Rechts: t -Verteilung mit $n = 10$ Freiheitsgraden und 0,975-Quantil

Mit zunehmender Anzahl n der Freiheitsgrade nähert sich die t -Verteilung der Standardnormalverteilung an. Mit wachsendem n werden folglich auch die p -Quantile beider Verteilungen immer ähnlicher.

Abbildung 13.8 veranschaulicht die letztgenannte Aussage anhand der t -Verteilungen, die schon in Abbildung 13.7 zusammen mit dem 0,95- bzw. dem 0,975-Quantil wiedergegeben waren. Zum Vergleich ist nun jeweils die Standardnormalverteilung mit den entsprechenden Quantilen $z_{0,95} = 1,6449$ resp. $z_{0,975} = 1,9600$ eingeblendet. Man sieht, dass sich die t -Verteilung mit 10 Freiheitsgraden und die Standardnormalverteilung schon nicht mehr sehr stark unterscheiden.

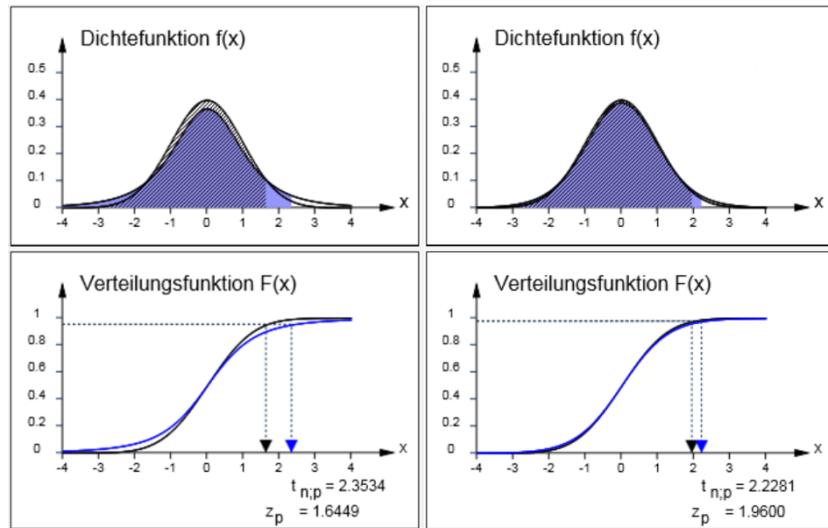


Abb. 13.8: Dichte- und Verteilungsfunktionen zweier t -Verteilungen und der Standardnormalverteilung. Links: t -Verteilung mit $n = 3$ Freiheitsgraden und Standardnormalverteilung mit 0,95-Quantilen. Rechts: t -Verteilung mit $n = 10$ Freiheitsgraden und Standardnormalverteilung mit 0,975-Quantilen

Für große n kann man die Quantile $t_{n;p}$ durch Quantile z_p der Standardnormalverteilung approximieren. Tabelle 13.1 illustriert die Größenordnung der Quantile $t_{n;p}$ und z_p für einige Werte n und p . In der Tabelle sind die vier in Abbildung 13.8 betonten Quantile grau hinterlegt.



Aufgabe 13.4

p	$t_{3;p}$	$t_{10;p}$	$t_{15;p}$	$t_{30;p}$	$t_{40;p}$	z_p
0,95	2,3534	1,8125	1,7531	1,6973	1,6839	1,6449
0,975	3,1824	2,2281	2,1314	2,0423	2,0211	1,9600
0,99	4,5407	2,7638	2,6025	2,4573	2,4233	2,3263

Tab. 13.1: Quantile der t -Verteilung und der Standardnormalverteilung

Der Vergleich der Werte $t_{30;p}$ und $t_{40;p}$ mit den Werten z_p zeigt, dass die Approximation von $t_{n;p}$ durch z_p ab $n = 30$ schon recht gut ist. Weitere Quantile der t -Verteilung sind in Tabelle 19.5 des Anhangs zu finden. Sie sind dort auf nur 3 Dezimalstellen genau ausgewiesen.



Eine Verteilung, die sich aus der χ^2 -Verteilung ableitet und häufig beim Testen von Hypothesen in der Regressions- und Varianzanalyse benötigt wird, ist die **F-Verteilung**. Sind X_1 und X_2 zwei unabhängige Zufallsvariablen mit $X_1 \sim \chi_m^2$ und $X_2 \sim \chi_n^2$, so folgt die Zufallsvariable

$$Y := \frac{X_1/m}{X_2/n} \quad (13.30)$$

Interaktives Objekt
„Quantile der
F-Verteilung“

Charakterisierung der
F-Verteilung

einer F -Verteilung mit m und n Freiheitsgraden. Man schreibt dann $Y \sim F_{m;n}$ (liest: Y ist F -verteilt mit m und n Freiheitsgraden). Formeldarstellungen für die Dichtefunktion der F -Verteilung findet man bei ZUCCHINI / SCHLEGEL / NENADIC / SPERLICH (2009, Abschnitt 6.4.2).

In Abbildung 13.9 sind beispielhaft die Dichte- und Verteilungsfunktionen zweier $F_{m;n}$ -verteilter Zufallsvariablen mit je einem p -Quantil $F_{m;n;p}$ visualisiert. Der linke Teil der Abbildung veranschaulicht die F -Verteilung mit $m = 10$ und $n = 15$ Freiheitsgraden sowie das zugehörige 0,95-Quantil $F_{10;15;0,95} = 2,544$. Der rechte Teil zeigt die F -Verteilung mit $m = 5$ und $n = 20$ Freiheitsgraden und deren 0,90-Quantil $F_{5;20;0,90} = 2,158$.

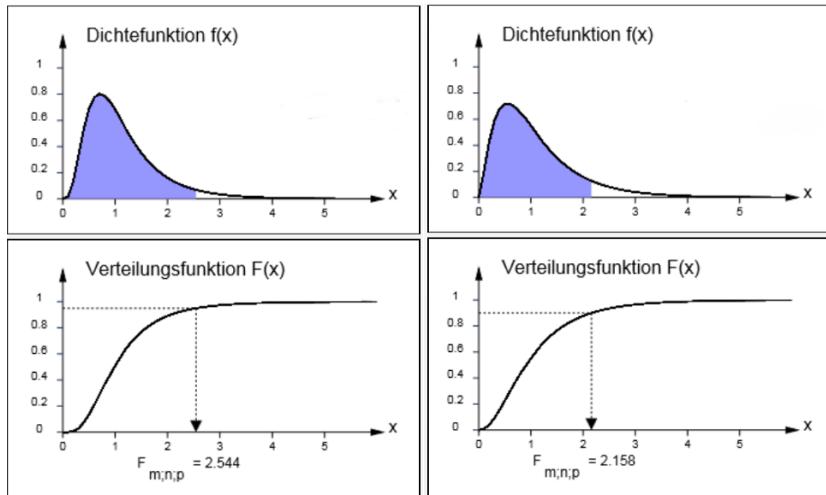


Abb. 13.9: Dichte- und Verteilungsfunktion zweier F -Verteilungen. Links: F -Verteilung mit $m = 10$ und $n = 15$ Freiheitsgraden und 0,95-Quantil. Rechts: F -Verteilung mit $m = 5$ und $n = 20$ Freiheitsgraden und 0,90-Quantil

Die Quantile bezeichnen jeweils die Positionen auf der Abszissenachse, bei der die markierten Flächen unter den Dichten enden. Die Inhalte dieser Flächen betragen 0,95 im Falle des 0,95-Quantils resp. 0,90 beim 0,90-Quantil. Der Wert des im linken Teil von Abbildung 13.9 dargestellten Quantils ist auch der Tabelle 19.6 zu entnehmen. Diese weist Quantile $F_{m;n;0,95}$ und $F_{m;n;0,99}$ für ausgewählte Freiheitsgrade m und n aus.

Für den Erwartungswert und die Varianz einer $F_{m;n}$ -verteilter Zufallsvariablen Y seien noch unter Verweis auf BAMBERG / BAUR / KRAPP (2017, Abschnitt 11.2.3) die Gleichungen

$$\begin{aligned} E(Y) &= \frac{n}{n-2} & (n > 2) \\ V(Y) &= \frac{2n^2 \cdot (m+n-2)}{m \cdot (n-2)^2 \cdot (n-4)} & (n > 4). \end{aligned}$$

angeführt. Ist $Y \sim F_{m;n}$, so folgt $W := \frac{1}{Y}$ einer F-Verteilung mit n und m Freiheitsgraden, also $W \sim F_{n;m}$. Für die mit $F_{m;n;p}$ bezeichneten p -Quantile einer $F_{m;n}$ -verteilten Zufallsvariablen Y gilt⁵

$$F_{m;n;p} = \frac{1}{F_{n;m;1-p}}. \quad (13.31)$$

Bei der Tabellierung von Quantilen der F-Verteilung kann man sich daher auf Quantile $F_{m;n;p}$ mit $m \leq n$ beschränken.

Exkurs 13.3: Formparameter von Verteilungen

Für die in den Kapiteln 12 – 13 vorgestellten diskreten und stetigen theoretischen Verteilungen wurden jeweils der Erwartungswert μ und die Varianz σ^2 bzw. die Standardabweichung σ als Maßzahlen zur Charakterisierung von Lage bzw. Streuung wiedergegeben. Daneben lassen sich auch Kenngrößen anführen, die die *Gestalt* einer theoretischen Verteilung beschreiben. Ein solcher Formparameter ist z. B. die Schiefe. Ist X eine Zufallsvariable mit Erwartungswert μ und Standardabweichung σ und $Z = \frac{X-\mu}{\sigma}$ die nach (13.11) standardisierte Fassung, so bezeichnet

$$\gamma_1 := E(Z^3) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E[(X-\mu)^3]}{\sigma^3}$$

die Schiefe oder – präziser – die *theoretische Schiefe* (engl.: *skewness*) der Verteilung von X . Bei einer symmetrischen Verteilung ist $\gamma_1 = 0$, weil hier $E[(X-\mu)^3] = 0$ gilt. Ist $\gamma_1 > 0$, liegt eine *rechtsschiefe* oder *linkssteile Verteilung* vor, im Falle $\gamma_1 < 0$ eine *linksschiefe* oder *rechtssteile Verteilung*. Die beiden in Abbildung 13.8 wiedergegebenen F-Verteilungen und auch die in Abbildung 12.6 dargestellte Binomialverteilung mit $p = 0,25$ sind linkssteile Verteilungen. Binomialverteilungen mit $p > 0,5$ sind rechtssteil.

⁵Folgt eine Zufallsvariable Y einer χ^2 - oder t-Verteilung mit n Freiheitsgraden oder einer F-Verteilung mit m und n Freiheitsgraden, werden hier die Notationen $Y \sim \chi_n^2$ resp. $Y \sim t_n$ und $Y \sim F_{m;n}$ verwendet und für die p -Quantile $\chi_{n;p}^2$, $t_{n;p}$ und $F_{m;n;p}$. In anderen Lehrbüchern findet man auch die Notationen $Y \sim \chi^2(n)$, $Y \sim t(n)$ und $Y \sim F(m; n)$ sowie für die Quantile $\chi_p^2(n)$, $t_p(n)$ und $F_p(m; n)$.

Ein weiterer Formparameter ist die *theoretische Wölbung* oder *Kurtosis* (engl.: *kurtosis*). Sie ist durch

$$\gamma_2 := E(Z^4) = E \left[\left(\frac{X - \mu}{\sigma} \right)^4 \right] = \frac{E[(X - \mu)^4]}{\sigma^4}$$

definiert und quantifiziert, wie stark sich die Wahrscheinlichkeitsmasse einer Verteilung um den Erwartungswert konzentriert und wie stark die Flanken einer Verteilung besetzt sind. Für die Wölbung einer beliebigen Normalverteilung gilt $\gamma_2 = 3$. Dieser Wert wird oft als Referenzwert herangezogen. Der sog. *Exzess* misst die Abweichung der Wölbung einer Verteilung vom Wert 3. Wölbung und Exzess spielen u. a. bei der Modellierung der Renditen von Aktien eine Rolle (vgl. ZUCCHINI / SCHLEGEL / NENADIC / SPERLICH (2009, Abschnitt 4.4.3)).

Die vorstehenden Ausführungen lassen sich auch auf empirische Verteilungen beziehen (vgl. TOUTENBURG / HEUMANN (2009, Abschnitt 3.3)). Hat man einen Datensatz x_1, \dots, x_n für ein Merkmal X , so ist die mit $\hat{\gamma}_1$ bezeichnete *empirische Schiefe* der Verteilung des Datensatzes durch

$$\hat{\gamma}_1 = \frac{1}{n} \cdot \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

erklärt. Analog ist die *empirische Wölbung* definiert. Das Symbol $\hat{\gamma}_1$ deutet an, dass man einen Datensatz als Realisationen einer Zufallsvariablen auffassen und die empirische Schiefe als Schätzung der theoretischen Schiefe γ_1 dieser Verteilung interpretieren kann.



14 Bivariate Verteilungen



Vorschau auf
das Kapitel

In diesem Kapitel wird zunächst definiert, wann zwei Zufallsvariablen als unabhängig gelten. Wenn man eine Stichprobe zieht und deren Elemente als Zufallsvariablen interpretiert, wird nämlich meist – z. B. bei der Verdichtung von Stichprobeninformation zu einer Stichprobenfunktion – Unabhängigkeit der Stichprobenvariablen unterstellt. Als Stichprobenfunktionen werden hier der Stichprobenmittelwert und die Stichprobenvarianz erwähnt. Modelliert man die Stichprobenelemente als Realisationen unabhängiger normalverteilter Zufallsvariablen, kann man Verteilungsaussagen für die genannten Stichprobenfunktionen ableiten. Beim Testen von Hypothesen werden für die Testentscheidung nur Quantile der Verteilungen von Stichprobenfunktionen benötigt.

Zur Messung des Zusammenhangs zwischen Zufallsvariablen werden die theoretische Kovarianz als nicht-normiertes und der Korrelationskoeffizient ρ als normiertes Maß vorgestellt. Am Ende des Kapitels wird anhand eines klassischen Modells zur Portfoliooptimierung verdeutlicht, dass diese Zusammenhangsmaße u. a. in Kapitalmarktmodellen eine wichtige Rolle spielen.

14.1 Unabhängigkeit von Zufallsvariablen

In Abschnitt 11.3 wurde der Begriff der Unabhängigkeit von *Ereignissen* erklärt. Zwei Ereignisse A und B gelten als unabhängig, wenn das Eintreten eines Ereignisses keinen Einfluss auf das jeweils andere Ereignis hat. Formal lässt sich Unabhängigkeit gemäß (11.16) definieren. Danach sind A und B unabhängig, wenn die Wahrscheinlichkeit $P(A \cap B)$ für das gleichzeitige Eintreten von A und B als Produkt der Eintrittswahrscheinlichkeiten $P(A)$ und $P(B)$ der Einzelereignisse darstellbar ist.

Zufallsvariablen nehmen Werte an, die sich als Ergebnisse von Zufallsvorgängen interpretieren lassen. Wenn eine diskrete Zufallsvariable eine bestimmte Ausprägung oder eine stetige Zufallsvariable eine Realisation innerhalb eines bestimmten Intervalls annimmt, sind auch dies Ereignisse mit bestimmten Eintrittswahrscheinlichkeiten. Der Unabhängigkeitsbegriff für Ereignisse lässt sich daher direkt auf Zufallsvariablen übertragen.

Eine Zufallsvariable X , gleich ob diskret oder stetig, lässt sich durch die Verteilungsfunktion $F(x) = P(X \leq x)$ beschreiben. Hat man *zwei* beliebige Zufallsvariablen X und Y , so lässt sich die gemeinsame Verteilung beider Variablen durch deren **gemeinsame Verteilungsfunktion**

Gemeinsame
Verteilung zweier
Zufallsvariablen

$$F(x; y) := P(X \leq x; Y \leq y) \quad (14.1)$$

charakterisieren. Sind $F_X(x) = P(X \leq x)$ und $F_Y(y) = P(Y \leq y)$ die Verteilungsfunktion von X bzw. Y , so nennt man X und Y **unabhängig** oder auch **stochastisch unabhängig**, wenn sich deren gemeinsame Verteilungsfunktion $F(x; y)$ analog zu (11.16) für alle Elemente der Trägermengen von X und Y als Produkt

$$F(x; y) = F_X(X \leq x) \cdot F_Y(Y \leq y) \quad (14.2)$$

der Verteilungsfunktionen $F_X(x)$ und $F_Y(y)$ der beiden Einzelvariablen darstellen lässt.

Beispiel 14.1: Unabhängige und abhängige Zufallsvariablen

Wenn man einen Würfel n -mal wirft, so kann man jeden Wurf durch eine Zufallsvariable X_i modellieren ($i = 1, 2, \dots, n$), wobei diese Variablen bei Verwendung eines „fairen“ Würfels diskret gleichverteilt sind mit gleichen Eintrittswahrscheinlichkeiten $p = \frac{1}{6}$. Die Zufallsvariablen X_i sind hier unabhängig. Das n -malige Würfeln mit einem Würfel entspricht in der Terminologie des Urnenmodells dem n -maligen Ziehen einer Kugel aus einer Urne mit 6 nummerierten Kugeln, wobei die Ziehung jeweils *mit Zurücklegen* erfolgt.

Wirft man zweimal und verwendet die Bezeichnungen X und Y anstelle von X_1 und X_2 , so ist die Wahrscheinlichkeit $F(2; 3)$ dafür, dass der erste Wurf eine Augenzahl X bis höchstens 2 und der zweite Wurf eine Augenzahl Y bis höchstens 3 erzielt, durch das Produkt der beiden Einzelwahrscheinlichkeiten $F_X(2) = P(X \leq 2) = \frac{1}{3}$ und $F_Y(3) = P(Y \leq 3) = \frac{1}{2}$ gegeben, d. h. durch den Wert $\frac{1}{6}$. Dieses Ergebnis erhält man auch anhand kombinatorischer Überlegungen – von den 36 möglichen Augenzahl-Paaren genügen genau 6 Paare gleichzeitig den genannten Obergrenzen für die Augenzahlen X und Y .

Zieht man aus einer Urne mit nummerierten Kugeln n -mal jeweils eine Kugel *ohne Zurücklegen* und modelliert die einzelnen Ziehungen anhand von Zufallsvariablen X_i , so sind diese Zufallsvariablen nicht mehr stochastisch unabhängig. Die Ziehung der Lottozahlen ist ein Beispiel für ein solches Experiment.

Exkurs 14.1: Charakterisierung bivariater Verteilungen

Neben der Verteilungsfunktion $F(x; y)$ lässt sich zur Charakterisierung der gemeinsamen Verteilung zweier Zufallsvariablen X und Y auch – wie bei univariaten theoretischen Verteilungen – die Wahrscheinlichkeitsfunktion (diskreter Fall) resp. die Dichtefunktion (stetiger Fall) heranziehen.

Hat man zwei *diskrete* Zufallsvariablen X und Y mit der Trägermenge x_1, \dots, x_k resp. y_1, \dots, y_l und bezeichnet $p_{ij} := P(X = x_i; Y = y_j)$ die Eintrittswahrscheinlichkeit für die Realisation $(x_i; y_j)$, so lautet das bivariate Analogon zur

Wahrscheinlichkeitsfunktion (12.1)

$$f(x; y) = \begin{cases} p_{ij} & \text{für } (x; y) = (x_i; y_j); \quad i = 1, 2, \dots, k; j = 1, 2, \dots, l; \\ 0 & \text{für alle anderen } (x; y). \end{cases}$$

Diese bivariate Wahrscheinlichkeitsfunktion heißt *gemeinsame Wahrscheinlichkeitsfunktion* von X und Y . Deren Werte lassen sich in *Kontingenztafeln für Wahrscheinlichkeiten* darstellen und aus diesen kann man – genau wie bei den bivariaten empirischen Verteilungen – *Randverteilungen* und *bedingte Wahrscheinlichkeiten* ableiten.

Liegen hingegen zwei *stetige* Zufallsvariablen X und Y vor, so lässt sich die gemeinsame Verteilung beider Variablen durch die Dichtefunktion $f(x; y)$ charakterisieren. Deren Werte sind stets nicht-negativ. Die Dichtefunktion $f(x; y)$ ist analog zu (13.2) dadurch definiert, dass sie die Eigenschaft hat, dass sich jeder Wert $F(x; y)$ der Verteilungsfunktion aus (13.1) durch Integration der Dichte bis zur Stelle $(x; y)$ ergibt:

$$F(x; y) = \int_{-\infty}^x \int_{-\infty}^y f(s; t) ds dt \quad \text{für alle reellwertigen Paare } (x; y).$$

Auch bei bivariaten stetigen Verteilungen kann man *Randverteilungen* einer Variablen betrachten, die sich bei Vernachlässigung der jeweils anderen Variablen ergeben, und *bedingte Dichtefunktionen* bestimmen. Randdichten sind die Dichten der Einzelpunkten und bedingte Dichten resultieren – analog zu (9.7) oder (9.8) bei bivariaten empirischen Verteilungen – nach Division der gemeinsamen Dichtefunktion $f(x; y)$ durch eine der beiden Randdichten.

Eine detailliertere Darstellung dieser hier nur angerissenen Begriffe findet man z. B. bei FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Kapitel 8) oder TOUTENBURG / HEUMANN (2008, Abschnitt 3.7).

Der Begriff der Unabhängigkeit spielt eine zentrale Rolle beim Schätzen von Modellparametern und beim Testen von Hypothesen. zieht man aus einer Grundgesamtheit eine n -elementige Stichprobe, so wird diese in der schließenden Statistik durch Zufallsvariablen X_1, X_2, \dots, X_n modelliert, für die man Realisationen x_1, x_2, \dots, x_n beobachtet und verwertet. Die Zufallsvariablen X_1, X_2, \dots, X_n werden meist nicht direkt herangezogen, sondern anhand einer **Stichprobenfunktion** aggregiert:

$$X_1, X_2, \dots, X_n \xrightarrow{\text{Verdichtung der Stichprobeninformation}} g(X_1, X_2, \dots, X_n)$$

Wichtige Stichprobenfunktionen

Wenn eine Stichprobenfunktion im Kontext der Schätzung verwendet wird, spricht man sie auch als **Schätzfunktion** an, beim Testen als **Testfunktion** oder **Prüfstatistik**.

Eine wichtige Stichprobenfunktion ist der **Stichprobenmittelwert**

$$\bar{X} := \frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n) = \frac{1}{n} \cdot \sum_{i=1}^n X_i, \quad (14.3)$$

der auf der Datenebene seine Entsprechung in (5.2) findet. Eine weitere Stichprobenfunktion, die beim Schätzen und Testen oft gebraucht wird, ist die **mittlere quadratische Abweichung**

$$S^2 := \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2 \quad (14.4)$$

bzw. die **Stichprobenvarianz**

$$S^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} \cdot S^2, \quad (14.5)$$

die in (5.6) und (5.9) ihre empirischen Entsprechungen haben.¹ Etwas komplexere Stichprobenfunktionen, bei denen noch spezielle Verteilungsannahmen ins Spiel kommen und zur Definition von χ^2 -, t - und F -Verteilung führen, wurden bereits in (13.27), (13.28) resp. (13.30) vorgestellt.

Verteilung des Stichprobenmittelwerts Wenn die Stichprobenvariablen X_1, X_2, \dots, X_n alle unabhängig $N(\mu; \sigma^2)$ -verteilt sind, so kann man auch für die Stichprobenfunktionen (14.3) und (14.5) Verteilungsaussagen ableiten, die u. a. beim Testen von Hypothesen eine wichtige Rolle spielen. Überträgt man (13.17) auf die Summation von n normalverteilten Zufallsvariablen ($n \geq 2$) mit gleichem Erwartungswert μ und gleicher Varianz σ^2 , so folgt zunächst für die Summe der n Stichprobenvariablen, dass ihr Erwartungswert durch $n \cdot \mu$ und ihre Varianz durch $n \cdot \sigma^2$ gegeben ist. Für den Stichprobenmittelwert \bar{X} verifiziert man dann mit (13.16), wenn man dort $a = \frac{1}{n}$ und $b = 0$ einsetzt, dass er normalverteilt ist mit Erwartungswert $E(\bar{X}) = \mu$ und Varianz $V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$, also²

$$\bar{X} \sim N(\mu; \sigma_{\bar{X}}^2) \quad \text{mit} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}. \quad (14.6)$$

Standardisiert man den Stichprobenmittelwert gemäß (13.11), folgt

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} \sim N(0; 1). \quad (14.7)$$

Für die aus n unabhängigen $N(\mu; \sigma^2)$ -verteilten Stichprobenvariablen X_i gebildete Stichprobenvarianz lässt sich eine Beziehung zur χ^2 -Verteilung ableiten. Standardisiert man die Variablen X_i , so ist die Summe der

¹Beim Schätzen und Testen wird vor allem die Stichprobenvarianz (14.5) verwendet, die im Vergleich zu (14.4) günstigere Schätzegenschaften hat.

²Die Formeln für den Erwartungswert und die Varianz von \bar{X} sind nicht an die Normalverteilungsannahme gebunden, wie in Abschnitt 15.2 noch gezeigt wird.

Quadrat der resultierenden Variablen $Z_i = \frac{X_i - \mu}{\sigma}$ nach (13.27) χ^2 -verteilt mit n Freiheitsgraden:

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2. \quad (14.8)$$

Hieraus kann man mit einigen Überlegungen ableiten, dass die mit dem Faktor $\frac{n}{\sigma^2}$ multiplizierte mittlere quadratische Abweichung S^2 bzw. – äquivalent – die mit $\frac{n-1}{\sigma^2}$ multiplizierte Stichprobenvarianz S^{*2} einer χ^2 -Verteilung mit $n-1$ Freiheitsgraden folgt:

$$\frac{n \cdot S^2}{\sigma^2} = \frac{(n-1) \cdot S^{*2}}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2. \quad (14.9)$$

Ferner lässt sich mit (14.9) zeigen, dass eine Ersetzung von σ in (14.7) durch die **Stichprobenstandardabweichung** $S^* := \sqrt{S^{*2}}$ zu einer t -Verteilung mit $n-1$ Freiheitsgraden führt:

$$\frac{\bar{X} - \mu}{S} \cdot \sqrt{n-1} = \frac{\bar{X} - \mu}{S^*} \cdot \sqrt{n} \sim t_{n-1}. \quad (14.10)$$

Auf einen Beweis der beiden letzten Verteilungsaussagen, die beide auf der Voraussetzung unabhängiger und normalverteilter Stichprobenvariablen beruhen und beim Schätzen und Testen vielfach gebraucht werden, sei hier verzichtet. Man findet eine Herleitung von (14.10) z. B. bei MOSLER / SCHMID (2011, Abschnitt 4.3.1).

Exkurs 14.2: Der Zentrale Grenzwertsatz

Aussage (14.6) bezieht sich auf die Verteilung eines Stichprobenmittelwerts \bar{X} , der aus n unabhängigen, mit gleichem Erwartungswert μ und gleicher Varianz σ^2 normalverteilten Zufallsvariablen X_1, X_2, \dots, X_n gebildet ist, während (14.7) eine Verteilungsaussage für den aus \bar{X} abgeleiteten standardisierten Stichprobenmittelwert liefert. Ein direkt an diese Aussagen anknüpfender bedeutender Satz der Wahrscheinlichkeitsrechnung ist der **Zentrale Grenzwertsatz**. Er beinhaltet, dass die beiden genannten Aussagen für große Werte von n immerhin noch näherungsweise gültig bleiben, wenn die Variablen X_1, X_2, \dots, X_n zwar unabhängig sind und bezüglich Erwartungswert und Varianz übereinstimmen, aber nicht mehr normalverteilt sind.

Seien also X_1, X_2, \dots, X_n unabhängige Zufallsvariablen mit gleichem Erwartungswert μ und gleicher Varianz σ^2 . Die Summe $Y_n := \sum_{i=1}^n X_i$ der n Zufallsvariablen hat dann den Erwartungswert $n \cdot \mu$ und die Varianz $n \cdot \sigma^2$. Wenn man zur genaueren Kennzeichnung des Stichprobenmittelwerts \bar{X} hier noch einen Index anträgt, also die Bezeichnung \bar{X}_n verwendet, so sagt der Zentrale

Verteilung der
Stichprobenvarianz

Grenzwertsatz, dass die Verteilungsfunktion der standardisierten Summe

$$Z_n := \frac{Y_n - E(Y_n)}{\sqrt{V(Y_n)}} = \sum_{i=1}^n \frac{X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}}$$

unter diesen Voraussetzungen für $n \rightarrow \infty$ gegen die Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung konvergiert. Hieraus lässt sich folgern, dass bei großen Werten n näherungsweise

$$Z_n \sim N(0; 1)$$

gilt und dass die Summe Y_n der Zufallsvariablen X_1, X_2, \dots, X_n approximativ $N(n \cdot \mu; n \cdot \sigma^2)$ -verteilt ist.



Interaktives Objekt
„Approximation der
Binomialverteilung“

Wählt man speziell die n Variablen X_1, X_2, \dots, X_n wie in (12.19), also als identisch bernoulli-verteilt, so folgt aus den vorausgegangenen Ausführungen und bei Beachtung von (12.15) und (12.16), dass die binomialverteilte Zählvariable $Y_n := X$ aus (12.19) bei großem n approximativ $N(n \cdot p; n \cdot p(1-p))$ -verteilt ist. Die Verteilungsfunktion $F(x) = P(X \leq x)$ einer $B(n; p)$ -verteilten Zufallsvariablen X kann demnach für große n durch die Verteilungsfunktion $F(x)$ einer $N(n \cdot p; n \cdot p(1-p))$ -verteilten Zufallsvariable approximiert werden. Für den Wert $F(a)$ der Verteilungsfunktion $F(x)$ einer $B(n; p)$ -verteilten Zufallsvariablen an der Stelle $x = a$ gilt also, wie man durch Standardisierung der $N(n \cdot p; n \cdot p(1-p))$ -Verteilung verifiziert, dass $F(a) = \Phi(a^*)$ gilt mit $a^* = \frac{a - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$. Hinreichende Approximationsgüte wird in der Praxis meist als gegeben angesehen, wenn die Bedingungen $n \cdot p \geq 5$ und $n \cdot (1-p) \geq 5$ erfüllt sind.

14.2 Kovarianz und Korrelation

In den Abschnitten 12.2 und 13.2 wurden univariante Wahrscheinlichkeitsverteilungen von Zufallsvariablen anhand von Kenngrößen charakterisiert. Als Lageparameter für die Verteilung von X wurde hier der durch (12.6) resp. (13.9) definierte Erwartungswert $\mu = E(X)$ aufgeführt und als Streuungsparameter die Varianz $V(X) = \sigma^2 = E[(X - \mu)^2]$ aus (12.8) oder die Standardabweichung $\sigma = \sqrt{V(X)}$ aus (13.10).

Ein nicht-normiertes
Zusammenhangsmaß

Hat man *zwei* Zufallsvariablen X und Y mit Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$ sowie Varianzen $\sigma_X^2 = V(X)$ und $\sigma_Y^2 = V(Y)$, so ist man auch daran interessiert, einen möglichen Zusammenhang zwischen den Verteilungen der beiden Zufallsvariablen zu quantifizieren. Ein nicht-normiertes Maß für einen linearen Zusammenhang ist die mit $Cov(X; Y)$ abgekürzte **Kovarianz** von X und Y , die zwecks Unterscheidung von der empirischen Kovarianz (10.9) auch **theoretische Kovarianz** genannt

wird. Sie ist definiert als Erwartungswert von $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X; Y) := E[(X - E(X))(Y - E(Y))]. \quad (14.11)$$

Durch Ausmultiplizieren der beiden Differenzterme $X - \mu_X$ und $Y - \mu_Y$ und anschließende gliedweise Anwendung des Erwartungswertoperators gewinnt man aus (14.11) noch die äquivalente Darstellung

$$\text{Cov}(X; Y) = E(XY) - E(X) \cdot E(Y). \quad (14.12)$$

Ähnlich wie bei der empirischen Kovarianz gilt auch bei der theoretischen Kovarianz, dass sie positiv ist, wenn X und Y eine gleichgerichtete Tendenz haben und negativ bei gegenläufiger Tendenz. Im Falle $\text{Cov}(X; Y) = 0$ kann nicht von einem *linearen* Zusammenhang zwischen den Zufallsvariablen X und Y ausgegangen werden. Wenn X und Y unabhängig sind, hat ihre Kovarianz stets den Wert 0, d. h. es gilt

$$X \text{ und } Y \text{ sind unabhängig} \Rightarrow \text{Cov}(X; Y) = 0. \quad (14.13)$$

Sind X und Y zwei Zufallsvariablen mit der Kovarianz $\text{Cov}(X; Y)$, so gilt für die Varianz ihrer Summe

$$V(X + Y) = V(X) + V(Y) + 2 \cdot \text{Cov}(X; Y). \quad (14.14)$$

Wie die empirische Kovarianz ist auch die theoretische Kovarianz maßstabsabhängig. Sie hat daher keine untere oder obere Schranke. Eine zur Definition (10.10) des empirischen Korrelationskoeffizienten r analoge Normierung wird erreicht, wenn man die Kovarianz durch das Produkt der Standardabweichungen σ_X und σ_Y dividiert. Dies führt zum Korrelationskoeffizienten ρ (lies: *rho*) für die Zufallsvariablen X und Y :³

$$\rho = \frac{\text{Cov}(X; Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}. \quad (14.15)$$

Ein normiertes
Zusammenhangsmaß

Der **Korrelationskoeffizient** ρ liegt wie sein empirisches Analogon r stets zwischen -1 und $+1$, d. h. es gilt

$$-1 \leq \rho \leq 1. \quad (14.16)$$

Es gilt $|\rho| = 1$ (lies: *rho-Betrag* = 1) genau dann, wenn die beiden Zufallsvariablen X und Y linear abhängig sind, etwa $Y = aX + b$. Dabei wird die obere Schranke $\rho = 1$ im Falle $a > 0$ angenommen (gleichsinnige Tendenz von X und Y) und die untere Schranke $\rho = -1$ für $a < 0$ (gegensinnige Tendenz). Im Falle $\rho = 0$ spricht man von **Unkorreliertheit**, im Falle

³Man verwendet anstelle von ρ auch die Schreibweise $\rho(X; Y)$ oder ρ_{XY} , wenn man betonen will, dass es um ein Zusammenhangsmaß für X und Y geht.

$\rho \neq 0$ von **Korreliertheit** der Variablen X und Y . Aus (14.13) folgt, dass Unabhängigkeit von X und Y stets Unkorreliertheit impliziert:

$$X \text{ und } Y \text{ sind unabhängig} \Rightarrow \rho = 0. \quad (14.17)$$

Der Umkehrschluss gilt nicht, d.h. unkorrelierte Zufallsvariablen sind nicht zwingend auch stochastisch unabhängig.

Beispiel 14.2: Berechnung des Korrelationskoeffizienten

Beim dreimaligen Werfen einer Münze könnte man die Anzahl der Ausgänge mit „Zahl“ durch eine Zufallsvariable X und die der Ausgänge mit „Kopf“ durch eine Zufallsvariable Y modellieren. Für den Korrelationskoeffizienten dieser beiden Zufallsvariablen gilt $\rho = -1$, d. h. X und Y sind maximal negativ korreliert. Man kann dieses Ergebnis bei diesem einfachen Illustrationsbeispiel auch ohne Rückgriff auf (14.15) leicht verifizieren. Würde man das Experiment durchführen, so wären hier für $(X; Y)$ nur vier Realisationen $(x; y)$ möglich, nämlich $(0; 3), (1; 2), (2; 1), (3; 0)$, die alle auf einer fallenden Geraden liegen. Der Wert $\rho = -1$ leitet sich hier aus dem Modellzusammenhang ab und nicht – wie bei der Berechnung des empirischen Korrelationskoeffizienten r nach Bravais-Pearson – aus Daten.



Aufgabe 14.1

Exkurs 14.3: Minimum-Varianz-Portfolio nach Markowitz



Harry M.
MARKOWITZ

Varianz, Kovarianz und Korrelation von Zufallsvariablen spielen auch in der **Portfoliotheorie** eine zentrale Rolle. Diese schon 1952 entwickelte Theorie geht auf Harry M. MARKOWITZ (geb. 1927) zurück, der für seine bahnbrechenden Arbeiten 1990 mit dem Wirtschaftsnobelpreis ausgezeichnet wurde. Die Portfoliotheorie beweist, dass sich das Anlegerrisiko für ein Wertpapierportfolio minimieren lässt, wenn die Korrelation zwischen den Risiken der einzelnen Anlagen berücksichtigt wird. Es wird gezeigt, dass Risikoreduktion durch Diversifikation erreichbar ist, wobei die besten Ergebnisse bei fehlender oder geringer Korrelation der Depotkomponenten erzielt werden.

Die Grundidee des Theoriegebäudes lässt sich anhand eines aus nur zwei Anlagen A und B bestehenden Portfolios P verdeutlichen. Die Anlagen – z. B. zwei Aktien oder eine Aktie und Gold – lassen sich durch Renditen R_A und R_B charakterisieren. Dies sind Zufallsvariablen mit Erwartungswert $\mu_A = E(R_A)$ bzw. $\mu_B = E(R_B)$ und Varianz $\sigma_A^2 = V(R_A)$ bzw. $\sigma_B^2 = V(R_B)$. Die Varianzen σ_A^2 und σ_B^2 resp. die Standardabweichungen σ_A und σ_B spiegeln die Volatilität von A und B wider, repräsentieren somit die mit diesen Anlagen verbundenen Risiken. Ziel des Anlegers ist eine möglichst hohe Rendite für das Portfolio bei möglichst geringem Risiko.

Besteht das Portfolio zu einem Anteil x_A aus Anlage A und zu einem Anteil $x_B = 1 - x_A$ aus Anlage B , ist der Erwartungswert μ_P der Portfolio-Rendite wegen (12.11) und (12.13) gegeben durch

$$\mu_P = x_A \cdot \mu_A + x_B \cdot \mu_B \quad \text{mit} \quad x_B = 1 - x_A.$$

Für die Standardabweichung σ_P erhält man mit (14.14) und (12.12)

$$\sigma_P = \sqrt{(x_A \cdot \sigma_A)^2 + (x_B \cdot \sigma_B)^2 + 2 \cdot (x_A \cdot x_B) \cdot \text{Cov}_{AB}},$$

wenn Cov_{AB} die Kovarianz der Zufallsvariablen R_A und R_B bezeichnet. Das unnormierte Zusammenhangsmaß Cov_{AB} ist nach (14.15) durch $\sigma_A \cdot \sigma_B \cdot \rho_{AB}$ gegeben, wobei ρ_{AB} der Korrelationskoeffizient der Renditen der Portfoliokomponenten ist.

Wenn man für alle möglichen Mischungsverhältnisse x_A/x_B der Anlagen A und B die Werte (μ_P, σ_P) grafisch darstellt und μ_P auf der Abszissenachse, σ_P auf der Ordinatenachse abträgt, erhält man eine Kurve, die die Punkte (μ_A, σ_A) und (μ_B, σ_B) verbindet. Die Gestalt der Kurve hängt von der Kovarianz Cov_{AB} ab. Die beiden Punkte (μ_A, σ_A) und (μ_B, σ_B) resultieren, wenn das Portfolio im Grenzfall nur aus der Anlage A bzw. nur aus B bestünde. Gesucht wird der Punkt MVP auf der Kurve, für den das Risiko σ_P ein Minimum annimmt (MVP = Minimum-Varianz-Portfolio). Man kann mit Hilfe der Differentialrechnung zeigen, dass der Abszissenwert μ_{MVP} des Punktes MVP bei der Mischung $x_A/x_B = x_A/(1 - x_A)$ erreicht wird, für die

$$x_A = \frac{\sigma_B^2 - \text{Cov}_{AB}}{\sigma_A^2 + \sigma_B^2 - 2 \cdot \text{Cov}_{AB}}$$

gilt. Auf einen Beweis dieser Formel sei hier verzichtet. In der Praxis sind die Erwartungswerte, Varianzen und Kovarianzen bzw. Korrelationskoeffizienten von Renditen aus historischen Daten zu schätzen.

Ein Beispiel für ein aus zwei Anlagen A und B bestehendes Portfolio ist eine Kombination von Gold (Anlage A) mit Amazon-Aktien (Anlage B). Der Korrelationskoeffizient der Renditen R_A und R_B ist hier relativ klein, weil Gold ein Rohstoff ist, der in Krisenzeiten – anders als Unternehmensanteile – stark gefragt ist. Mit Daten der Jahre 2000 – 2019 errechneten sich für die Renditen μ_A und μ_B die Schätzwerte 0,03039 bzw. 0,30002, für die Varianzen σ_A^2 und σ_B^2 die Werte 0,06978 und 0,17819 sowie für die Kovarianz $-0,000819$ bzw. $-0,00735$ für den Korrelationskoeffizienten. Mit diesen Werten erhält man für den varianzminimierenden Anteil x_A von Gold

$$x_A = \frac{0,171819 + 0,000819}{0,06978 + 0,17819 + 2 \cdot 0,000819} \approx 0,717.$$

Das Portfolio mit minimaler Varianz besteht demnach zu 71,7 % aus Gold und zu 28,3 % aus Amazon-Aktien. Der Erwartungswert der Rendite μ_{MVP} dieses Portfolios errechnet sich zu

$$\mu_{MVP} = 0,717 \cdot 0,03039 \mu_A + 0,283 \cdot 0,30002 \approx 0,107.$$

Der Fall eines Portfolios mit nur zwei Komponenten

Bestimmung der risikominimierenden Anlagenmischung

Beispiel eines Zwei-Komponenten-Portfolios

Abbildung 14.1 zeigt die Erwartungswerte μ_P und Standardabweichungen σ_P für alle denkbaren Portfoliokombinationen in Form einer Kurve, die im Punkt (μ_A, σ_A) beginnt und in (μ_B, σ_B) endet. Der Punkt (μ_A, σ_A) repräsentiert den Grenzfall, bei dem das gesamte Portfolio nur aus Gold besteht, der Punkt (μ_B, σ_B) den Fall eines nur Amazon-Aktien enthaltenden Depots. Der Punkt MVP kennzeichnet die Portfoliozusammensetzung, bei der die Varianz σ_P^2 bzw. die Standardabweichung σ_P ein Minimum erreicht. Auf der Abszissenachse sind die Erwartungswerte μ_A und μ_B der Renditen für Gold und Amazon-Aktien markiert sowie die Rendite μ_{MVP} des varianzminimierenden Portfolios. Auf der Ordinatenachse sind die durch die Standardabweichungen σ_A und σ_B repräsentierten Risiken für Gold und Amazon-Aktien betont.

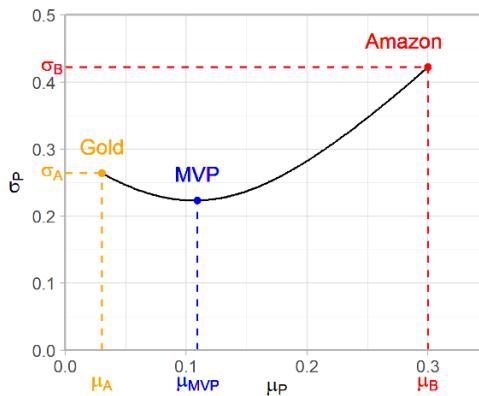


Abb. 14.1: Risiko σ_P eines aus Gold und Amazon-Aktien bestehenden Portfolios in Abhängigkeit vom Erwartungswert μ_p der Rendite

Man erkennt, dass durch die Kombination von Gold und Amazon-Aktien die Einzelrisiken σ_A und σ_B gesenkt werden können. Jede Portfoliokombination auf der Kurve ist möglich. Die Kombinationen auf dem linken Kurvenast sind aber ineffizient, weil sie alle schlechter sind (kleinere Werte für μ_P bei gleichzeitig höheren Risiken σ_P) als die durch den Punkt MVP repräsentierte Portfoliozusammensetzung. Jeder Punkt rechts von MVP beinhaltet ein größeres Risiko σ_P , dafür aber einen größeren Erwartungswert μ_P für die Rendite des Portfolios. Ob MVP oder ein weiter rechts liegender Punkt auf der Kurve gewählt wird, hängt von der Risikoeinstellung eines Anlegers ab.

Es sei erwähnt, dass das beschriebene Portfolio-Modell nicht alle Entscheidungsgrößen berücksichtigt, die für einen Anleger relevant sein können – etwa Transaktionskosten oder Lagerkosten für Gold.



15 Schätzung von Parametern



Vorschau auf
das Kapitel

Vorgestellt wird zunächst das Konzept der *Punktschätzung*. Bei dieser wird eine Stichprobenfunktion herangezogen, um einen unbekannten Parameterwert möglichst genau zu treffen. Da die Stichprobenfunktion als Zufallsvariable modelliert wird, bestimmt die Verteilung dieser Zufallsvariablen die Güte der Schätzung. Die Verteilung der zur Schätzung verwendeten Stichprobenfunktion lässt sich wiederum durch den Erwartungswert und die Varianz charakterisieren. Sowohl die Varianz als auch die als Verzerrung bezeichnete Abweichung zwischen dem Erwartungswert und dem zu schätzenden Parameter sollen möglichst klein sein. Der mittlere quadratische Fehler ist ein Gütemaß, das Verzerrung und Varianz einer Schätzfunktion verknüpft.

Als Stichprobenfunktionen werden für Punktschätzungen häufig der Stichprobenmittelwert und die Stichprobenvarianz herangezogen. Für die Verteilung beider Stichprobenfunktionen werden unter der Voraussetzung unabhängig normalverteilter Stichprobenvariablen Kenngrößen abgeleitet, insbesondere der Erwartungswert.

Als Alternative zur Punktschätzung wird abschließend die *Intervallschätzung* erläutert. Bei dieser berechnet man ein Intervall, das den zu schätzenden Parameter mit einer Wahrscheinlichkeit $1 - \alpha$ überdeckt (α klein). Das Prinzip wird anhand der Schätzung von Anteilswerten illustriert.

In Abschnitt 3.2 wurde bereits die Ziehung von Stichproben im Kontext der beschreibenden Statistik behandelt. Es wurde dargelegt, dass man anhand von Stichprobendaten Aussagen für Merkmale in einer umfassenderen Grundgesamtheit ableiten will. Wie man diesen Brückenschlag von der Stichprobe zur Grundgesamtheit bewerkstelligen kann, wird erst jetzt – im Rahmen der schließenden Statistik – verständlich. Um von der Stichprobeninformation auf die Grundgesamtheit zu schließen, verwendet man i. d. R. Verteilungsmodelle, die das Verhalten eines zu untersuchenden Merkmals X in der Grundgesamtheit charakterisieren. Von Interesse ist dann ein hier ganz allgemein mit θ (lies: *theta*) bezeichneter unbekannter Parameter der Verteilung von X . Dieser Parameter kann z. B. der Erwartungswert μ von X sein, die Varianz σ^2 von X oder ein Anteilswert p . Die Stichprobeninformation wird zu einer Stichprobenfunktion aggregiert und deren Ausprägung für die Schätzung des unbekannten Parameters herangezogen. Da die Informationsbasis bei Verwendung von Stichproben schmäler ist als bei Erfassung der Merkmalsausprägungen aller Elemente der Grundgesamtheit, sind die aus Stichproben abgeleiteten Schlüsse natürlich nicht fehlerfrei. Bei zufälliger Auswahl der Stichprobenelemente kann man Fehlerwahrscheinlichkeiten aber unter Kontrolle halten. Es leuchtet ein, dass größere Stichproben mehr Informationen liefern und

die aus ihnen abgeleiteten Schlüsse tendenziell zuverlässiger sind als bei kleinen Stichproben.

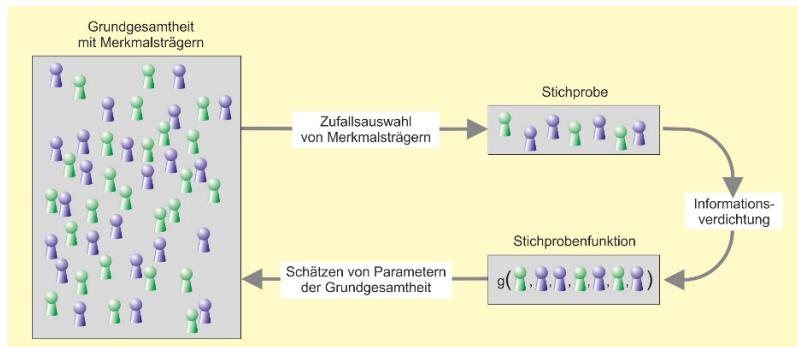


Abb. 15.1: Vorgehensweise bei der Schätzung

Punkt- und Intervallschätzung Wenn man für ein stochastisches Merkmal X ein geeignetes Verteilungsmodell spezifiziert hat, kommen für die Schätzung des interessierenden Parameters der Verteilung zwei Ansätze in Betracht, nämlich die Punkt- und die Intervallschätzung. Mit einer **Punktschätzung** will man einen unbekannten Parameter möglichst gut treffen, während eine **Intervallschätzung** einen als **Konfidenzintervall** bezeichneten Bereich festlegt, in dem der Parameter mit einer Wahrscheinlichkeit von mindestens $1 - \alpha$ liegt, wobei die Irrtumswahrscheinlichkeit α vorgegeben ist.

Beispiel 15.1: Schätzprobleme in der Praxis

Nachstehend ist eine kleine Auswahl von Schätzproblemen genannt. Es ließen sich leicht weitere Beispiele auflisten.

- In der *Klimaforschung* verwendet man Zeitreihen für Schadstoffkonzentrationen in der Atmosphäre, um Veränderungen der Luftverschmutzung abzuschätzen und daraus empirisch fundierte Prognosen abzuleiten.
- In der *Marktforschung* verwertet man u. a. Stichprobendaten zum Fernsehverhalten von Zuschauern. Aus diesen will man z. B. die Verweildauer bei Werbeblöcken für verschiedene Altersgruppen schätzen.
- Bei den ersten *Hochrechnungen bei Bundestagswahlen* geht es darum, auf der Basis einzelner Wahlkreise eine Schätzung des Anteils von Wählern zu erhalten, die eine bestimmte Partei gewählt haben.
- Bei der *industriellen Serienfertigung* schätzt man auf der Basis von Stichproben die mittlere Ausprägung von Qualitätsmerkmalen.
- In der *Medizin* versucht man, den Anteil der Personen, die gegenüber einer Viruserkrankung immun sind, anhand von Antikörpertests zu schätzen. Für die Schätzung geeignet wäre eine Zufallsstichprobe von Laborbefunden, die für eine größere Population repräsentativ ist.

15.1 Punktschätzungen und ihre Eigenschaften

Will man für einen Parameter θ (lies: *theta*) der Verteilung eines Merkmals in einer Grundgesamtheit eine Punktschätzung anhand von Stichprobendaten x_1, x_2, \dots, x_n gewinnen, verwendet man die Realisation einer **Stichprobenfunktion** $g(x_1, x_2, \dots, x_n)$ als Schätzwert. Da die Stichprobendaten als Ausprägungen von Zufallsvariablen X_1, X_2, \dots, X_n interpretiert werden, ist auch der aus ihnen errechnete Schätzwert eine Realisation einer Zufallsvariablen $g(X_1, X_2, \dots, X_n)$, die hier **Schätzstatistik**, **Schätzfunktion** oder kurz **Schätzer** genannt wird. Im Folgenden wird, wenn von der Schätzung eines nicht näher spezifizierten Parameters θ die Rede ist, bei der Notation nicht zwischen dem Schätzer und dem Schätzwert unterschieden; beide werden mit $\hat{\theta}$ angesprochen (lies: *theta-Dach*). Die Verwendung von $\hat{\cdot}$ über einer Kenngröße ist in der Statistik für die Kennzeichnung von Schätzungen üblich.

Bevor Schätzer $\hat{\theta}$ für Kenngrößen θ vorgestellt werden, ist zu klären, was eine „gute“ Schätzung ausmacht. Ein einleuchtendes Gütekriterium ist die **Erwartungstreue** oder **Unverzerrtheit**. Diese beinhaltet, dass der Schätzer „im Mittel“ den zu schätzenden Wert θ genau trifft, d. h.¹

$$E(\hat{\theta}) = \theta. \quad (15.1)$$

Wenn ein Schätzer $\hat{\theta}$ nicht erwartungstreu ist, heißt die Differenz

$$B(\hat{\theta}) := E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta) \quad (15.2)$$

Verzerrung oder **Bias** (engl.: *bias*). Ein Schätzer für θ ist also genau dann erwartungstreu, wenn seine Verzerrung Null ist. Manchmal ist ein Schätzer $\hat{\theta}$ zwar verzerrt, besitzt aber eine Verzerrung, die gegen Null strebt, wenn der Umfang n des zur Berechnung von $\hat{\theta}$ verwendeten Datensatzes gegen ∞ (lies: *unendlich*) strebt:

$$\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta. \quad (15.3)$$

Ein Schätzer $\hat{\theta}$ mit dieser Eigenschaft heißt **asymptotisch erwartungstreu** oder **asymptotisch unverzerrt**.

Neben der Erwartungstreue ist die anhand der **Varianz** oder der **Standardabweichung** ausgedrückte Variabilität einer Schätzung als Präzisionsmaß von Interesse. Die auch als **Standardfehler** (engl: *standard error*) bezeichnete Standardabweichung einer Schätzfunktion wird von



Video
„Punktschätzung“

Gütekriterien für
Schätzfunktionen:

- keine oder geringe
Verzerrung

- kleine Varianz

¹Der Erwartungswert $E(\hat{\theta})$ wird unter der Annahme bestimmt, dass der zu schätzende unbekannte Parameter den Wert θ hat. Gelegentlich wird dies durch die Notation $E_\theta(\hat{\theta})$ betont (analoge Indizierung für andere Schätzercharakteristika). Erwartungstreue beinhaltet, dass (15.1) für alle möglichen Werte von θ gilt.

Statistiksoftwarepaketen bei der Anwendung von Schätzprozeduren routinemäßig neben den Schätzwerten ausgewiesen (vgl. Abbildung 17.4).



Video „Beurteilung von Schätzern“

Abbildung 15.2 zeigt die – hier als symmetrisch angenommenen – Dichtefunktionen für drei Schätzfunktionen, wobei die ersten beiden Schätzer, etwa $\hat{\theta}_1$ und $\hat{\theta}_2$, den Erwartungswert $E(\hat{\theta}_1) = E(\hat{\theta}_2) = a$ und der dritte Schätzer $\hat{\theta}_3$ den Erwartungswert $E(\hat{\theta}_3) = b$ habe. Geht man davon aus, dass der zu schätzende unbekannte Parameter θ den Wert $\theta = b$ hat, so ist $\hat{\theta}_3$ erwartungstreu, während $\hat{\theta}_1$ und $\hat{\theta}_2$ verzerrte Schätzer sind. Allerdings hat $\hat{\theta}_2$ eine kleinere Varianz als $\hat{\theta}_3$. Da sich die Schätzer $\hat{\theta}_1$ und $\hat{\theta}_2$ nicht bezüglich des Erwartungswerts unterscheiden, ist von beiden der Schätzer mit der geringeren Streuung (steilere Dichtekurve) vorzuziehen, also $\hat{\theta}_2$.

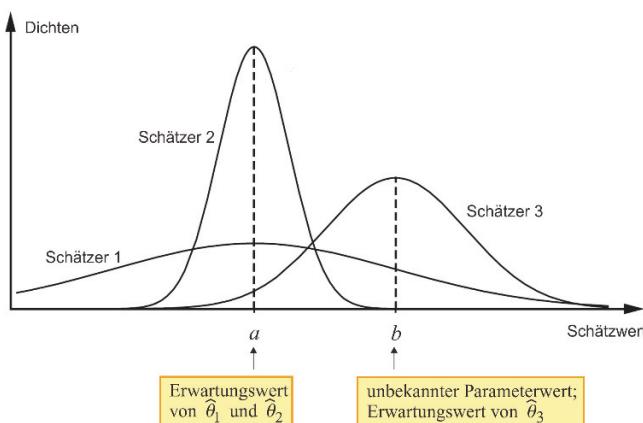


Abb. 15.2: Vergleich dreier Schätzfunktionen für $\theta = b$

- kleiner MSE In der Praxis hat man häufig verzerrte Schätzer. Wie aber soll man sich zwischen zwei Schätzern entscheiden, wenn – wie in Abbildung 15.2 anhand der Schätzer $\hat{\theta}_3$ und $\hat{\theta}_2$ illustriert – ein Schätzer bezüglich des Kriteriums „Verzerrung“ schlechter, dafür aber beim Streuungsvergleich besser abschneidet? Man benötigt noch ein Kriterium, das sowohl die Verzerrung als auch die Streuung berücksichtigt. Ein solches Gütemaß ist der mit **MSE** abgekürzte **mittlere quadratische Fehler** (engl.: *mean squared error*)

$$MSE(\hat{\theta}) := E \left[(\hat{\theta} - \theta)^2 \right]. \quad (15.4)$$

- MSE-Zerlegungsformel Nach elementaren Umformungen erhält man mit (15.2) und der Varianzdefinition (12.8) die äquivalente Darstellung

$$MSE(\hat{\theta}) = E \left[(\hat{\theta} - E(\hat{\theta}))^2 \right] + [E(\hat{\theta}) - \theta]^2 = V(\hat{\theta}) + B(\hat{\theta})^2. \quad (15.5)$$

Der MSE repräsentiert eine additive Verknüpfung von Varianz und quadrierter Verzerrung. Man wird von den beiden Schätzern $\hat{\theta}_2$ und $\hat{\theta}_3$ in Abbildung 15.2 denjenigen als „besser“ ansehen, dessen MSE kleiner ausfällt. Bei erwartungstreuen Schätzern sind MSE und Varianz identisch.

15.2 Punktschätzung von Erwartungswerten und Varianzen

Will man den Erwartungswert μ einer Zufallsvariablen anhand der Ausprägungen unabhängiger Stichprobenvariablen X_1, X_2, \dots, X_n schätzen, bietet sich als Stichprobenfunktion der in (14.3) eingeführte **Stichprobenmittelwert** \bar{X} an. Da man die Erwartungswertbildung nach (12.13) auf die Stichprobenvariablen einzeln anwenden kann, gilt

$$E(\bar{X}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \dots + E(X_n)] = \frac{1}{n} \cdot n \cdot \mu = \mu. \quad (15.6)$$

Schätzung des Erwartungswerts

Der Stichprobenmittelwert liefert also eine *unverzerrte* Schätzung für den Erwartungswert. Wenn die Stichprobenvariablen X_1, X_2, \dots, X_n unabhängig sind und die feste Varianz σ^2 haben, kann man für die Varianz $V(\bar{X}) = \sigma_{\bar{X}}^2$ von \bar{X} mit (12.12) und (12.14) die Darstellung

$$V(\bar{X}) = \frac{\sigma^2}{n} \quad (15.7)$$

ableiten, d. h. durch $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ ist der Standardfehler der Schätzfunktion \bar{X} gegeben. Wegen der Unverzerrtheit von \bar{X} stimmt $V(\bar{X})$ mit dem mittleren quadratischen Fehler $MSE(\bar{X})$ von \bar{X} überein. Die Qualität des Schätzers \bar{X} verbessert sich bei Erhöhung des Stichprobenumfangs.

Zur Schätzung der Varianz σ^2 einer Zufallsvariablen kommt zunächst die **mittlere quadratische Abweichung** S^2 aus (14.4) in Betracht. Mit elementaren Umformungen und Anwendung der Zerlegungsformel (12.9) auf die Varianz von \bar{X} kann man zeigen, dass ²

$$E(S^2) = \frac{n-1}{n} \cdot \sigma^2. \quad (15.8)$$

Schätzung der Varianz

Die mittlere quadratische Abweichung liefert also eine *verzerrte* Schätzung für σ^2 . Der nach (5.6) errechnete Schätzwert s^2 unterschätzt wegen $\frac{n-1}{n} < 1$ den wahren Wert von σ^2 , wobei die Verzerrung mit zunehmendem Stichprobenumfang n gegen Null strebt. Die Schätzfunktion S^2 ist demnach nur asymptotisch erwartungstreu. Um eine Schätzung für σ^2 zu erhalten, die nicht nur asymptotisch, sondern auch für endliches n



Aufgabe 15.1

²Die Zerlegung der Varianz von \bar{X} hat die Gestalt $\sigma_{\bar{X}}^2 = E(\bar{X}^2) - \mu^2$. Eine Herleitung von (15.8) findet man bei MOSLER / SCHMID (2011, Abschnitt 5.1.4).

unverzerrt ist, verwendet man anstelle von S^2 zur Varianzschätzung die **Stichprobenvarianz** S^{*2} aus (14.5). Für sie gilt

$$E(S^{*2}) = \frac{n}{n-1} \cdot E(S^2) = \sigma^2. \quad (15.9)$$

Beispiel 15.2: Schätzung von Prozessfähigkeit in der Industrie

Bei der industriellen *Qualitätssicherung* werden Fertigungsprozesse kontinuierlich beobachtet und dokumentiert. Ziel ist es, stabile Prozessverläufe zu gewährleisten und nicht-spezifikationskonforme Produkte zu vermeiden. Abbildung 15.3 zeigt ein Foto aus der Fertigungsüberwachung. Hier werden die Ausprägungen mehrerer qualitätsrelevanter geometrischer Merkmale (Durchmesser von Wellen) gleichzeitig erfasst und die Ergebnisse fortlaufend als Zeitreihe gespeichert. Die Ampelfarben links im Bild signalisieren den Grad der Einhaltung von Toleranzvorgaben. Bei Bedarf kommen intervenierende Maßnahmen zum Zuge, etwa das Nachjustieren einer Fertigungseinrichtung.

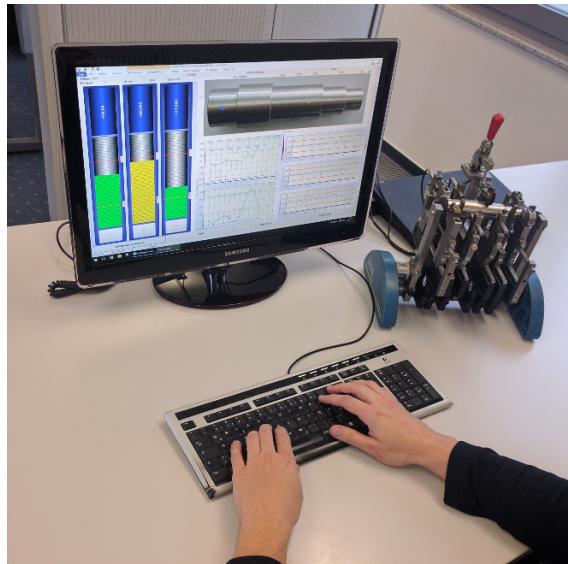


Abb. 15.3: Simultane Erfassung dreier Merkmale eines Serienteils bei der industriellen Fertigungsüberwachung (Quelle: Fa. Q-DAS)

Die Stabilität von Fertigungsprozessen und der Grad der Einhaltung von Toleranzspezifikationen für ein Qualitätsmerkmal X werden bei der industriellen Fertigung anhand sog. *Prozessfähigkeitsindizes* erfasst. Man geht i. d. R. von einem normalverteilten Merkmal aus, dessen Ausprägungen sich innerhalb eines Toleranzintervalls $[UG; OG]$ bewegen sollen. Der Zielwert für den Erwartungswert μ ist meist die Mitte M des Toleranzintervalls. Damit das Merkmal nicht zu sehr streut, wird gefordert, dass das Sechsfache der Standardabweichung σ von X weniger als 75% der Länge $OG - UG$ des Toleranzintervalls ausmacht. Dies ist äquivalent mit der Forderung, dass der Prozessfähigkeitsindex $C_p := (OG - UG)/(6 \cdot \sigma)$ die Bedingung $C_p > \frac{4}{3}$ erfüllt („C“ steht für das

englische Wort „capability“, „p“ für „process“). Da in diesen Index nur σ eingeht, erlaubt er nur eine Aussage zur Variabilität (Stabilität eines Prozesses) und keinen Rückschluss auf das Prozessniveau μ .

Ein Index, der sowohl auf unerwünschte Entwicklungen der Prozessstreuung σ als auch der Prozesslage μ reagiert, ist der Index C_{pk} (das „k“ steht für „katayori“, das japanische Wort für „Abweichung“). Dieser vergleicht den Abstand $\min(\mu - UG; OG - \mu)$ von μ zur nächstgelegenen Spezifikationsgrenze mit der halben Länge $(OG - UG)/2$ des Toleranzintervalls. Er ist definiert als $C_{pk} := \min(\mu - UG; OG - \mu)/[(OG - UG)/2]$. Auch für diesen Prozessfähigkeitsindex wird meist verlangt, dass er die Bedingung $C_{pk} > \frac{4}{3}$ erfüllt.

In der Praxis sind der Erwartungswert μ und die Varianz σ^2 von X nur in Form von Schätzungen (14.3) und (14.5) bekannt. Man erhält anstelle von C_p und C_{pk} nach Einsetzen von $\hat{\mu} = \bar{X}$ und $\hat{\sigma} = S^*$ die Zufallsvariablen

$$\hat{C}_p = \frac{OG - UG}{6 \cdot S^*} \quad \hat{C}_{pk} = \frac{\min(\bar{X} - UG; OG - \bar{X})}{0,5 \cdot (OG - UG)}.$$

In der industriellen Fertigung hat man Ausprägungen von \bar{X} und S^* , die sich aus Stichprobendaten errechnen. Damit hat man auch für die Zufallsvariablen \hat{C}_p und \hat{C}_{pk} Realisationen, die man mit dem geforderten Mindestwert $\frac{4}{3}$ vergleicht.



Die Verteilungen der zur Bewertung von Prozessfähigkeit verwendeten Schätzfunktionen \hat{C}_p und \hat{C}_{pk} und auch die weiterer Prozessfähigkeitsindizes sind bei RINNE / MITTAG (1999) detailliert beschrieben.

Exkurs 15.1: Schätzung von Geschäftsführergehältern

Selbst die beste Schätzmethode kann zu umbrauchbaren Ergebnissen führen, wenn die Daten von zweifelhafter Qualität sind. Diese Aussage lässt sich anhand eines Fallbeispiels illustrieren, das unter dem Schlagwort „Maserati-Affäre“ durch die Presse ging (Bericht im *Spiegel* vom 25. 2. 2010). Der Geschäftsführer einer in der Obdachlosenhilfe tätigen gemeinnützigen Berliner Organisation geriet in die Schlagzeilen, weil er einen Maserati als Dienstwagen fuhr. Dabei kam ans Licht, dass er ein Jahresbruttogehalt von über 400 000 Euro bezog. Der Landesverband Berlin des Paritätischen Wohlfahrtsverbands kam rasch unter öffentlichen Druck. Er ließ daher bei den Mitgliedsorganisationen die Geschäftsführergehälter per Fragebogen ermitteln.

Mit den Ergebnissen der Erhebung sah der Auftraggeber den Beweis als erbracht, dass die Affäre nur einen Einzelfall betraf. Aus den eingegangenen Fragebögen hatte sich für das Bruttojahresgehalt der nicht-ehrenamtlich tätigen Geschäftsführer ein Mittelwert von ca. 56 100 Euro und ein Median von ca. 53 300 Euro ergeben. Die Schätzwerte für Mittelwert und Median errechneten sich aus den Angaben aus 246 Fragebögen, die von den 650 angeschriebenen Trägern zurück kamen (Rücklaufquote von nur 38%). Dass der Median unterhalb des Mittelwerts liegt, ist plausibel und mit den Befunden aus Abbildung 4.8 kompatibel. Bei näherer Betrachtung des Fragebogens war festzustellen, dass

die Variable „Geschäftsführergehalt“ sehr missverständlich operationalisiert war. Allein dies sprach gegen eine ausreichende Reliabilität und Validität der Ergebnisse. Bedenklich war neben der niedrigen Rücklaufquote zudem, dass alle im Fragebogen abgefragten statistischen Informationen anonym und freiwillig geliefert wurden.

Was als „angemessenes“ Gehaltsniveau gelten kann, lässt sich natürlich nicht von der Statistik her beantworten. Die Statistik als Wissenschaft kann nur eine Bewertung von Design und Methodik der Befragung und eine Aussage zur Tragfähigkeit der statistischen Analyse liefern.

15.3 Punktschätzung von Anteilswerten

Der Stichprobenmittelwert findet auch bei der Schätzung des Erwartungswerts $p = E(X)$ bernoulli-verteilter Merkmale X Anwendung. Die Bernoulli-Verteilung charakterisiert ein Zufallsexperiment mit zwei möglichen Ausgängen A und \bar{A} , die mit Wahrscheinlichkeit $p = P(A)$ resp. $1 - p = P(\bar{A})$ auftreten. Wenn man ein Bernoulli-Experiment n -mal durchführt, kann man den Ausgang jedes Einzelexperiments anhand der Indikatorvariablen (12.18) modellieren, die gesamte Bernoulli-Kette also durch eine Folge unabhängiger Stichprobenvariablen X_1, X_2, \dots, X_n . Der hieraus gebildete Stichprobenmittelwert \bar{X} lässt sich zur Schätzung des Erwartungswerts p heranziehen. Letzterer repräsentiert den zu erwartenden *Anteil* der Ausgänge mit A . Für die Schätzfunktion $\hat{p} := \bar{X}$ gilt analog zu (15.6)

$$E(\hat{p}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \dots + E(X_n)] = \frac{1}{n} \cdot n \cdot p = p. \quad (15.10)$$

Die bernoulli-verteilten Variablen X_i besitzen nach (12.16) die Varianz $\sigma^2 = p(1 - p)$. Für die Varianz $V(\hat{p})$ von $\hat{p} = \bar{X}$ folgt daher

$$V(\hat{p}) = V(\bar{X}) = \frac{p \cdot (1 - p)}{n}. \quad (15.11)$$

Beispiele zur Schätzung von Anteilswerten Als besonders einfaches Beispiel für die Schätzung eines Anteils lässt sich die Schätzung der Wahrscheinlichkeit p für „Zahl“ bei einem Münzwurf anführen. Als Schätzung \hat{p} für den bei wiederholter Durchführung eines Münzwurfs zu erwartenden Anteil der Ausgänge mit „Zahl“ bietet sich hier die relative Häufigkeit von „Zahl“ an. Aus Abbildung 12.4 ist zu ersehen, dass sich diese mit zunehmender Länge n der Münzwurfserie tendenziell dem unbekannten Parameter p immer weiter annähert.

Auch beim Politbarometer geht es um die Schätzung von Anteilswerten, nämlich um die Schätzung des Anteils von Wählern mit Präferenz für

eine bestimmte Partei. Mit den in Tabelle 9.3 wiedergegebenen Daten des Politbarometers vom 8. Dezember 2017 würde man den Anteil p der Frauen in Deutschland mit SPD-Präferenz durch die relative Häufigkeit $\hat{p} = \frac{134}{627} = 0,214$ schätzen (21,4%). Auch hier wird die Qualität der Schätzung tendenziell besser, wenn man die Anzahl der Frauen in der Wählerstichprobe erhöht. Ein weiteres Beispiel ist die Schätzung von Armutgefährdungsquoten auf der Basis von Haushaltsbefragungen. Will man hier brauchbare Schätzungen für kleinere räumliche Einheiten gewinnen, etwa für Landkreise, muss für jede einzelne Einheit eine hinreichend große Anzahl n von Haushaltsstichproben verfügbar sein.

Exkurs 15.2: Schätzung bei sehr kleinen Stichproben

Wenn man anhand einer Stichprobe eine Schätzung durchführt und diese für einen Rückschluss auf eine umfassendere Grundgesamtheit nutzt, sollte die Stichprobe für die Grundgesamtheit möglichst repräsentativ sein. Eine Stichprobe erfüllt diese Bedingung nur näherungsweise. Außerdem streut die Ausprägung der verwendeten Schätzfunktion von Stichprobe zu Stichprobe. Dies schlägt sich in der Standardabweichung der Schätzfunktion (*Standardfehler*) nieder, die mit Abnahme des Stichprobenumfangs n zunimmt. Wenn der Stichprobenumfang n sehr klein ist, insbesondere im Verhältnis zum Umfang N der Grundgesamtheit, kann die Qualität der Schätzung fragwürdig sein.

Diese Aussage sei an einer Schätzung von Armutgefährdungsquoten in den Kreisen des Bundeslandes Rheinland-Pfalz illustriert. Die Schätzung basiert auf Daten des Jahres 2010 aus der europaweiten Erhebung *EU-SILC* (European Union Statistics on Income and Living Conditions). Diese Erhebung umfasst für ganz Deutschland nur knapp 15 000 Haushalte – zu wenig, um für das Bundesland Rheinland-Pfalz auf Kreisebene Schätzungen von Armutgefährdungsquoten in ausreichender Qualität zu gewinnen.

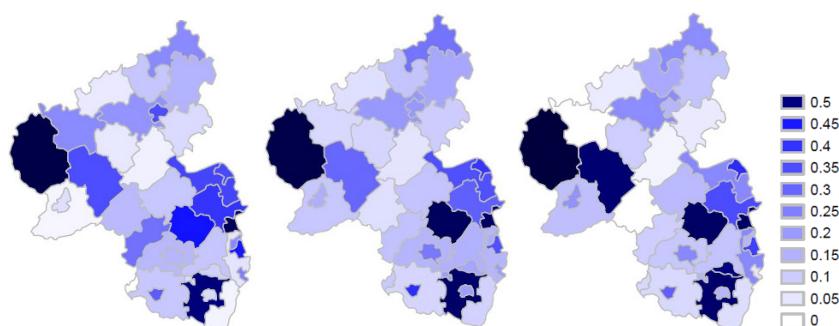


Abb. 15.4: Schätzung von Armutgefährdungsquoten in den Kreisen von Rheinland-Pfalz (Karte: Bundesamt für Kartographie und Geodäsie / Geodatenzentrum)

Um dies zu illustrieren, wurden an der Universität Trier im Rahmen einer kleinen Simulationsstudie wiederholt einfache Zufallsstichproben aus syntheti-

schen, auf EU-SILC basierenden Daten generiert. Dabei wurden, ähnlich wie bei EU-SILC, in den Landkreisen und kreisfreien Städten jeweils knapp 0,04% aller Haushalte in der Stichprobe berücksichtigt. Auf der Grundlage dieser sehr kleinen Stichproben erfolgte eine Schätzung der Armutgefährdungsquoten. Die Ergebnisse der Schätzung sind in Abbildung 15.4 für drei Stichproben anhand eingefärbter Landkarten dargestellt.

Die unterschiedlichen Färbungen repräsentieren Klassen für die Armutgefährdungsquote. Die Definition der Klassengrenzen ist der Legende zu entnehmen. Der Wert 0,2 etwa ist ein Anteilswert und als 20% zu interpretieren. Vergleicht man die drei Landkarten, stellt man fest, dass es erhebliche Unterschiede in den Färbungen gibt. Als Schwellenwert für die Unterscheidung zwischen „armutsgefährdet“ und „nicht armutsgefährdet“ gilt 60% des Medians der Einkommensverteilung für Deutschland (vgl. auch Exkurs 5.1).

Man kann aus den drei Karten auch ohne Angabe der genauen Schätzergebnisse erkennen, dass die Zuverlässigkeit der Schätzungen bei kleinen Auswahlsätzen $\frac{n}{N}$ nicht gesichert ist. Schätzergebnisse für eine größere Region, die auf einer großen Stichprobe fußen, lassen sich jedenfalls nicht ohne Weiteres übertragen auf kleinere räumliche Einheiten, für die nur eine kleine Datenbasis vorliegt.

15.4 Intervallschätzung für Erwartungswerte

Eine **Punktschätzung** $\hat{\theta}$ für einen Parameter θ liefert einen einzigen Schätzwert, der meist mit θ nicht exakt übereinstimmt. Zur Beurteilung der Güte einer Punktschätzung spielt die Verzerrung (15.2) eine Rolle, daneben aber auch die Varianz oder die Standardabweichung des Schätzers. Beide gehen in den als Gütemaß für Schätzer verwendeten mittleren quadratischen Fehler MSE aus (15.4) ein.

Bei einer **Intervallschätzung** werden die beiden Aspekte „mittlere Lage“ und „Streuung“ einer Schätzfunktion auf andere Weise verknüpft, nämlich durch Ermittlung eines Intervalls, das den zu schätzenden Parameter θ mit einer Wahrscheinlichkeit von mindestens $1 - \alpha$ enthält.³ Das Intervall, dessen Grenzen sich aus den Stichprobendaten errechnen, soll natürlich möglichst schmal sein, also eine geringe Länge aufweisen.

Das Konzept der Intervallschätzung sei anhand der Schätzung für den Erwartungswert $\mu = E(X)$ eines $N(\mu; \sigma^2)$ -verteilten Merkmals X illustriert. Es sei zunächst vorausgesetzt, dass die Varianz $\sigma^2 = V(X)$ bekannt sei. Die Stichprobenwerte x_1, x_2, \dots, x_n werden als Ausprägungen unabhängiger $N(\mu; \sigma^2)$ -verteilter Zufallsvariablen X_1, X_2, \dots, X_n interpretiert. Die

Schätzung bei
normalverteiltem

Merkmal und
bekannter Varianz

³Bei Intervallschätzungen von Kenngrößen θ stetiger Verteilungen kann man den Zusatz „mindestens“ streichen – hier lässt sich das Intervall exakt so bestimmen, dass es θ mit Wahrscheinlichkeit $1 - \alpha$ überdeckt.

Zufallsvariable $Z := \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$ ist dann gemäß (14.7) standardnormalverteilt. Damit liegt sie mit Wahrscheinlichkeit $1 - \alpha$ in dem durch die Quantile $z_{\alpha/2} = -z_{1-\alpha/2}$ und $z_{1-\alpha/2}$ begrenzten Intervall $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$, das in Abbildung 13.5 veranschaulicht ist. Es gilt also für den standardisierten Stichprobenmittelwert Z die Wahrscheinlichkeitsaussage

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} \leq z_{1-\alpha/2}\right) = 1 - \alpha. \quad (15.12)$$

Wenn man die drei Terme der Ungleichungskette in der Klammer mit $\frac{\sigma}{\sqrt{n}}$ erweitert, dann jeweils \bar{X} subtrahiert und mit -1 multipliziert, folgt

$$P\left(\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \quad (15.13)$$

Das mit KI bezeichnete Intervall

$$KI = \left[\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \quad (15.14)$$

enthält demnach den unbekannten Verteilungsparameter μ mit Wahrscheinlichkeit $1 - \alpha$. Das Intervall KI ist das **Konfidenzintervall** zum **Konfidenzniveau** $1 - \alpha$ für μ . Es repräsentiert eine Intervallschätzung für μ . Die Berechnung von (15.14) setzt voraus, dass die Varianz σ^2 bzw. die Standardabweichung σ der $N(\mu; \sigma^2)$ -verteilten Variablen X bekannt ist, also nicht erst über eine Schätzung zu ermitteln ist.

In (15.13) bzw. (15.14) geht die Ausprägung $\hat{\mu} = \bar{x}$ des Schätzers \bar{X} ein, die von Stichprobe zu Stichprobe variiert. Die Intervallgrenzen sind also *zufallsabhängig*. Die Länge des Konfidenzintervalls ist fest und durch

$$\text{Länge}(KI) = 2 \cdot z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad (15.15)$$

gegeben, hängt also von der Irrtumswahrscheinlichkeit α und vom Stichprobenumfang n ab. Mit abnehmender Irrtumswahrscheinlichkeit α (wachsendem Konfidenzniveau $1 - \alpha$) nimmt die Länge (15.15) zu, weil das Quantil $z_{1-\alpha/2}$ dann größere Werte annimmt (vgl. Tabelle 19.3). Mit zunehmendem n wird das Konfidenzintervall schmäler.

Abbildung 15.5 zeigt Konfidenzintervalle, die nach (15.14) mit $1 - \alpha = 0,95$ berechnet wurden und jeweils auf n per Simulation generierten Stichprobendaten basieren. Bei der Simulation wurde für die Stichprobenvarianz $\sigma^2 = 1$ und für den zu schätzenden Erwartungswert $\mu = 0$ gewählt (Generierung von n Werten aus einer Standardnormalverteilung). Der Vorgang wurde k -mal ausgeführt mit $k = 100$. Insgesamt wurden somit jeweils $k = 100$ Konfidenzintervalle per Simulation erzeugt. Eine Simulation bietet den Vorteil, dass der üblicherweise unbekannte Parameter μ , für



Interaktives Objekt
„Konfidenzintervalle
für μ (Varianz
bekannt)“

den man Intervallschätzungen berechnen will, ausnahmsweise bekannt ist (kontrollierte Laborsituation).

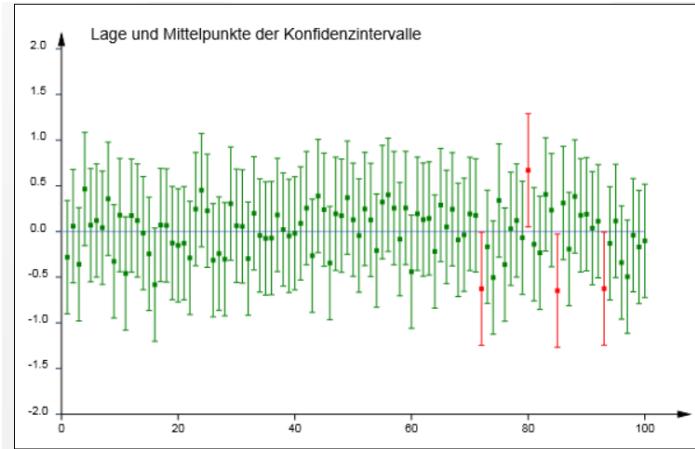


Abb. 15.5: Konfidenzintervalle für μ (berechnet aus standardnormalverteilten Stichprobendaten, per Simulation erzeugt, $n = 10$)

Für die Länge der Konfidenzintervalle erhält man mit (15.15) und Tabelle 19.3 im Falle $n = 10$ den Wert $\frac{2 \cdot 1,96}{\sqrt{10}} \approx 1,24$. Im Falle $n = 50$ sind die Konfidenzintervalle erwartungsgemäß schmäler (breitere Datenbasis). Ihre Länge errechnet sich zu $\frac{2 \cdot 1,96}{\sqrt{50}} \approx 0,55$. Eine Verdopplung von n führt nach (15.15) dazu, dass die Länge des Konfidenzintervalls sich um den Faktor $\frac{1}{\sqrt{2}} \approx 0,71$ verändert, also auf ca. 71% der vorherigen Länge schrumpft. In Abbildung 15.5 überdecken vier (4%) der per Simulationsexperiment erzeugten $k = 100$ Intervalle den Parameter $\mu = 0$ nicht. Bei Wiederholung des Experiments kann eine andere Anzahl nicht-überdeckender Intervalle resultieren, z. B. sechs (6%) oder fünf (5%). Der theoretische Wert für die Anzahl nicht-überdeckender Intervalle, den man approximativ bei Wahl eines sehr großen Wertes für k erreicht, ist $\alpha = 0,05$ (5%). Es ist jedenfalls festzuhalten, dass ein konkretes Konfidenzintervall den unbekannten Parameter – auch bei klein gewählter Irrtumswahrscheinlichkeit α – nicht zwingend überdeckt.

Schätzung bei normalverteiltem Merkmal und unbekannter Varianz	Die vorstehenden Ableitungen sind leicht zu modifizieren, wenn man die Varianz σ^2 nur in Form einer Schätzung $\hat{\sigma}^2$ kennt. Ausgangspunkt ist hier nicht mehr der standardisierte Stichprobenmittelwert aus (14.7), sondern die mit $n - 1$ Freiheitsgraden t -verteilte Zufallsvariable aus (14.10). Man erhält anstelle von (15.14)
---	---

$$KI = \left[\bar{X} - t_{n-1;1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}}; \bar{X} + t_{n-1;1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}} \right]. \quad (15.16)$$

Diese Formel unterscheidet sich von (15.14) darin, dass statt der festen Größe σ nun die Zufallsvariable S^* erscheint und statt zweier Quantile der

**Aufgabe 15.2**

Standardnormalverteilung die entsprechenden Quantile der t -Verteilung mit $n - 1$ Freiheitsgraden. Die erste Ersetzung hat wegen (15.9) im Mittel keinen Effekt auf die Länge des Konfidenzintervalls. Das $(1 - \alpha/2)$ -Quantil der t -Verteilung mit $n - 1$ Freiheitsgraden ist allerdings stets größer als das $(1 - \alpha/2)$ -Quantil der Standardnormalverteilung, wobei die Unterschiede mit wachsendem n kleiner werden (vgl. Tabelle 13.1). Das Konfidenzintervall (15.16) ist folglich im Mittel länger. Ein wesentlicher Unterschied gegenüber (15.14) besteht auch darin, dass die Länge

$$\text{Länge}(KI) = 2 \cdot t_{n-1; 1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}} \quad (15.17)$$

des Konfidenzintervalls (15.16) nicht nur von der Irrtumswahrscheinlichkeit α und dem Stichprobenumfang n , sondern auch von der jeweiligen Ausprägung von S^* abhängt, also zufallsabhängig ist.

15.5 Intervallschätzung für Anteilswerte

Auch für die Schätzung von Anteilswerten kommt die Verwendung von Konfidenzintervallen in Betracht. Ausgangspunkt ist eine Grundgesamtheit, deren Elemente bezüglich eines Merkmals mit Wahrscheinlichkeit p eine Ausprägung A und mit Wahrscheinlichkeit $1 - p$ eine komplementäre Ausprägung \bar{A} aufweisen (Bernoulli-Verteilung mit unbekanntem Verteilungsparameter p). Entnimmt man einer solchen dichotomen Grundgesamtheit eine Stichprobe des Umfangs n (Durchführung von n voneinander unabhängigen Bernoulli-Experimenten) und zählt die Anzahl X der Ausgänge A , so gilt $X \sim B(n; p)$ (lies: X ist binomialverteilt mit den Parametern n und p). Erwartungswert und Varianz der Zählvariablen X sind nach (12.20) und (12.21) durch $E(X) = n \cdot p$ resp. durch $V(X) = n \cdot p \cdot (1 - p)$ gegeben. Der Anteil $\hat{p} := \frac{X}{n}$ der Ausgänge A in der Stichprobe (Punktschätzung für p) hat den Erwartungswert $E(\hat{p}) = p$ und die Varianz $V(\hat{p}) = \frac{p \cdot (1-p)}{n}$. Nach dem Zentralen Grenzwertsatz (Exkurs 14.2) gilt für die standardisierte Fassung

$$Z_n := \frac{\hat{p} - E(\hat{p})}{\sqrt{V(\hat{p})}} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \quad (15.18)$$

des Anteils \hat{p} , dass sie bei hinreichend großem Stichprobenumfang approximativ standardnormalverteilt ist. Für die Zufallsvariable Z_n gilt somit die Wahrscheinlichkeitsaussage

$$P(z_{\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) = P\left(-z_{1-\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Die Ungleichungskette in der letzten Klammer ist gleichbedeutend mit



Interaktives Objekt
„Intervallschätzung
für Anteilswerte
(Münzwurf)“

Nach Auflösen nach p einschließlich quadratischer Ergänzung erhält man – vgl. SCHLITTGEN (2012, Abschnitt 14.5.2) – mit der Abkürzung $z := z_{1-\alpha/2}$ als **Konfidenzintervall** zum **Konfidenzniveau** $1 - \alpha$ für den unbekannten Anteilswert p die Näherungsformel

$$KI \approx \left[\frac{\hat{p} + \frac{z^2}{2n} - z \cdot \sqrt{(\frac{z}{2n})^2 + \frac{\hat{p} \cdot (1 - \hat{p})}{n}}}{1 + \frac{z^2}{n}}, \frac{\hat{p} + \frac{z^2}{2n} + z \cdot \sqrt{(\frac{z}{2n})^2 + \frac{\hat{p} \cdot (1 - \hat{p})}{n}}}{1 + \frac{z^2}{n}} \right]. \quad (15.19)$$

Das Intervall wird nach dem US-amerikanischen Mathematiker Edwin Bidwell WILSON (1879 – 1964) auch **Wilson-Intervall** genannt. Für die Anwendung von (15.19) sollte $n \cdot \hat{p} \cdot (1 - \hat{p}) \geq 5$ gelten. Wenn n noch größer ist, etwa $n \cdot \hat{p} \cdot (1 - \hat{p}) \geq 9$, kann man eine einfachere Formel heranziehen, bei der $z_{1-\alpha/2}$ wieder ausführlich notiert ist:

$$KI \approx \left[\hat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right]. \quad (15.20)$$

Die Intervallgrenzen hängen von der Ausprägung der Zufallsvariablen \hat{p} ab, sind also *zufallsabhängig*.

Die Formel (15.20) kann man z. B. anwenden, um aus den Ergebnissen von n voneinander unabhängigen Münzwürfen eine Intervallschätzung der Wahrscheinlichkeit p für das Auftreten von „Zahl“ zu gewinnen. Mittelpunkt des nach (15.20) berechneten approximativen Konfidenzintervalls ist dabei der Wert \hat{p} für die nach n Würfen ermittelte relative Häufigkeit von „Zahl“.

Ein Konfidenzintervall für den unbekannten Anteilswert p ist nicht zu verwechseln mit dem im Folgenden mit SI bezeichneten **zentralen Schwankungsintervall** für die zur Schätzung von p verwendete Schätzfunktion \hat{p} . Ein approximatives Schwankungsintervall für \hat{p} ergibt sich, wenn man bei der direkt nach (15.18) folgenden Wahrscheinlichkeitsaussage für Z_n die Ungleichungskette in der letzten Klammer nicht nach p auflöst, sondern nach \hat{p} . Es resultiert das Intervall

$$SI \approx \left[p - z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}; p + z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \right], \quad (15.21)$$

dessen Grenzen *nicht zufallsabhängig* sind. Eine Realisation von \hat{p} ist in diesem Intervall näherungsweise mit Wahrscheinlichkeit $1 - \alpha$ enthalten. Bei einer Münzwurfserie mit n Würfen und „fairer“ Münze liegt demnach die beobachtete relative Häufigkeit (Realisation von \hat{p}) näherungsweise mit Wahrscheinlichkeit 0,95 innerhalb des Intervalls $\left[0,5 - \frac{0,98}{\sqrt{n}}; 0,5 + \frac{0,98}{\sqrt{n}} \right]$.



Interaktives Objekt
„Schwankungsintervall
für Anteilswerte
(Münzwurf)“

Beispiel 15.3: Schwankungsintervall für Schätzer von Anteilswerten

Der Anteil von Mädchen in einer Population von Neugeborenen ist eine Zufallsvariable Y , deren Ausprägung von Stadt zu Stadt variieren kann. Abbildung 15.6 zeigt 35 Datenpunkte $(x; y)$, wobei $x = n$ die Anzahl der Neugeborenen in einer von insgesamt 35 ausgewählten kleineren oder mittelgroßen Städten in Nordrhein-Westfalen bezeichnet und y die Anzahl der im Jahr 2015 in der jeweiligen Stadt registrierten Neugeborenen.



Interaktives Objekt
„Schwankungsintervall
für Anteilswerte
(Geburten in NRW)“

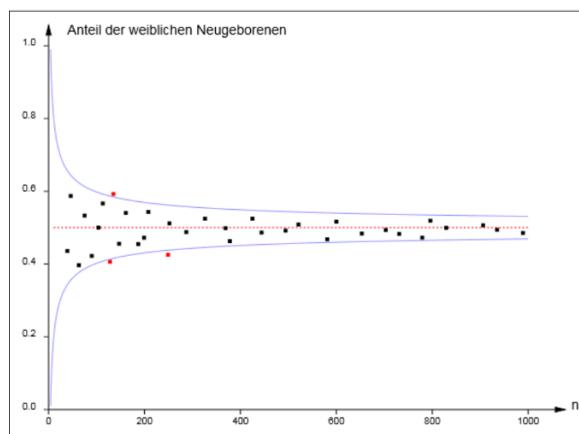


Abb. 15.6: Anteil weiblicher Neugeborener in NRW-Städten mit Schwankungsintervall zum Sicherheitsniveau 0,95 für den Anteilswert (n = Anzahl der Geburten im Jahr 2015 in den einzelnen Städten). Datenquelle: IT.NRW

Eingezeichnet ist auch – als Funktion von n – das durch (15.20) und $p = 0,5$ definierte approximative Schwankungsintervall zum Niveau 0,95 für Y (zweiteilige Trichterkurve). Datenpunkte, die außerhalb des Trichters liegen, sind rot markiert. Für ein festes n ergeben sich die Grenzen des Schwankungsintervalls als Schnittpunkte eines an der Stelle n ausgeführten vertikalen Schnittes durch den Trichter.

Exkurs 15.3: Dosis-Eskalations-Studien in der klinischen Forschung

Vor der Zulassung neuer Medikamente zur Behandlung aggressiver Tumore ist in klinischen Studien eine Dosis zu ermitteln, die bei Abwägung der Nebenwirkungen für die betroffenen Patienten gerade noch verträglich erscheint. Diese maximal verträgliche Dosis wird in der einschlägigen Literatur als *MTD* bezeichnet (engl.: *maximum tolerated dose*) und der Prozess der MTD-Bestimmung als Phase I der Studie, der sich weitere Phasen anschließen (u. a. Bestimmung der Effizienz des Therapieregimes und Vergleich mit alternativen Behandlungsformen). Bei der Planung der Phase I ist vorab in einem **Versuchsplan** festzulegen, ab welcher – z. B. anhand von Laborbefunden quantifizierbarer – Intensität von Nebenwirkungen eine Dosis für einen einzelnen Patienten als nicht mehr verträglich angesehen wird. Die Dosis, bei der diese individuelle kritische Schwelle erreicht wird, wird mit *DLT* abgekürzt (engl.: *dose-limiting toxicity*) und das Erreichen der Schwelle als DLT-Ereignis bezeichnet.

Nachdem das Medikament entwickelt und in Tierversuchen getestet wurde (präklinische Tests), beginnt man die Phase I zumeist mit einer sehr gerin- gen Startdosis („Microdosing“), bei der noch keine stärkeren Nebenwirkungen erwartet werden. Danach wird die Dosis in Stufen und maximal bis zu ei- nem im Versuchsplan festgelegten Höchstniveau gesteigert. Die an der Studie beteiligten Patienten sind typischerweise solche, für die keine bekannten Be- handlungsansätze Erfolg versprechen. Sie müssen natürlich umfassend über die Nebenwirkungen aufgeklärt sein, die mit dem Einsatz des neuen Medikaments verbunden sein können.

Ein klassisches Verfahren zur Dosisbestimmung Die am häufigsten angewandte Strategie zur Dosisfestsetzung und zur Er- mittlung der MTD ist das **3+3-Studiendesign**. Bei diesem Design wird die Startdosis zunächst an drei Patienten verabreicht. Kommt es innerhalb dieser Patientenkohorte zu keinen DLT-Ereignissen ($DLT = 0/3$), bekommen drei Patienten einer neuen Kohorte die nächsthöhere Dosis. Entwickelt genau eine dieser drei Testpersonen eine DLT ($DLT = 1/3$), bekommen drei weitere Pati- enten dieselbe Dosis. Gibt es in der auf sechs Personen erweiterten Kohorte keine zusätzlichen DLTs ($DLT = 1/6$), kann zur nächsthöheren Dosis gegangen werden. Sollten allerdings auch bei der hinzugekommenen Patientengruppe DLTs erkannt werden ($DLT > 1/6$), wird entweder in die nächstniedrigere Dosis abgeschwächt oder – falls in der niedrigeren Dosis bereits sechs Patienten getestet waren – diese als MTD festgestellt. Dasselbe geschieht, wenn von den ersten drei Patienten mehr als einer eine DLT erfährt. Abbildung 15.7 verdeutlicht den Algorithmus.

Vorgehensweise bei
der Dosisbestimmung
für neue
Medikamente

Ein klassisches
Verfahren zur
Dosisbestimmung

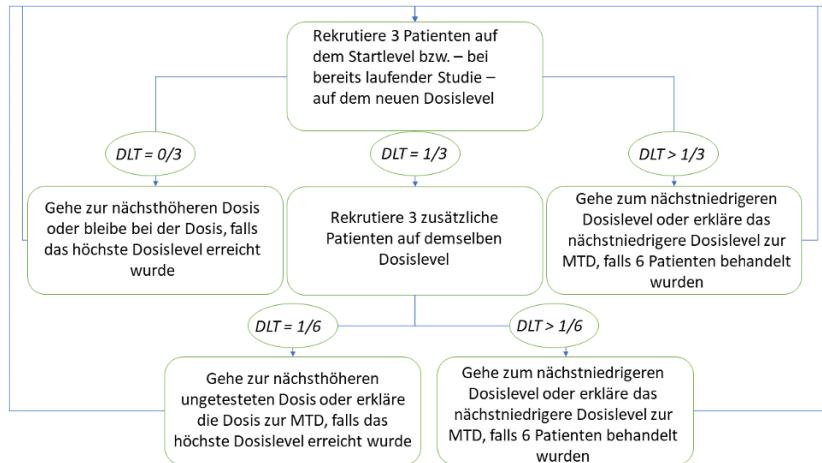


Abb. 15.7: Schema des 3 + 3-Studiendesigns

Bis heute werden über 90 % der Dosis-Eskalations-Studien auf der Basis des 3+3-Studiendesigns durchgeführt. Inzwischen gibt es allerdings zunehmend Kritik an dessen statistischer Aussagekraft und der Fähigkeit zur Bestimmung der MTD. Ein Kritikpunkt ist unter anderem, dass lediglich die letzten drei bzw. sechs DLT-Reaktionen von Patienten und nicht die gesamte verfügbare Information zur Bestimmung des weiteren Vorgehens genutzt wird. Dies verdeutlicht der folgende, beispielhafte Ablauf, bei dem nach Beobachtung eines DLT-Ereignisses bei Kohorte 5 die MTD mit 8 mg/kg bestimmt ist:

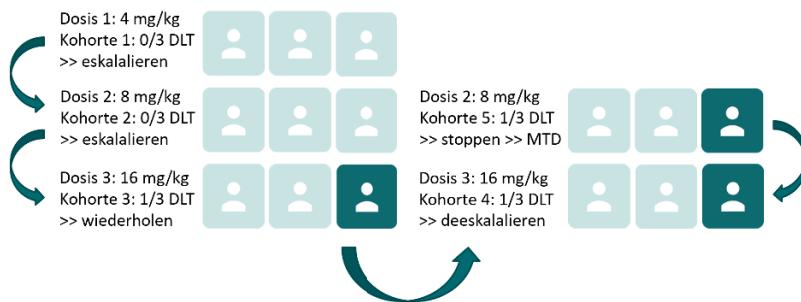


Abb. 15.8: Ablaufbeispiel eines 3 + 3-Studiendesigns.

Lediglich die Information aus sechs Patienten (Kohorten 2 und 5) wird hier herangezogen, um die MTD von 8 mg/kg Körpergewicht und deren geschätztes Toxizitätsrisiko in Gestalt der relativen Häufigkeit von DLT-Ereignissen zu bestimmen ($\frac{1}{6}$ bzw. 16,7 %). Das 3+3-Studiendesign beschränkt sich demnach auf eine Punktschätzung.

Vergleicht man die Punktschätzer für die drei in Abbildung 15.8 aufgeführten Dosisstufen, so ergeben sich Toxizitätsrisiken von 0 (0 %) für die Dosis 4 mg/kg, 0,167 (16,7 %) bei 8 mg/kg und 0,333 (33,3 %) bei Verabreichung von 16 mg/kg

Körpergewicht. Da in allen drei Dosisgruppen $n \cdot \hat{p} \cdot (1 - \hat{p}) < 9$ gilt, kann Formel (15.20) nicht angewandt werden. Der approximative Intervallschätzer (15.19) (**Wilson-Intervall**) liefert $[0; 0,561]$ für die Dosisstufe 4 mg/kg, $[0,001; 0,564]$ für 8 mg/kg und $[0,097; 0,700]$ bei 16 mg/kg.

Schätzung von Anteilswerten mit R

Um solche Berechnungen komfortabel durchführen zu können, bietet sich der Einsatz einer Statistik-Software wie R an. R ermöglicht es, inzwischen weit über 10.000 Erweiterungspakete einzubinden. Im Paket `Hmisc` finden sich zahlreiche Funktionen zur Datenanalyse. Wir nutzen hier die Funktion `binconf` zur Berechnung von Konfidenzintervallen für den Anteilswert p einer Binomialverteilung in Dosisgruppe 2 mit $k = 1$ DLTs bei $n = 6$ Patienten. Als Konfidenzniveau wählen wir $(1 - \alpha) = 0,95$. R gibt sowohl den Punktschätzer (**PointEst**) als auch die untere (**Lower**) und obere (**Upper**) Konfidenzschanke an, d.h. die Grenzen des Konfidenzintervalls:

```
> binconf(1, 6, alpha=0.05, method="wilson")
 PointEst      Lower      Upper
 0.1666667  0.008548882  0.5635028
```

Abb. 15.9: Punktschätzer \hat{p} mit approximativem Konfidenzintervall

Da sogar für alle drei Dosisstufen $n \cdot \hat{p} \cdot (1 - \hat{p}) < 5$ gilt, ist selbst die Anwendung des Wilson-Intervalls kritisch zu hinterfragen. R ermöglicht es auch, die exakten Konfidenzintervalle zu ermitteln, die auf der Binomialverteilung beruhen. Hierzu wird der R-Code folgendermaßen abgeändert:

```
> binconf(1, 6, alpha=0.05, method="exact")
 PointEst      Lower      Upper
 0.1666667  0.004210745  0.6412346
```

Abb. 15.10: Punktschätzer \hat{p} mit exaktem Konfidenzintervall

Die Punktschätzung verändert sich nicht, die Intervallschätzung jedoch erheblich. Als Intervallschätzer für die drei Gruppen ergeben sich mit der exakten Methode $[0; 0,708]$ für die Dosis 4 mg/ kg, $[0; 0,641]$ bei 8 mg/kg sowie $[0,043; 0,777]$ für 16 mg/kg. Offensichtlich überlappen sich die Konfidenzintervalle weitgehend. Das bedeutet, die Ergebnisse der drei Dosisgruppen sind statistisch kaum ununterscheidbar. Es gibt demnach kein rationales Argument, die MTD-Bestimmung auf dem 3+3-Studiendesign aufzubauen.

Alternative Ansätze zur Dosisbestimmung

Aufgrund der Kritik am 3 + 3-Studiendesign werden heute neuere Methoden verwendet, die auf Verfahren der **Bayes-Statistik** basieren. Diese versteht sich als Alternative zur klassischen schließenden Statistik. Während letztere zum Schätzen von Parametern und zum Testen von Hypothesen allein auf Stichprobendaten zurückgreift, verwendet die Bayes-Statistik noch zusätzliche Informationen, z. B. Vorinformationen über einen zu schätzenden Parameter in Form einer Dichtefunktion. Eine Einführung in bayesianische Statistik würde den Rahmen dieses Buches sprengen (vgl. hierzu z. B. TSCHIRK (2018)).

16 Statistische Testverfahren



Vorschau auf
das Kapitel

Dieses Kapitel beginnt mit einer Klassifikation von Testverfahren nach unterschiedlichen Kriterien, u. a. nach der Annahme / dem Fehlen der Annahme einer bestimmten Verteilung in der Grundgesamtheit (parametrische vs. nicht-parametrische Tests) oder nach der Verteilung der Prüfgröße (z. B. Gauß-Test, t -Test, χ^2 -Test). Am Beispiel des Gauß-Tests und des t -Tests für den Erwartungswert einer normalverteilten Grundgesamtheit wird die Vorgehensweise beim Testen von Hypothesen erläutert. Da die Entscheidung bei einem Test nur auf Stichprobeninformation beruht, sind Fehlentscheidungen grundsätzlich nicht auszuschließen. Es wird zwischen zwei Fehlerarten unterschieden, nämlich dem über das Testdesign kontrollierten Fehler 1. Art und dem Fehler 2. Art. Als Instrument zur Beurteilung der Leistungsfähigkeit eines Tests wird die Gütekfunktion präsentiert. An dieser lassen sich die Eintrittswahrscheinlichkeiten für beide Fehlerarten ablesen.

Neben der Behandlung von Tests für den Erwartungswert bei normalverteilter Grundgesamtheit – auch solchen, die Information aus zwei Stichproben nutzen – wird ein Test vorgestellt, der sich auf die Varianz bezieht. Abschließend wird noch ein Test vorgestellt, der auf die Prüfung der Unabhängigkeit zweier Merkmale abzielt.

In der Forschungspraxis will man nicht nur Modellparameter schätzen, sondern häufig auch Hypothesen H_0 und H_1 auf der Basis von Daten überprüfen. Ausgangspunkt ist eine Fragestellung, die sich oft auf die Verteilung eines Merkmals bzw. auf eine Kenngröße der Verteilung eines Merkmals in einer Grundgesamtheit von Merkmalsträgern bezieht. Man zieht eine Zufallsprobe, aggregiert die Stichprobenvariablen zu einer Stichprobenfunktion und nutzt deren Ausprägung dazu, eine Entscheidung bezüglich der zu testenden Hypothesen zu fällen:

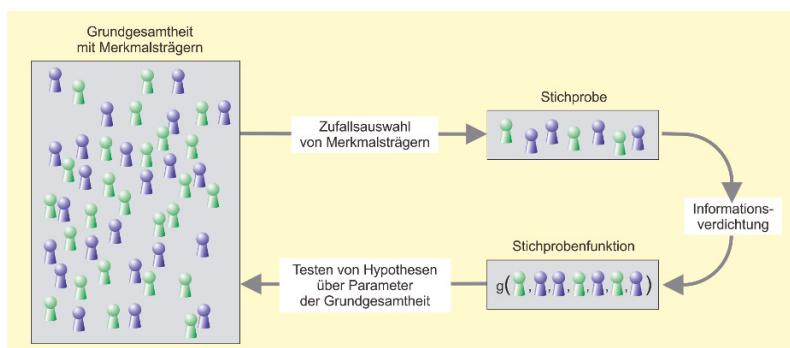


Abb. 16.1: Vorgehensweise beim Testen von Hypothesen

16.1 Arten statistischer Tests

Klassifikationen für statistische Tests:

- nach der Anzahl der Stichproben

- in Abhängigkeit von Verteilungsannahmen

- nach dem Inhalt der Hypothesen

- nach der Verteilung der Prüfstatistik

- nach der Hypothesenformulierung

Man spricht von einem **Einstichproben-Test**, wenn ein Test die Information nur einer Stichprobe verwendet. Manchmal testet man auch Hypothesen, die Information aus zwei Stichproben nutzen und sich auf *zwei* Zufallsvariablen beziehen, z. B. auf die Erwartungswerte oder Varianzen zweier Variablen X und Y . Solche Tests, in die zwei Stichproben eingehen, heißen **Zweistichproben-Tests**. Es gibt auch Tests, die mit k Stichproben arbeiten ($k > 2$) und k Zufallsvariablen betreffen. Diese werden entsprechend als **k -Stichproben-Tests** etikettiert.

Wenn man für die Teststatistik die Kenntnis des Verteilungstyps in der Grundgesamtheit voraussetzt, liegt ein **parametrischer Test** vor, andernfalls ein **verteilungsfreier** oder **nicht-parametrischer Test**.

Man kann Tests auch danach klassifizieren, worauf sich die Hypothesen beziehen. So gibt es **Tests für Erwartungswerte**, **Tests für Varianzen** oder **Tests für Anteile** von Populationen. Für die drei genannten Fälle gibt es Ein- und Mehrstichproben-Tests, d. h. die aufgeführten Testklassifikationen überschneiden sich. Mit **Anpassungstests** prüft man, ob eine Zufallsvariable einer bestimmten Verteilung folgt, z. B. der Normalverteilung. Bei **Unabhängigkeitstests** will man eine Aussage darüber gewinnen, ob zwei Zufallsvariablen unabhängig sind.

Häufig werden statistische Tests, deren Prüfstatistik einer bestimmten diskreten oder stetigen Verteilung folgt, zu einer Gruppe zusammengefasst. So gibt es ganz unterschiedliche Tests, die mit einer χ^2 -, t - oder F -verteilten Testgröße operieren. Diese Tests werden dann als **χ^2 -Tests**, **t -Tests** resp. als **F -Tests** angesprochen. Ein Test mit standardnormalverteilter Prüfstatistik wird als **Gauß-Test** bezeichnet. Der t -Test kommt z. B. beim Testen von Hypothesen über Erwartungswerte normalverteilter Grundgesamtheiten ins Spiel, findet aber auch Anwendung beim Testen von Hypothesen über Regressionskoeffizienten bei normalverteilten Störvariablen. Es gibt also nicht *den* t -Test, sondern ganz unterschiedliche t -Tests, deren Gemeinsamkeit darin besteht, dass die Prüfstatistik bei Gültigkeit gewisser Annahmen einer t -Verteilung folgt.

Bei der Prüfung von Hypothesen über Parameter kann es darauf kommen, Veränderungen nach beiden Seiten zu entdecken (**zweiseitiger Test**) oder auch nur in eine Richtung (**einseitiger Test**). Wenn zwei Hypothesen direkt aneinandergrenzen, wie etwa im Falle der Hypothesen $H_0 : \mu = \mu_0$ und $H_1 : \mu \neq \mu_0$, spricht man von einem **Signifikanztest**. Andernfalls, etwa im Falle $H_0 : \mu = \mu_0$ und $H_1 : \mu = \mu_1$ ($\mu_0 < \mu_1$), liegt ein **Alternativtest** vor.

Im Folgenden stehen ausgewählte parametrische Signifikanztests für Erwartungswerte, Varianzen und Anteilswerte im Vordergrund, an denen

die Vorgehensweise beim Testen von Hypothesen erläutert wird. Eine ausführlichere Behandlung statistischer Tests findet man u. a. bei MOSLER / SCHMID (2011, Kapitel 6) oder SCHLITTGEN (2012, Kapitel 15 - 16).



Beispiel 16.1: Hypothesentests in der Praxis

Es fällt nicht schwer, Anwendungsfelder und Beispiele für Hypothesentests aus unterschiedlichen Bereichen aufzuführen:

- Anhand der Daten der von Eurostat alle 4 Jahre durchgeführten Verdienststrukturerhebung wird untersucht, ob sich in einzelnen Branchen das Verdienstniveau für Frauen und Männer unterscheidet. Solche Informationen sind der Ausgangspunkt für Strategien zur Verringerung eines geschlechtsspezifischen Verdienstgefälles.
- Im Umweltbereich will man aus Daten Informationen gewinnen, ob bestimmte Variablen, bei denen man einen Effekt auf Schadstoffemissionen vermutet, wirklich zur Emissionsreduktion beitragen.
- In der Medizin ist man daran interessiert zu prüfen, ob ein neues Medikament tatsächlich eine von einem Pharmakonzern behauptete Wirkung erzielt. In der Kieferchirurgie will man testen, ob der Erwartungswert des Merkmals „Lebensdauer von Implantaten“ bei Verwendung von Titan oder bei keramischen Werkstoffen verschieden ist und ob Rauchen die Lebensdauer der Implantate beeinflusst.
- In einigen Ländern ist es von besonderem Interesse, aus Daten Informationen über ein etwaiges ungleiches Verhältnis der Geschlechter innerhalb der Bevölkerung zu gewinnen. Dies gilt z. B. für Indien, wo vorgeburtliche Geschlechterselektion durch Traditionen begünstigt wird.
- Bei der Serienfertigung will man die mittlere Lage $\mu = E(X)$ eines häufig als normalverteilt spezifizierten Qualitätsmerkmals X überwachen. Anhand von regelmäßig entnommenen Stichproben lässt sich eine Aussage darüber ableiten, ob der Verteilungsparameter μ noch auf einem Sollniveau μ_0 liegt. Bei Eintritt eines Shifts soll möglichst rasch korrigierend in den Fertigungsprozess eingegriffen werden.

In der *Medizin* wird der Begriff „Test“ als technisches Verfahren zur Diagnose von Erkrankungen (Labordiagnostik) verstanden, in der *Psychologie* als routinemäßig einsetzbares Messinstrument zur Erfassung latenter Variablen bzw. hypothetischer Konstrukte. Im letztgenannten Fall geht es um die Bestimmung der relativen Position von Individuen oder Gruppen bezüglich bestimmter Persönlichkeitsmerkmale, etwa „Leistungsmotivation“, „Intelligenz“ oder „Teamfähigkeit“, meist unter Verwendung von Fragebögen. Oft werden in der Psychologie anstelle

Tests in Medizin
und Psychologie

solcher Einzelmerkmale auch ganze Bündel von Persönlichkeitsmerkmalen anhand eines einzigen Tests gemessen, etwa die sog. *Big Five* der Persönlichkeitsspsychologie.¹

Statistische Tests

Im Folgenden ist mit „Test“ ein statistischer Test gemeint. Charakteristisch für einen statistischen Test ist, dass eine Forschungshypothese mit Daten konfrontiert wird mit dem Ziel, Aufschluss darüber zu gewinnen, ob die Hypothese mit vorhandenen Beobachtungen verträglich ist und daher bis auf weiteres beizubehalten ist oder ob sie aufgrund des empirischen Befunds zu verwerfen ist. Eine Verwerfung der Ausgangshypothese erfolgt, wenn das Stichprobenergebnis in signifikantem Gegensatz zur betreffenden Hypothese steht. Bei einem statistischen Test wird aber bei der Konfrontation zweier sich ausschließender Hypothesen mit empirischen Befunden eine Hypothese *nie* in dem Sinne bewiesen, dass ihre Gültigkeit ohne jede Möglichkeit des Irrtums erwiesen ist. Die am Ende eines statistischen Tests stehende Testentscheidung schließt stets die Möglichkeit einer Fehlentscheidung ein (vgl. Tabelle 16.1).

16.2 Grundbegriffe und Gauß-Test für Erwartungswerte

Null- und Alternativhypothese Die anhand eines Tests zu untersuchende Fragestellung wird in Form einer Nullhypothese H_0 und einer Alternativhypothese H_1 formuliert. Die **Nullhypothese** H_0 beinhaltet eine bisher als akzeptiert geltende Aussage über den Zustand des Parameters einer Grundgesamtheit. Von dieser Hypothese geht man aus und will ihren Wahrheitsgehalt anhand eines Tests empirisch absichern.

Die **Alternativhypothese** H_1 beinhaltet die eigentliche Forschungshypothese. Sie formuliert das, was gezeigt werden soll. Will man etwa bei der industriellen Serienproduktion ein stetiges Merkmal überwachen, dessen mittleres Niveau μ einen Wert μ_0 nicht überschreiten soll, wird man $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ testen (rechtsseitiger Test). Bei der Überwachung des mittleren Füllvolumens für Tinte bei Tintendruckerpatronen, wird sich das Eichamt vor allem für Unterschreitungen des angegebenen Füllvolumens interessieren und $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$ testen (linksseitiger Test). Der Hersteller wird hingegen das mittlere Niveau μ möglichst genau auf dem Zielwert μ_0 halten wollen, um den gesetzlichen Vorschriften zu genügen und gleichzeitig nichts zu verschenken (weder

¹Die „Big Five“ sind Offenheit gegenüber neuen Erfahrungen (engl.: *openness to new experience*), Gewissenhaftigkeit (*conscientiousness*), Extraversion (*extraversion*), Verträglichkeit (*agreeableness*) und Neurotizismus (*neuroticism*). Sie werden nach den englischsprachigen Faktorenbezeichnungen auch mit OCEAN abgekürzt – zu Details und Modifikationen vgl. z. B. ASENDORPF / NEYER (2018).

Unter- noch Überschreitung des etikettierten Füllvolumens), d. h. er wird $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ testen (zweiseitiger Test).

Ein Test basiert auf einer **Prüfvariablen**, auch **Prüf-** oder **Teststatistik** genannt, deren Ausprägung sich im Ein-Stichprobenfall aus einer Stichprobe x_1, x_2, \dots, x_n ergibt. Letztere wird als Realisation von Stichprobenvariablen X_1, X_2, \dots, X_n interpretiert. Die Stichprobenvariablen werden nicht direkt verwendet; man aggregiert sie vielmehr anhand einer Stichprobenfunktion $g(X_1, X_2, \dots, X_n)$, z. B. anhand des Stichprobenmittelwerts \bar{X} oder der Stichprobenvarianz S^2 bzw. S^{*2} . Da die Stichprobenvariablen Zufallsvariablen sind, gilt dies auch für die Teststatistik. Die Testentscheidung hängt also von der Ausprägung $g(x_1, x_2, \dots, x_n)$ der herangezogenen Stichprobenfunktion ab.

Teststatistiken sind Zufallsvariablen

Beispiel 16.2: Anwendung von Qualitätsregelkarten

Eine effiziente und weitverbreitete Methode zur Vermeidung von Fehlern in der industriellen Massenfertigung ist die statistische Prozessregelung (engl.: *statistical process control*, kurz *SPC*) mit sog. *Qualitätsregelkarten*. Deren Anwendung entspricht der wiederholten Durchführung eines Hypothesentests. Man geht meist von einem normalverteilten Qualitätsmerkmal aus – z. B. der Länge oder dem Durchmesser eines Serienteils – und überwacht fortlaufend durch regelmäßige Entnahme von Stichproben, ob sich das mittlere Niveau und die Streuung des Merkmals während der Produktion in unerwünschter Weise verändern. Für den Lageparameter μ gibt es i. d. R. einen aus Designvorgaben resultierenden Sollwert μ_0 .

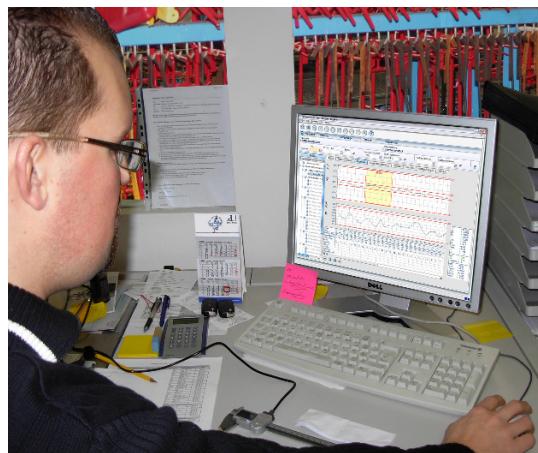


Abb. 16.2: Anwendung einer Qualitätsregelkarte (Serienproduktion von Platinen; Quelle: Fa. Böhme und Weihs Systemtechnik)

Abbildung 16.2 zeigt eine solche Qualitätsregelkarte, die hier zur Überwachung der Länge von Platinen eingesetzt wird. Auf dem Bildschirm sind zwei Zeitreihen dargestellt. Die obere Grafik zeigt die beobachteten Stichprobenmittelwerte.

Sobald ein Wert \bar{x} eine untere Linie unter- bzw. eine obere Linie überschreitet, erfolgt ein prozesskorrigierender Eingriff. Die beiden Linien werden in der Qualitätssicherung *Eingriffsgrenzen* genannt. Die untere Grafik dient der Überwachung der Prozessstreuung.

- Test für den Erwartungswert:
- Die im vorstehenden Beispiel beschriebene Vorgehensweise entspricht der wiederholten Durchführung eines Tests. Ausgangspunkt ist eine Zufallsvariable X , die im Folgenden als exakt oder zumindest approximativ normalverteilt angenommen wird mit Erwartungswert $\mu = E(X)$ und Varianz $\sigma^2 = V(X)$. Getestet wird im zweiseitigen Fall, ob das mittlere Niveau $\mu = E(X)$ von X mit einem Zielwert μ_0 übereinstimmt oder nicht. Der Test bezieht sich demnach auf die Hypothesen
- zweiseitiger Fall

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0. \quad (16.1)$$

- einseitiger Fall
- Beim einseitigen Hypothesentest für den Erwartungswert μ bezieht sich die Nullhypothese nicht nur auf einen einzigen Wert, sondern auf alle Werte unterhalb oder oberhalb eines bestimmten Schwellenwertes μ_0 . Man testet beim *rechtsseitigen* Test

$$H_0 : \mu \leq \mu_0 \quad \text{gegen} \quad H_1 : \mu > \mu_0 \quad (16.2)$$

und im *linksseitigen* Fall

$$H_0 : \mu \geq \mu_0 \quad \text{gegen} \quad H_1 : \mu < \mu_0. \quad (16.3)$$

Um einen zwei- oder einseitigen Test durchführen zu können, benötigt man Daten aus einer Stichprobe x_1, \dots, x_n . In Beispiel 16.2 wurde sie der laufenden Produktion entnommen. Die Stichprobeninformation ermöglicht es, für den unbekannten Lageparameter μ eine Schätzung $\hat{\mu}$ zu gewinnen. Als Schätzfunktion bietet sich der Stichprobenmittelwert $\hat{\mu} = \bar{X}$ an, der als Prüf- oder Testgröße fungiert. Wenn $\mu = \mu_0$ gilt, kann man die Verteilung der Prüfstatistik angeben. Aus der Kenntnis der Verteilung lässt sich ein Intervall ableiten, in das die Prüfgröße mit einer hohen Wahrscheinlichkeit $1 - \alpha$ fällt. Der Wert α ist ein vorab festzulegender Designparameter des Tests. Man wählt für α immer einen relativ kleinen Wert, z. B. $\alpha = 0,05$ oder $\alpha = 0,01$.

Für den Test der Hypothesen (16.1), (16.2) oder (16.3) kann man die Prüfgröße \bar{X} direkt verwenden. So wurde in Beispiel 16.2 verfahren. Es ist aber naheliegend, \bar{X} erst einmal gemäß

$$Z := \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} \quad (16.4)$$

zu standardisieren. Ein mit der standardisierten Prüfvariablen Z arbeitender Test wird auch als **Gauß-Test** bezeichnet.² Die Prüfvariable Z ist im Falle $\mu = \mu_0$ standardnormalverteilt, weil dann $\bar{X} \sim N(\mu_0; \sigma_{\bar{X}}^2)$ gilt mit $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$. Unter der Voraussetzung $\mu = \mu_0$ liegt demnach eine Realisation der Zufallsvariablen Z mit Wahrscheinlichkeit $1 - \alpha$ in dem durch das $\frac{\alpha}{2}$ -Quantil $z_{\alpha/2} = -z_{1-\alpha/2}$ und das $(1 - \frac{\alpha}{2})$ -Quantil $z_{1-\alpha/2}$ der Standardnormalverteilung definierten Intervall. Nur wenn die aus Stichprobendaten errechnete Ausprägung der Prüfgröße (16.4) in dieses Intervall fällt, wird beim zweiseitigen Gauß-Test weiter von der Gültigkeit der Nullhypothese H_0 ausgegangen. Andernfalls wird die Nullhypothese verworfen, also für

$$|z| > z_{1-\alpha/2}. \quad (16.5)$$

Der Ablehnbedingung (16.5) entspricht beim *rechtsseitigen* Gauß-Test

$$z > z_{1-\alpha}, \quad (16.6)$$

und beim *linksseitigen* Gauß-Test

$$z < z_{\alpha} = -z_{1-\alpha}. \quad (16.7)$$

Die durch (16.5), (16.6) bzw. (16.7) definierten Bereiche definieren den **Ablehnbereich**, die hierzu komplementären Bereiche den **Annahmebereich** für die Nullhypothese H_0 aus (16.1), (16.2) resp. (16.3). Die Werte, die die Grenze zwischen Ablehn- und Annahmebereich markieren, werden als **kritische Werte** bezeichnet. Beim zweiseitigen Gauß-Test sind dies die Quantile $z_{\alpha/2} = -z_{1-\alpha/2}$ und $z_{1-\alpha/2}$. Beim einseitigen Gauß-Test gibt es nur einen kritischen Wert, nämlich $-z_{1-\alpha}$ beim linksseitigen und $z_{1-\alpha}$ beim rechtsseitigen Test.

Ein statistischer Test führt entweder zur Ablehnung der Nullhypothese H_0 (Entscheidung für H_1) oder zur Nicht-Verwerfung von H_0 (Beibehaltung von H_0 mangels Evidenz für H_1). Jede der beiden Testentscheidungen kann richtig oder falsch sein. Es gibt somit vier denkbare Fälle: Die Nullhypothese H_0 kann zu Recht beibehalten oder zu Recht verworfen werden (korrekte Entscheidungen) oder sie wird fälschlich verworfen – diesen Fehler bezeichnet man als **Fehler 1. Art** oder α -**Fehler** – oder sie wird fälschlich nicht verworfen. Im letztgenannten Fall spricht man vom **Fehler 2. Art** oder auch β -**Fehler**. Bei der in Beispiel 16.2 beschriebenen Überwachung des mittleren Niveaus eines Qualitätsmerkmals in der Serienfertigung ist der Fehler 1. Art als blinder Alarm zu interpretieren (Eingriff in einen ungestört laufenden Fertigungsprozess), der Fehler 2. Art als unterlassener Alarm (Ausbleiben einer notwendigen Korrektur).

Ablehnbedingung für die Nullhypothese beim Gauß-Test:

– im zweiseitigen Fall

– im einseitigen Fall

Fehlerarten beim Testen

²Die Verwendung der standardisierten Prüfgröße (16.4) ist äquivalent damit, dass man für den Referenzwert μ_0 die Setzung $\mu_0 = 0$ vornimmt.

Testentscheidung	tatsächlicher Zustand	
	Nullhypothese richtig	Nullhypothese falsch
Nullhypothese wird nicht verworfen	richtige Entscheidung	Fehler 2. Art (β -Fehler)
Nullhypothese wird verworfen	Fehler 1. Art (α -Fehler)	richtige Entscheidung

Tab. 16.1: Ausgänge bei einem Hypothesentest

Tabelle 16.1 bewertet die möglichen Testausgänge. Die Wahrscheinlichkeiten für Fehlentscheidungen sind bedingte Wahrscheinlichkeiten:

$$P(\text{Fehler 1. Art}) = P(\text{Ablehnung von } H_0 | H_0 \text{ ist wahr}) \quad (16.8)$$

$$P(\text{Fehler 2. Art}) = P(\text{Nicht-Verwerfung von } H_0 | H_1 \text{ ist wahr}). \quad (16.9)$$

Veranschaulichung der Fehlerarten beim zweiseitigen Test Abbildung 16.3 zeigt die beiden Fehlerarten für den *zweiseitigen Gauß-Test*. Ein Fehler 1. Art kann nur eintreten, wenn $\mu = \mu_0$ gilt, ein Fehler 2. Art nur für $\mu \neq \mu_0$. Die Abbildung zeigt im oberen Teil die Dichte der Variablen Z für $\mu = 0$, also unter H_0 . Die Eintrittswahrscheinlichkeit für den Fehler 1. Art ist durch die beiden rot markierten Flächen repräsentiert, die jeweils den Inhalt $\frac{\alpha}{2}$ besitzen.

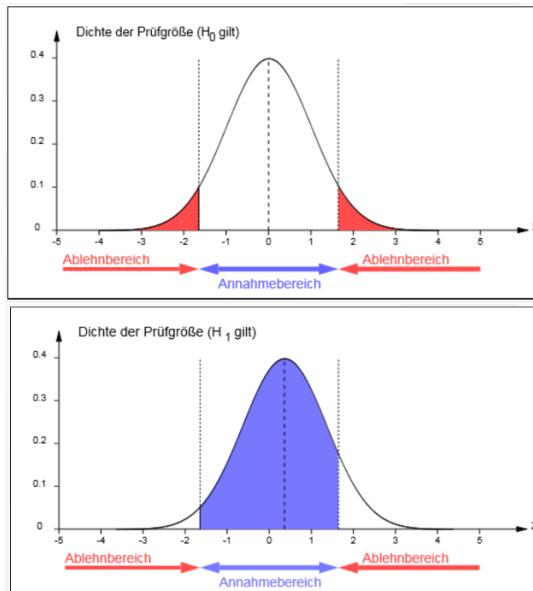


Abb. 16.3: Fehler beim Gauß-Test von $H_0 : \mu = 0$ gegen $H_1 : \mu \neq 0$; $\alpha = 0.05$. Oben: H_0 trifft zu (rot: Wahrscheinlichkeit für den Eintritt eines Fehlers 1. Art). Unten: H_1 trifft zu (blau: Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art)

Die Eintrittswahrscheinlichkeit α für den Fehler 1. Art definiert das **Signifikanzniveau** des Tests. Da hier $\alpha = 0,05$ gewählt wurde und $z_{0,975} = 1,96$ gilt, ist der Annahmebereich durch das Intervall $[-1,96; 1,96]$ gegeben. Im unteren Teil von Abbildung 16.3 ist eine Situation mit $\mu \neq 0$ gewählt, d. h. eine Situation, in der H_1 zutrifft. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art ist hier durch den Inhalt der blau markierten Fläche gegeben. In beiden Abbildungsteilen ist der Erwartungswert der standardisierten Prüfvariablen anhand einer dünnen Strichellinie gekennzeichnet. Die kritischen Werte sind anhand dünn gepunkteter Linien hervorgehoben.

Abbildung 16.4 veranschaulicht die Wahrscheinlichkeit für den Eintritt eines Fehlers 1. Art beim einseitigen Gauß-Test, wobei wir uns hier auf den *rechtsseitigen* Fall beschränken.

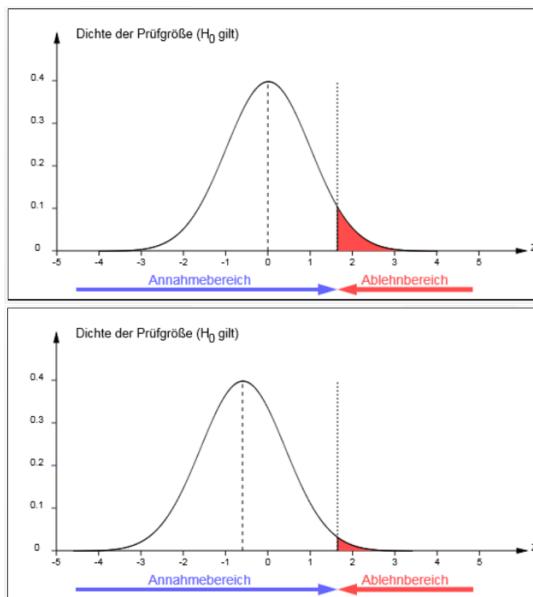


Abb. 16.4: Wahrscheinlichkeit für den Eintritt eines Fehlers 1. Art (rote Flächen) beim Gauß-Test von $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$; $\alpha = 0,05$. Oben: H_0 trifft gerade noch zu ($\mu = 0$). Unten: H_0 trifft zu ($\mu < 0$).

Im oberen Teil der Abbildung ist eine Situation dargestellt, in der H_0 gerade noch gilt. Das **Signifikanzniveau** α entspricht dem Inhalt der bei dieser Situation rot markierten Fläche. Der untere Teil von Abbildung 16.4 zeigt, dass das **Signifikanzniveau** α bei einem einseitigen Test als *obere Schranke* für den Eintritt eines Fehlers 1. Art zu interpretieren ist. Der Erwartungswert der Prüfvariablen Z und die Grenze zwischen Annahme- und Ablehnbereich sind wie in Abbildung 16.3 betont. Bei



Interaktives Objekt
„Fehler 1. und 2. Art
(zweiseitiger
Gauß-Test)“

Veranschaulichung
der Fehlerarten beim
einseitigen Test

der hier getroffenen Wahl von $\alpha = 0,05$ ist der kritische Wert durch $z_{0,95} = 1,6449$ gegeben.



Interaktives Objekt
„Fehler 1. und 2. Art
(einseitiger
Gauß-Test)“

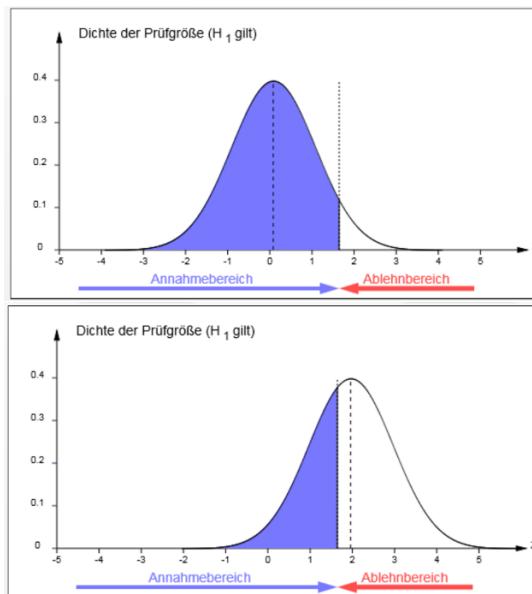


Abb. 16.5: Fehler 2. Art beim Gauß-Test von $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$; $\alpha = 0,05$. Oben: H_0 ist nur knapp verfehlt. Unten: H_0 ist deutlicher verfehlt. Blau in beiden Teilgrafiken: Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art

Es ist plausibel, dass die Wahrscheinlichkeit für die fälschliche Nicht-Verwerfung von H_0 besonders hoch ist, wenn sich μ sehr nahe am Gültigkeitsbereich von H_0 befindet (oberer Teil der Grafik). Wenn μ sich von diesem Bereich weiter entfernt, kann der Test besser zwischen H_0 und H_1 diskriminieren. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art nimmt ab (unterer Grafikteil).

16.3 Gütfunktion des Gauß-Tests

Ausgangspunkt für den Gauß-Test war ein normalverteiltes Merkmal X mit Erwartungswert μ und einer als bekannt vorausgesetzten Varianz σ^2 . Bei dem Test, bei dem der Stichprobenmittelwert \bar{X} oder die standardisierte Variable Z aus (16.4) als Prüfgröße fungiert, sollte die Wahrscheinlichkeit für die Ablehnung der Nullhypothese H_0 möglichst klein sein, falls H_0 zutrifft, und möglichst groß, wenn H_1 zutrifft. Um die Leistungsfähigkeit des Tests und insbesondere auch den Einfluss des Stichprobenumfangs n zu bewerten, zieht man die sog. **Gütfunktion**

Bewertung der Leistungsfähigkeit eines Tests

$$G(\mu) = P(\text{Ablehnung von } H_0 | \mu) \quad (16.10)$$

heran. Diese gibt *für jeden möglichen Wert* des Erwartungswerts μ des normalverteilten Merkmals X die Wahrscheinlichkeit für die Verwerfung der Nullhypothese an, spezifiziert also die Ablehnwahrscheinlichkeit für H_0 als Funktion von μ . Da $G(\mu)$ unter H_0 als Wahrscheinlichkeit für den Eintritt eines Fehlers 1. Art und $1 - G(\mu)$ für alle Werte μ im Bereich von H_1 als Wahrscheinlichkeit für das Testrisiko „Fehler 2. Art“ zu interpretieren ist, kann man anhand der Gütfunktion die Fehlerwahrscheinlichkeiten für jeden Wert μ ablesen. Von zwei mit dem Signifikanzniveau α arbeitenden Tests wird man den Test bevorzugen, dessen Gütfunktion unter H_1 einen steileren Verlauf aufweist, also geringere Wahrscheinlichkeiten für den Eintritt eines Fehlers 2. Art aufweist. Man sagt, dass dieser Test eine größere **Trennschärfe** aufweist.

Für die Gütfunktion des rechtsseitigen Gauß-Tests gilt, wie in Exkurs 16.1 gezeigt wird, die Darstellung

Gütfunktion des rechtsseitigen Tests

$$G(\mu) = 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right). \quad (16.11)$$

Dabei bezeichnet $\Phi(\cdot)$ wieder die Verteilungsfunktion der Standardnormalverteilung. Mit (16.11) lässt sich für einen beliebigen Wert μ die Wahrscheinlichkeit $G(\mu)$ für die Verwerfung der Nullhypothese berechnen, wenn α , μ_0 , σ und n vorgegeben sind. Man erkennt, dass man die Gütfunktionen auch als Funktion der relativen Abweichung $d := \frac{\mu - \mu_0}{\sigma}$ betrachten kann. Dem Wert $\mu = \mu_0$ entspricht dann $d = 0$ und $\mu = \mu_0 + \sigma$ entspricht $d = 1$. Wenn man die Gütfunktion als Funktion von d formuliert, hat dies den Vorzug, dass man von den jeweiligen Werten μ_0 und σ abstrahieren kann.

Den Verlauf der Funktion (16.11) für $\alpha = 0,05$ und für $n = 5$ sowie für $n = 10$ zeigt Abbildung 16.6. Die Grafik zeigt, dass eine Erhöhung von n für alle Werte $\mu \neq \mu_0$ zu einer Reduzierung beider Testrisiken führt. Man verifiziert insbesondere, dass die Gütfunktion für den Test mit dem

Einfluss des Stichprobenumfangs

größeren Stichprobenumfang unter H_1 einen steileren Verlauf aufweist, also trennschärfert ist.

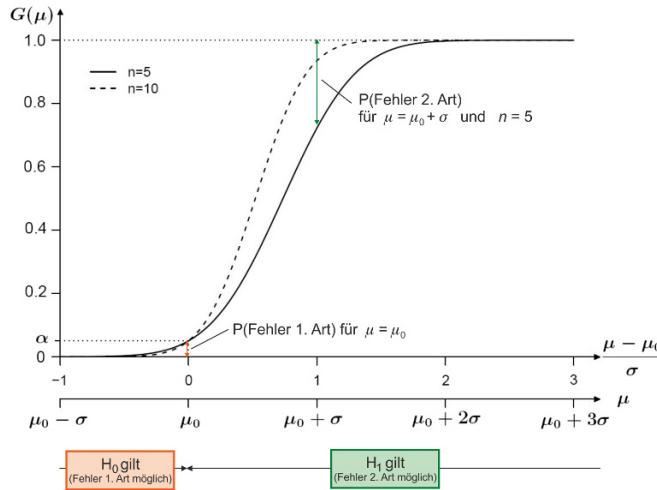


Abb. 16.6: Gütfunktion für den rechtsseitigen Gauß-Test ($\alpha = 0,05$)

Abbildung 16.6 weist auch aus, in welchem Bereich ein Fehler 1. Art oder 2. Art überhaupt möglich ist. Sie zeigt ferner, dass der Designparameter α des Tests eine in $\mu = \mu_0$ erreichte Obergrenze für die Wahrscheinlichkeit des Eintritts eines Fehlers 1. Art darstellt. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art ist in der Abbildung exemplarisch für $\mu = \mu_0 + \sigma$ und $n = 5$ anhand eines vertikalen Doppelpfeils dargestellt.

Gütfunktion des linksseitigen Tests

Für den linksseitigen Test gilt analog zum rechtsseitigen Fall, dass die Gütfunktion eine von 1 nach 0 streng *monoton fallende* Funktion ist und in μ_0 ebenfalls den Wert $G(\mu_0) = \alpha$ annimmt (vgl. Aufgabe 16.2). Ihre Formeldarstellung ist gegeben durch

$$G(\mu) = \Phi\left(-z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right). \quad (16.12)$$

Die Herleitung gleicht der Herleitung der Gütfunktion (16.11).

Gütfunktion des zweiseitigen Tests

Beim zweiseitigen Gauß-Test ist die Verwerfung der Nullhypothese eine Fehlentscheidung, die nur für $\mu = \mu_0$ und dort mit Wahrscheinlichkeit α eintreten kann. Trifft hingegen H_0 nicht zu, so sind zwei Werte μ , die gleich weit von μ_0 entfernt liegen, mit demselben Wert $G(\mu)$ verknüpft, d. h. die Gütfunktion ist *symmetrisch* bezüglich μ_0 . Sie verläuft bis μ_0 streng monoton fallend und danach streng monoton steigend. Die hier

ohne Beweis angegebene Formel lautet

$$G(\mu) = \Phi\left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right). \quad (16.13)$$



Interaktives Objekt
„Gütfunktion
(zweiseitiger
Gauß-Test mit
 $\mu_0 = 100$ “

Abbildung 16.7 zeigt, dass $G(\mu)$ für $\mu \neq \mu_0$ um so größere Werte annimmt, je weiter μ von μ_0 entfernt ist, um schließlich den Wert $G(\mu) = 1$ zu erreichen. Die Wahrscheinlichkeit $1 - G(\mu)$ für den Eintritt eines Fehlers 2. Art – in der Grafik für den Fall $\mu = \mu_0 + \sigma$ und $n = 5$ beispielhaft anhand eines langen vertikalen Pfeils veranschaulicht – nähert sich also um so mehr dem Wert 0, je weiter μ von μ_0 entfernt liegt. Auch in Abbildung 16.7 ist eine zweite Abszissenachse eingezeichnet (Darstellung der Gütfunktion als Funktion der relativen Abweichung d).

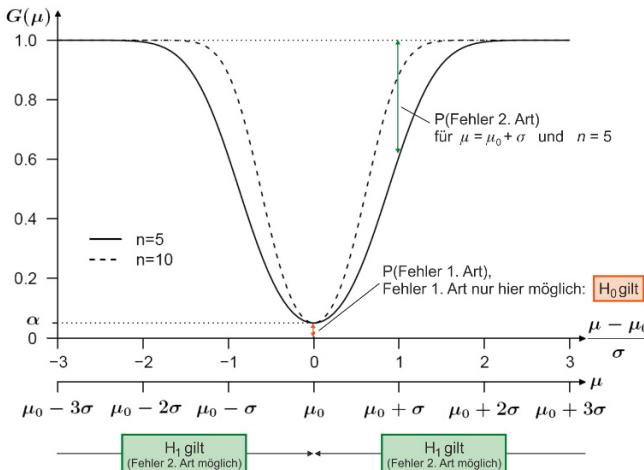


Abb. 16.7: Gütfunktion für den zweiseitigen Gauß-Test ($\alpha = 0,05$)

In der Praxis liefern Gütfunktionen eine Entscheidungshilfe, wenn man sich bei einem Test für einen Stichprobenumfang n zu entscheiden hat oder zwischen zwei mit unterschiedlichen Prüfgrößen operierenden Tests. Analysiert man die Gütfunktion eines Tests, so hat dies mit der Anwendung des Tests noch nichts zu tun. Insbesondere werden hier noch keine Stichprobendaten benötigt. Man blickt quasi aus der Vogelperspektive auf den Test und sieht, wie sensitiv er auf hypothetische Veränderungen des zu testenden Parameters reagiert.

Wozu braucht man
eine Gütfunktion?

Exkurs 16.1: Gütfunktion beim rechtsseitigen Gauß-Test

Beim rechtsseitigen Gauß-Test erfolgt die Ablehnung der Nullhypothese H_0 , wenn für die Realisation z der Prüfgröße Z aus (16.4) die Bedingung $z > z_{1-\alpha}$ erfüllt ist. Für die Gütfunktion (16.11) beinhaltet dies, dass

$$\begin{aligned} G(\mu) &= P(Z > z_{1-\alpha} | \mu) = P\left(\frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} > z_{1-\alpha} | \mu\right) \\ &= P\left(\frac{\bar{X} - \mu_0}{\sigma} \cdot \sqrt{n} > z_{1-\alpha} | \mu\right). \end{aligned}$$

Wenn man im Zähler des vor dem Ungleichheitszeichen stehenden Bruchs μ addiert und gleichzeitig subtrahiert, kann man nach einfachen Umformungen erreichen, dass der Term vor dem Ungleichheitszeichen von μ abhängt:

$$\begin{aligned} G(\mu) &= P\left(\frac{\bar{X} - \mu_0 + \mu - \mu}{\sigma} \cdot \sqrt{n} > z_{1-\alpha} | \mu\right) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} > z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n} | \mu\right) \\ &= 1 - P\left(\frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} < z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n} | \mu\right). \end{aligned}$$

Da $\frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$ standardnormalverteilt ist, folgt mit (13.20) die herzuleitende Darstellung (16.11):

$$G(\mu) = 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right).$$

16.4 Signifikanzniveau und p -Wert

- Was sagt der p -Wert aus? Es gibt eine Alternative für die Durchführung von Hypothesentests. Der Unterschied besteht hier darin, dass die Testentscheidung nicht auf dem Vergleich von Testvariablenwerten und kritischen Werten beruht, sondern auf dem Vergleich des vorgegebenen Signifikanzniveaus α mit dem sogenannten **p -Wert** α' (engl: *probability value*). Letzterer wird auch als **empirisches Signifikanzniveau** bezeichnet. Der p -Wert gibt das Niveau α' an, bei dem die Nullhypothese bei Verwendung des jeweiligen Datensatzes *gerade noch* verworfen würde. Wäre also der beim Testen verwendete Datensatz auf dem Signifikanzniveau α' getestet worden, so läge der Wert der Teststatistik am Rande des Verwerfungsbereichs. Gilt für das tatsächlich verwendete Signifikanzniveau α die Bedingung $\alpha' \leq \alpha$, ist die Nullhypothese H_0 abzulehnen, im Falle $\alpha' > \alpha$ hingegen nicht.

Man wird die Nullhypothese genau dann (in mathematischer Schreibweise: \Leftrightarrow) verwerfen, wenn der p -Wert α' den Wert α nicht überschreitet:

$$\text{Ablehnung von } H_0 \Leftrightarrow p\text{-Wert } \alpha' \text{ erfüllt die Bedingung } \alpha' \leq \alpha \quad (16.14)$$

Diese Aussage sei beispielhaft anhand eines mit $\alpha = 0,05$ arbeitenden F -Tests illustriert. Nimmt man etwa an, dass die Prüfgröße unter H_0 mit $m = 10$ und $n = 15$ Freiheitsgraden F -verteilt ist, so wird H_0 bei Überschreitung des 0,95-Quantils $F_{10;15;0,95}$ verworfen. Die Dichte der Prüfgröße und der kritische Wert $F_{10;15;0,95} = 2,544$ waren schon im linken Teil von Abbildung 13.9 dargestellt. In Abbildung 16.8 sind die Dichte und der kritische Wert erneut wiedergegeben. Dabei ist in Teil a für die Prüfgröße ein rechts von $F_{10;15;0,95}$ liegender Wert eingezeichnet. Das empirische Signifikanzniveau α' ist hier kleiner als α . Die Nullhypothese wäre hier abzulehnen. In Teil b ist hingegen für die Ausprägung der Prüfstatistik ein links von $F_{10;15;0,95}$ liegender Wert unterstellt. Der p -Wert α' ist dann größer als α und die Nullhypothese wird nicht verworfen.

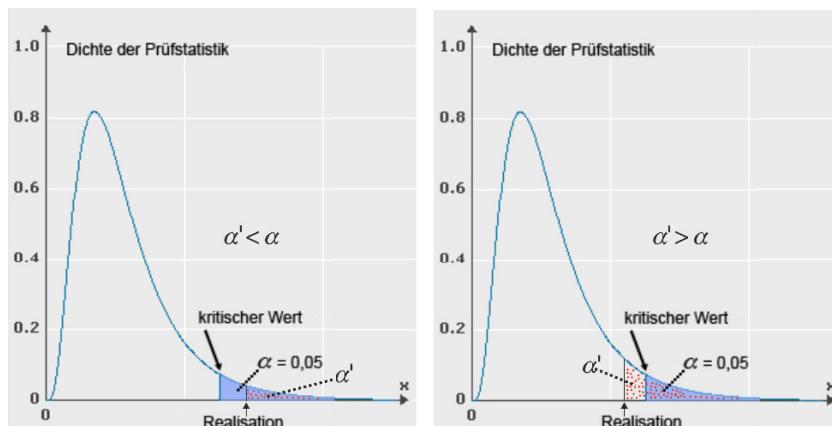


Abb. 16.8: p -Wert bei einem F -Test. Links: H_0 ist abzulehnen. Rechts: H_0 ist nicht abzulehnen

Der p -Wert wird von gängigen Statistik-Softwarepaketen, etwa SPSS, STATA oder JMP, bei Hypothesentests automatisch ausgewiesen. Wenn man testet, indem man das Signifikanzniveau vorgibt und dann den Stichprobenbefund mit von α abhängigen kritischen Werten vergleicht, gleicht dies einer Null-Eins-Entscheidung (Ablehnung oder Nicht-Ablehnung). Es spielt ja bei der Testentscheidung keine Rolle, wie weit das Stichprobenergebnis vom kritischen Wert entfernt liegt. Bei einer Testdurchführung, bei der der p -Wert α' mit dem Signifikanzniveau α verglichen wird, erhält man nuanciertere Informationen.

Beispiel 16.3: Test auf Einhaltung von Füllgewichten

In einer Fabrik wird Zucker in Tüten abgefüllt, auf denen das Füllgewicht X auf der Packung mit 2 kg angegeben ist. Aus Voruntersuchungen ist bekannt, dass X normalverteilt ist mit Standardabweichung $\sigma = 0,01$ kg. Aus einer Stichprobe von $n = 10$ Zuckertüten wurde für das Füllgewicht der Mittelwert $\bar{x} = 1,996$ kg errechnet. Es soll anhand eines statistischen Tests (16.3) mit $\mu_0 = 2$ kg eine Aussage darüber abgeleitet werden, ob der Stichprobenbefund als Indiz für eine systematische Unterschreitung des Soll-Füllgewichts angesehen werden darf. Die Wahrscheinlichkeit für das Eintreten eines Fehlers 1. Art soll den Wert $\alpha = 0,05$ nicht überschreiten. Die Ablehnung der Nullhypothese $H_0 : \mu \geq \mu_0$ erfolgt gemäß (16.7) genau dann, wenn die Ausprägung



$$z = \frac{\bar{x} - 2}{0,01} \cdot \sqrt{10} = 100 \cdot (1,996 - 2) \cdot \sqrt{10} \approx -1,2649$$

der Prüfstatistik Z aus (16.4) den Wert $z_{0,05} = -z_{0,95} = -1,6449$ unterschreitet.

Aufgabe 16.1-3

Da dies hier nicht zutrifft, kann H_0 nicht verworfen werden. Die Differenz zwischen $\bar{x} = 1,996$ kg und dem Soll-Füllgewicht von 2 kg ist also nur auf zufällige Abweichungen zurückzuführen und statistisch nicht signifikant.

Der p -Wert bezeichnet hier dasjenige Niveau α' , für das $z_{\alpha'} = -1,2649$ gilt, also $z_{1-\alpha'} = 1,2649$. Das letztgenannte Quantil ist charakterisiert durch die Gleichung $\Phi(1,2649) = 1 - \alpha'$, aus der man α' bestimmen kann. Mit Tabelle 19.2 erhält man $\Phi(1,2649) \approx 0,897$ und damit $\alpha' = 1 - 0,897 = 0,103$. Wegen $\alpha' = 0,103 > 0,05 = \alpha$ kann H_0 nicht abgelehnt werden.



Kritisch nachgefragt

In wissenschaftlichen Publikationen wird bei der Präsentation von Testergebnissen neben dem Signifikanzniveau α , für das häufig der Wert 0,05 verwendet wird, auch der p -Wert angegeben und zur Begründung der Ablehnung oder Nicht-Verwerfung der Nullhypothese herangezogen. Die sog. **Publikationsverzerrung** (engl.: *publication bias*) beschreibt die Tendenz wissenschaftlicher Journale, eher „positive“ Ergebnisse zu veröffentlichen, d. h. solche, die die Nullhypothese ablehnen, als „negative“ ohne signifikantes Ergebnis. Dies führt zu einem Druck auf Wissenschaftler, zur Rechtfertigung der eigenen Arbeit sowie der Verwendung von Forschungsgeldern aus der Industrie (Auftreten von Interessenkonflikten), statistisch signifikante Ergebnisse zu publizieren. Wissenschaftler können so bewusst oder unbewusst beeinflusst sein, Entscheidungen zu treffen, die die Wahrscheinlichkeit der Veröffentlichung erhöhen. Außerdem besteht die Gefahr, dass nicht-signifikante Ergebnisse häufiger erst gar nicht zur Veröffentlichung eingereicht werden.

Die Publikationsverzerrung kann z. B. in der Medizin dazu führen, dass die Wirksamkeit neuer Medikamente oder Therapien überschätzt wird, weil Studien mit nachgewiesener Wirksamkeit eher publiziert werden als solche, die die Wirksamkeit in Frage stellen.

16.5 *t*-Test für Erwartungswerte

Die Hypothesen (16.1), (16.2) und (16.3) beziehen sich auf den Erwartungswert eines normalverteilten Merkmals X . Die Verwendung der Prüfgröße (16.4) setzt voraus, dass die Varianz σ^2 bzw. die Standardabweichung σ von X bekannt ist. In der Praxis wird man aber meist nur auf eine Schätzung dieser Streuungsparameter zurückgreifen können. In (16.4) ist dann σ durch eine Schätzung $\hat{\sigma}$ zu ersetzen, wobei man wegen (15.9) anstelle der Stichprobenstandardabweichung S die korrigierte Stichprobenstandardabweichung $S^* = \hat{\sigma}$ wählt. Nach (14.10) ist die resultierende Prüfstatistik

$$T := \frac{\bar{X} - \mu_0}{S^*} \cdot \sqrt{n} \quad (16.15)$$

t-verteilt mit $n - 1$ Freiheitsgraden. Man kann den Annahme- und Ablehnungsbereich des mit (16.15) operierenden zweiseitigen Tests analog zu Abbildung 16.3 visualisieren, wenn man dort lediglich $z_{1-\alpha/2}$ durch das entsprechende Quantil $t_{n-1;1-\alpha/2}$ der *t*-Verteilung mit $n - 1$ Freiheitsgraden ersetzt. Da der Test mit einer *t*-verteilten Prüfstatistik arbeitet, wird er als **t-Test** angesprochen.

Abbildung 16.9 zeigt die Dichte der *t*-verteilten Variablen (16.15) mit 5 Freiheitsgraden und den Annahmebereich $[-t_{5;0,975}; t_{5;0,975}]$ des *zweiseitigen t-Tests* mit $\alpha = 0,05$. Der Wert α ist durch den Inhalt der beiden rot markierten Flächen repräsentiert. Die Dichte der Standardnormalverteilung und die Grenzen des Annahmebereichs $[-z_{0,025}; z_{0,975}]$ des analogen Gauß-Tests sind zu Vergleichszwecken ebenfalls eingezeichnet.



Interaktives Objekt
"t-Test"
(einschließlich
Vergleich mit dem
Gauß-Test)

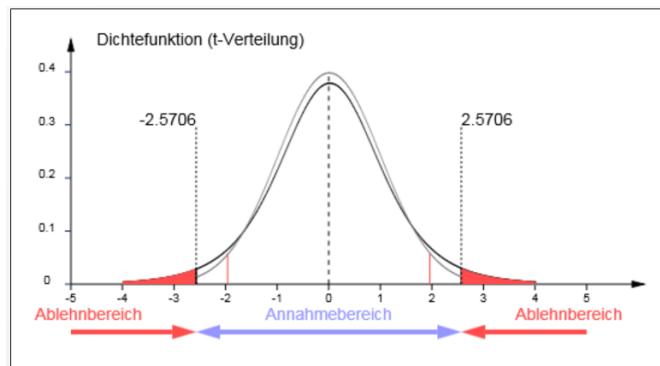


Abb. 16.9: *t*-Test von $H_0 : \mu = 0$ gegen $H_1 : \mu \neq 0$. Prüfvariable *t*-verteilt mit 5 Freiheitsgraden, $\alpha = 0,05$. Schwarze Kurve: Dichte der *t*-Verteilung mit Quantilen $t_{5;0,025}$ und $t_{5;0,975}$. Hellgrau Referenzkurve: Dichte der Standardnormalverteilung mit Quantilen $z_{0,025}$ und $z_{0,975}$ (rote vertikale Striche)

Der Annahmebereich $[-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$ des zweiseitigen t -Tests ist stets breiter als der Annahmebereich $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$ des Gauß-Tests, wobei der Unterschied mit zunehmender Anzahl von Freiheitsgraden abnimmt. Falls die Nullhypothese H_0 zutrifft, wird sie also bei dem mit der Prüfgröße (16.15) operierenden zweiseitigen Test verworfen, wenn die Prüfgröße außerhalb des in Abbildung 16.9 veranschaulichten Intervalls $[-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$ liegt, d. h. wenn

$$|t| > t_{n-1;1-\alpha/2} \quad (16.16)$$

gilt. Beim *rechtsseitigen* t -Test erfolgt die Ablehnung von H_0 im Falle

$$t > t_{n-1;1-\alpha} \quad (16.17)$$

und im *linksseitigen* Fall für

$$t < t_{n-1;\alpha} = -t_{n-1;1-\alpha}. \quad (16.18)$$

Abbildung 16.10 zeigt, wie Abbildung 16.9 im Falle des rechtsseitigen t -Tests zu modifizieren ist. Für α wurde hier der Wert 0,025 gewählt. Letzterer ist durch den Inhalt der rot markierten Fläche repräsentiert. Anstelle des $\frac{\alpha}{2}$ -Quantils und des $(1 - \frac{\alpha}{2})$ -Quantils, die in Abbildung 16.9 die Grenzen des Annahmebereichs markieren, wird der Annahmebereich beim *rechtsseitigen* Test durch das $(1 - \alpha)$ -Quantil und beim *linksseitigen* Test durch das α -Quantil vom Ablehnungsbereich getrennt.

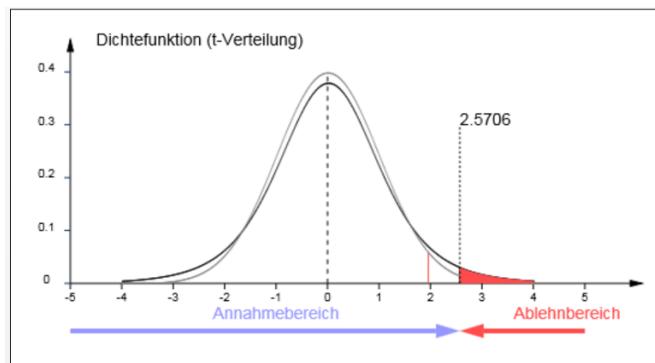


Abb. 16.10: t -Test von $H_0 : \mu \leq 0$ gegen $H_1 : \mu > 0$. Prüfvariable t -verteilt mit 5 Freiheitsgraden, $\alpha = 0,025$. Schwarze Kurve: Dichte der t -Verteilung mit Quantil $t_{5;0,975} = 2,5706$. Hellgrau Referenzkurve: Dichte der Standardnormalverteilung mit Quantil $z_{0,975} = 1,9600$ (roter vertikaler Strich)

Beispiel 16.4: Testen der Bierabfüllmenge auf dem Oktoberfest

Der „Verein gegen betrügerisches Einschenken e.V.“ kontrolliert jedes Jahr auf dem Münchener Oktoberfest eine Stichprobe von 1-Liter-Maßkrügen aus unterschiedlichen Zelten auf ihre Füllhöhe. Die Stadtverwaltung toleriert eine Abweichung von 15 Millimetern unter dem Eichstrich. Dies entspricht hier einer Unterschreitung von 100 ml – immerhin 10 % der nominellen Füllmenge. Die Münchener Boulevardzeitung *tz* hat 2013 die durchschnittlichen Füllmengen in zwölf Festzelten veröffentlicht (s. Spalte B der folgenden Tabelle). Die bittere Nachricht für den Oktoberfestbesucher lautete: „Eine 1-Liter-Maß wurde den Kontrolleuren kein einziges Mal serviert.“

Es liegt hier nahe zu testen, ob die gemessene Füllmenge im Durchschnitt wenigstens die großzügig bemessene Toleranzgrenze von $\mu_0 = 0,9$ Liter einhält oder ob sie sogar von dieser signifikant nach unten abweicht (linksseitiger Test der Hypothesen (16.3)).

Festzelt	Mittelwerte (in Liter)	Indikatorvariable
Armbrustschützen	0,87	0
Augustiner	0,86	0
Bräurosl	0,88	0
Fischer Vroni	0,86	0
Hacker Festzelt	0,92	1
Hippodrom	0,90	1
Käfer	0,80	0
Löwenbräu	0,94	1
Ochsenbraterei	0,87	0
Schottenhammel	0,81	0
Schützenzelt	0,84	0
Winzerer Fähndl	0,84	0

Tab. 16.2: Mittlere Abfüllmengen bei 12 Festzelten und Indikatorvariable

In Tabelle 16.2 sind in der letzten Spalte alle Festzelte, bei denen die mittlere Füllmenge den Mindestwert von 0,9 Litern unterschritt, mit 0, die anderen mit 1 gekennzeichnet. Man könnte die Anzahl der Ausprägungen 1 der binomialverteilten Indikatorvariablen als Testgröße verwenden. Ein solcher Test mit binomialverteilter Prüfgröße (sog. **Binomialtest**) hat den Nachteil, dass die in den Messwerten steckende Information durch die Transformation quantitativer in qualitative Daten stark reduziert wird. Es bietet sich stattdessen die Anwendung des *t*-Tests an, wobei von der Annahme normalverteilter Füllmengen auszugehen ist.

Mit Hilfe von Excel oder von *R* lässt sich der *t*-Test ohne großen Aufwand durchführen. Der *R*-Code ist in Abbildung 16.11 wiedergegeben. Man erhält für die Prüfgröße (16.15) den Wert $t = -2,871$. Die Testentscheidung führt bei Wahl von $\alpha = 0,05$ und einem Vergleich von $-2,871$ und $-t_{11;0,95} = -1,796$ gemäß (16.18) zu einer Ablehnung der Nullhypothese H_0 . Der *p*-Wert liegt bei 0,0076, was wegen $0,0076 < 0,05$ die Verwerfung von H_0 bestätigt.

```

> Oktoberfest
      Festzelt Mittelwerte Indikatorvariable
  1: Löwenbräu     0.94          1
  2: Hacker Festzelt    0.92          1
  3: Hippodrom     0.90          1
  4: Bräuros!      0.88          0
  5: Ochsenbraterei  0.87          0
  6: Armbrustschützen 0.87          0
  7: Augustiner     0.86          0
  8: Fischer Vroni   0.86          0
  9: Winzerer Fähndl 0.84          0
 10: Schützenzeit    0.84          0
 11: Schottenhammel  0.81          0
 12: Käfer          0.80          0
>
> ## linksseitiger t-Test (Alternativhypothese mu < 0) mit mu_0 = 0,9:
> t.test(Oktoberfest$Mittelwerte, alternative = "less", mu = .9)

One Sample t-test

data: Oktoberfest$Mittelwerte
t = -2.8712, df = 11, p-value = 0.007604
alternative hypothesis: true mean is less than 0.9
95 percent confidence interval:
-Inf 0.8872039
sample estimates:
mean of x
0.8658333

>
> ## Quantil der studentschen t-verteilung mit Signifikanzniveau alpha = 0,05 und
> ## 11 Freiheitsgraden (n = 12):
> qt(.05, 11)
[1] -1.795885

```

Abb. 16.11: R-Code für den t-Test auf Einhaltung von Bierabfüllmengen

Die Füllmengen der Münchner Oktoberfestwirte weichen also, statistisch gesehen, von der großzügig bemessenen Toleranzgrenze von 0,9 Litern signifikant nach unten ab.

16.6 χ^2 -Test für Varianzen

Hypothesen, Prüfgröße und Ablehnbedingungen Die Ausführungen aus Abschnitt 16.2 über das Testen zwei- und einseitiger Hypothesen für Erwartungswerte bei normalverteiltem Merkmal lassen sich leicht auf Hypothesen für Varianzen übertragen. Die zu (16.1) analogen Hypothesen für den zweiseitigen Fall lauten nun

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{gegen} \quad H_1 : \sigma^2 \neq \sigma_0^2. \quad (16.19)$$

Der Test wird durchgeführt mit der Prüfstatistik

$$T := \frac{n \cdot S^2}{\sigma_0^2} = \frac{(n-1) \cdot S^{*2}}{\sigma_0^2}. \quad (16.20)$$

Diese folgt nach (14.9) einer χ^2 -Verteilung mit $n-1$ Freiheitsgraden: $T \sim \chi_{n-1}^2$. Die Nullhypothese aus (16.19) wird bei diesem zweiseitigem **χ^2 -Test** mit Irrtumswahrscheinlichkeit α verworfen, wenn die Realisation t der Prüfgröße entweder kleiner als $\chi_{n-1;\alpha/2}^2$ oder größer als $\chi_{n-1;1-\alpha/2}^2$ ist, wenn also die berechnete Testgröße die Bedingung

$$t \notin [\chi_{n-1;\alpha/2}^2; \chi_{n-1;1-\alpha/2}^2] \quad (16.21)$$

(lies: t ist *nicht Element* des genannten Intervalls) erfüllt. Man beachte, dass die Intervallgrenzen – anders als die Grenzen des Ablehnbereichs $[t_{n-1;\alpha/2}; t_{n-1;1-\alpha/2}]$ aus Abbildung 16.9 – nicht symmetrisch zueinander liegen, weil die χ^2 -Verteilung asymmetrisch ist.

Für den einseitigen Fall hat man anstelle von (16.2) und (16.3)

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{gegen} \quad H_1 : \sigma^2 > \sigma_0^2 \quad (\text{rechtsseitiger Test}) \quad (16.22)$$

resp.

$$H_0 : \sigma^2 \geq \sigma_0^2 \quad \text{gegen} \quad H_1 : \sigma^2 < \sigma_0^2 \quad (\text{linksseitiger Test}). \quad (16.23)$$

Beim rechtsseitigen Test wird H_0 verworfen, wenn für die Realisation t der Testgröße T aus (16.20)

$$t > \chi_{n-1;1-\alpha}^2 \quad (16.24)$$

gilt. Die Ablehnbedingung für H_0 beim linksseitigen Test lautet

$$t < \chi_{n-1;\alpha}^2. \quad (16.25)$$

Die Ablehnbereiche lassen sich analog zu Abbildung 16.9 und Abbildung 16.10 veranschaulichen. Man muss nur die Dichte der χ^2 -Verteilung mit $n - 1$ Freiheitsgraden visualisieren (vgl. Abbildung 13.6) und dann in der Grafik beim zweiseitigen Test die Quantile $\chi_{n-1;\alpha/2}^2$ und $\chi_{n-1;1-\alpha/2}^2$, beim rechtsseitigen Test das Quantil $\chi_{n-1;1-\alpha}^2$ und beim linksseitigen Test das Quantil $\chi_{n-1;\alpha}^2$ auf der Abszissenachse markieren. Die Quantile sind jeweils Tabelle 19.4 des Anhangs zu entnehmen.

16.7 Zweistichproben-Tests für Erwartungswerte

Die bisher vorgestellten Tests bezogen sich auf **Einstichproben-Tests** für den Erwartungswert oder die Varianz eines normalverteilten Merkmals X . Bei den Tests für Erwartungswerte wurde unter der Voraussetzung einer bekannten Varianz der standardisierte Stichprobenmittelwert (16.4) als Prüfvariable herangezogen (**Gauß-Test**), bei geschätzter Varianz die in (16.15) eingeführte t -verteilte Teststatistik (**t-Test**).

In der Praxis hat man häufig den Fall, dass Daten für ein Merkmal vorliegen, die aus zwei Teilmengen einer Grundgesamtheit stammen. Man möchte dann prüfen, ob es bezüglich des interessierenden Merkmals eventuell Niveauunterschiede für die beiden Teilstichproben gibt. Man denke etwa an Daten zu Mathematikleistungen für Jungen und Mädchen oder an die Ergebnisse eines psychologischen Experiments, bei dem Daten in einer Versuchs- und in einer Kontrollgruppe anfallen.

Formal kann man hier die Daten als Ausprägungen zweier Zufallsvariablen X und Y interpretieren, für die zwei separate Stichproben des Umfangs n_1 bzw. n_2 vorliegen, und anhand eines **Zweistichproben-Tests** untersuchen, ob sich die Erwartungswerte $\mu_1 := E(X)$ und $\mu_2 := E(Y)$ beider Zufallsvariablen signifikant unterscheiden. Getestet wird also im hier ausschließlich betrachteten *zweiseitigen* Fall anstelle von (16.1)

$$H_0 : \mu_1 = \mu_2 \quad \text{gegen} \quad H_1 : \mu_1 \neq \mu_2. \quad (16.26)$$

Die Zufallsvariablen X und Y seien unabhängig, d. h. es wird unterstellt, dass **unabhängige Stichproben** vorliegen. Man spricht auch von **unverbundenen Stichproben**.

Wie bei den Einstichproben-Tests wird auch bei Zweistichproben-Tests für Erwartungswerte meist Normalverteilung unterstellt, also $X \sim N(\mu_1; \sigma_1^2)$ und $Y \sim N(\mu_2; \sigma_2^2)$. Man kann dann wieder zwischen den Fällen bekannter und geschätzter Varianzen σ_1^2 und σ_2^2 differenzieren. In beiden Fällen geht man bei der Konstruktion einer Prüfstatistik von der Differenz

$$D := \bar{X} - \bar{Y} \quad (16.27)$$

der Stichprobenmittelwerte aus. Nach (14.6) gilt $\bar{X} \sim N(\mu_1; \sigma_{\bar{X}}^2)$ und $\bar{Y} \sim N(\mu_2; \sigma_{\bar{Y}}^2)$. Für die Differenz D ergibt sich daraus mit (13.17) und der vorausgesetzten Unabhängigkeit der Stichproben

$$D \sim N(\mu_D; \sigma_D^2) \quad \text{mit} \quad \mu_D = \mu_1 - \mu_2; \quad \sigma_D^2 = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2. \quad (16.28)$$

Für die Varianz σ_D^2 kann man wegen (15.7) auch

$$\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (16.29)$$

Test bei Gleichheit von σ_1^2 und σ_2^2 schreiben. Bei Gültigkeit von H_0 ist $\mu_D = 0$, also $D \sim N(0; \sigma_D^2)$, so dass man unter der Voraussetzung bekannter Varianzen σ_1^2 und σ_2^2 den Test der Hypothesen (16.26) anhand der standardnormalverteilten Prüfgröße

$$Z = \frac{D}{\sigma_D} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (16.30)$$

durchführen kann. Haben die beiden Varianzen denselben Wert, etwa $\sigma^2 := \sigma_1^2 = \sigma_2^2$, vereinfacht sich (16.30) zu

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X} - \bar{Y}}{\sigma} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}.$$

Die Nullhypothese wird bei diesem **Zweistichproben-Gauß-Test** mit Irrtumswahrscheinlichkeit α verworfen, wenn $|z| > z_{1-\alpha/2}$ gilt. Dies gilt unabhängig davon, ob die Varianzen übereinstimmen oder nicht.

Bei unbekannten Varianzen σ_1^2 und σ_2^2 ist σ_D^2 zu schätzen. Die Vorgehensweise sei nur angerissen. Bezeichnet man die analog zu (14.5) definierten korrigierten Stichprobenvarianzen mit S_1^{*2} resp. S_2^{*2} , so liefert

$$\hat{\sigma}_D^2 := \frac{(n_1 - 1) \cdot S_1^{*2} + (n_2 - 1) \cdot S_2^{*2}}{(n_1 - 1) + (n_2 - 1)} \quad (16.31)$$

Test bei unbekannten Varianzen σ_1^2 und σ_2^2

eine erwartungstreue Schätzung für σ_D^2 , die die beiden Stichprobenvarianzen mit dem Umfang der Stichprobenumfänge gewichtet. Man erhält anstelle von (16.30) die Prüfstatistik

$$T = \frac{D}{\hat{\sigma}_D} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n_1 - 1) \cdot S_1^{*2} + (n_2 - 1) \cdot S_2^{*2}}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (16.32)$$

des **Zweistichproben-t-Tests**. Für die Prüfvariable (16.32) kann man bei Gleichheit der beiden Varianzen σ_1^2 und σ_2^2 zeigen, dass sie t -verteilt ist mit $n_1 + n_2 - 2$ Freiheitsgraden (vgl. MOSLER / SCHMID (2011, Abschnitt 6.2.2)). Die Nullhypothese wird dann zum Signifikanzniveau α verworfen, falls für die Realisation der Prüfgröße $|t| > t_{n_1+n_2-2; 1-\alpha/2}$ gilt.

Man spricht von **abhängigen Stichproben** oder **verbundenen Stichproben**, wenn zwei oder mehrere Messungen an denselben Merkmalsträgern vorgenommen werden. Typische Situationen, in denen verbundene Stichproben $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ auftreten, sind Vorher-Nachher-Messungen an Patienten, die eine bestimmte Therapie erhalten haben, oder die Erfassung von Abgaswerten für Fahrzeuge auf dem Prüfstand und im Realbetrieb auf der Straße. In der experimentellen Psychologie hat man verbundene Stichproben, wenn man für dieselben Personen zu zwei verschiedenen Zeitpunkten Daten erhebt, also eine Messwertwiederholung durchführt, etwa um Effekte intervenierender Maßnahmen zu verifizieren.

t-Test für
Erwartungswerte
bei verbundenen
Stichproben

Zur Analyse zweier abhängiger Messungen metrisch skalierten Merkmale dient der sogenannte **t-Test für verbundene Stichproben**. Hier stellt sich die Frage, ob sich der Mittelwert in den zwei Stichproben signifikant unterscheidet. Die Wertepaare einer jeden Messung lassen sich als Ausprägungen bivariater Zufallsvariablen $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ interpretieren, die als unabhängig identisch normalverteilt angenommen werden. Die Grundidee besteht hier darin, die Nullhypothese auf den Erwartungswert der Differenzen $\mu := E(\bar{X}_i - \bar{Y}_i)$ zu beziehen, das Testproblem also auf den Einstichproben-Fall zurückzuführen.

Beispiel 16.5: Testen der Bierabfüllmenge mit Messwiederholung

In Tabelle 16.2 waren mittlere Füllmengen von 1-Liter-Maßkrügen wiedergegeben, die in 12 Festzelten des Münchner Oktoberfests ermittelt wurden. Die Messungen wurden in der Folgewoche wiederholt. Es galt zu testen, ob sich die mittleren Abfüllmengen der beiden Wochen unterschieden. Nachstehend sind die Ergebnisse beider Messperioden wiedergegeben:

Festzelt	Woche 1	Woche 2
Armbrustschützen	0,90	0,90
Augustiner	0,86	0,86
Bräurosl	0,88	0,86
Fischer Vroni	0,83	0,88
Hacker Festzelt	0,94	0,86
Hippodrom	0,90	0,83
Käfer	0,84	0,86
Löwenbräu	0,85	0,89
Ochsenbraterei	0,88	0,95
Schottenhammel	0,85	0,88
Schützenzelt	0,87	0,8
Winzerer Fähndl	0,84	0,86

Tab. 16.3: Mittlere Abfüllmengen bei 12 Festzelten und Indikatorvariable

Die Tatsache, dass jedem Zelt jeweils zwei Messungen zugeordnet sind, lässt sich ausnutzen. Statt die Mittelwerte der beiden Stichproben miteinander zu vergleichen, werden die Differenzen der zusammengehörigen Messungen gebildet. Dann reduziert sich das Problem auf die Frage, ob sich der Mittelwert der Differenzen von 0 unterscheidet. Der t-Test für verbundene Stichproben ist demnach äquivalent zu einem Einstichproben-t-Test für den Mittelwert der Differenzen beider Stichproben. Abbildung 16.12 zeigt den R-Code für den zum Signifikanzniveau $\alpha = 0,05$ durchgeföhrten Test.

```
Paired t-test

data: Woche.1 and Woche.2
t = 0.057955, df = 11, p-value = 0.9548
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.03081459  0.03248125
sample estimates:
mean of the differences
0.0008333333
```

Abb. 16.12: R-Code für den t-Test mit Messwiederholung

Mit einem p-Wert von 0,95 können wir die Null-Hypothese bei dem gewählten Signifikanzniveau $\alpha = 0,05$ nicht ablehnen. Das bedeutet, dass die Abfüllmenge auf dem Oktoberfest sich nicht von Woche 1 zu Woche 2 unterschied. Die Zeile „mean of the differences“ im obigen R-Code weist darauf hin, dass der Test nicht mit den Rohdaten arbeitet, sondern zunächst die Differenzen gebildet werden. Das zeigt sich auch an der Zahl der Freiheitsgrade ($df = 11$). Die Freiheitsgrade

eines Tests entsprechen der um 1 verringerten Anzahl der Beobachtungen. Es gibt zwar 24 Messungen, aber nur 12 Differenzen, die in den Test eingehen.

16.8 Unabhängigkeitstests

Wenn man an n Untersuchungseinheiten, z. B. an n Personen, jeweils die Ausprägungen zweier diskreter Merkmale X und Y feststellt, hat man zwei verbundene Stichproben des Umfangs n für X und Y . In Kapitel 9 wurde die gemeinsame empirische Verteilung der beiden Merkmale in einer Kontingenztabelle zusammengefasst (vgl. Tabelle 9.1). Diese weist die Häufigkeiten h_{ij} für die möglichen Kombinationen der Ausprägungen a_1, \dots, a_k von X resp. b_1, \dots, b_m von Y aus.

Der durch (10.1) definierte **χ^2 -Koeffizient** wurde in Kapitel 10 als Zusammenhangsmaß eingeführt. In der schließenden Statistik bietet sich der χ^2 -Koeffizient als Testgröße für einen Test

$$\begin{aligned} H_0 : X \text{ und } Y \text{ sind unabhängig} &\quad \text{gegen} \\ H_1 : X \text{ und } Y \text{ sind abhängig} & \end{aligned} \tag{16.33}$$

auf Unabhängigkeit zweier diskreter Merkmale X und Y an. Dieser nicht-parametrische Test wird **χ^2 -Unabhängigkeitstest** genannt, gelegentlich auch **Kontingenztest**. Je größer der stets nicht-negative χ^2 -Koeffizient ausfällt, desto mehr spricht für die Hypothese H_1 . Man wird H_0 verwerfen, wenn die Teststatistik eine bestimmte Schranke überschreitet.

Die durch (10.1) erklärte Testvariable $T = \chi^2$ ist unter relativ schwachen Voraussetzungen unter H_0 in guter Näherung χ^2 -verteilt mit $(k-1) \cdot (m-1)$ Freiheitsgraden. Diese Aussage wird hier nicht hergeleitet und lediglich auf BÜNING / TRENKLER (1994) verwiesen, wo auch die Approximationsvoraussetzungen näher beschrieben sind. Die vorstehende Verteilungsaussage kann jedenfalls unter H_0 als gesichert angesehen werden, wenn alle Werte h_{ij} in der Kontingenztafel größer als Null sind und mindestens 80 % der Werte sogar die Bedingung $h_{ij} \geq 5$ erfüllen. Die Nullhypothese H_0 wird verworfen, wenn für die nach (10.1) errechnete Testgröße

$$\chi^2 > \chi^2_{(k-1) \cdot (m-1); 1-\alpha} \tag{16.34}$$

gilt. Das $(1 - \alpha)$ -Quantil der χ^2 -Verteilung mit $(k-1) \cdot (m-1)$ Freiheitsgraden entnimmt man Tabelle 19.4 des Anhangs.

Beispiel 16.6: Unabhängigkeitstest mit Politbarometer-Daten

In Beispiel 10.1 wurde anhand der Daten des ZDF-Politbarometers vom 8. Dezember 2017 für das Zusammenhangsmaß (10.1) der Wert $\chi^2 = 44,39$ errechnet. Da die Werte in der zugrunde liegenden Tabelle 9.9 alle deutlich größer als 5 sind, kann davon ausgegangen werden, dass die Testvariable $T = \chi^2$ unter H_0 , also bei Unabhängigkeit der Merkmale „Parteipräferenz X “ und „Geschlecht Y “, approximativ χ^2 -verteilt ist mit $(7 - 1)(2 - 1) = 6$ Freiheitsgraden. Die Nullhypothese H_0 aus (16.33) muss somit bei Vorgabe der Irrtumswahrscheinlichkeit $\alpha = 0,05$ verworfen werden, weil

$$\chi^2 = 44,39 > \chi_{5,0,95}^2 = 12,592$$

gilt, die Ablehnbedingung (16.34) also erfüllt ist. Selbst wenn man mit $\alpha = 0,01$ testet, kann H_0 abgelehnt werden. Es gilt dann

$$\chi^2 = 44,39 > \chi_{5,0,99}^2 = 16,812.$$



17 Das lineare Regressionsmodell



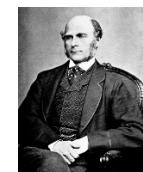
Vorschau auf
das Kapitel

Regressionsmodelle zielen darauf ab, die Werte eines Merkmals oder mehrerer Merkmale (unabhängige Variablen) zur Erklärung der Werte eines anderen Merkmals (abhängige Variable, Zielvariable) heranzuziehen. Im linearen Regressionsmodell wird der Zusammenhang über eine lineare Funktion vermittelt – im Falle nur einer unabhängigen Variablen (einfaches Regressionsmodell) also durch eine Gerade, bei mehr als einer unabhängigen Variablen (multiples Regressionsmodell) durch eine Ebene bzw. eine Hyperebene. Für die abhängige Variable wird Normalverteilung vorausgesetzt.

Die Parameter der den linearen Zusammenhang vermittelnden Regressionsfunktion können nach dem Prinzip der kleinsten Quadrate aus den Daten errechnet werden. Das an Beispielen illustrierte Verfahren beinhaltet die Minimierung der quadrierten Abstände zwischen den Datenpunkten und der Regressionsfunktion. Als Maß für die Beurteilung der Anpassungsgüte der Regressionsfunktion an die Daten wird das Bestimmtheitsmaß R^2 eingeführt. Dieses Maß stimmt beim einfachen Regressionsmodell mit dem Quadrat des Korrelationskoeffizienten r nach Bravais-Pearson überein.

Am Ende des Kapitels werden verallgemeinerte lineare Regressionsmodelle angesprochen, bei denen die abhängige Variable nicht mehr notwendigerweise normalverteilt ist. Als Beispiel aus dieser Modellklasse werden Modelle mit dichotomer Zielvariablen vorgestellt.

Sir Francis GALTON (1822 - 1911), Sohn einer Quäkerfamilie und Halbbruder von Charles DARWIN (1809 - 1882), war ein vielseitiger Naturforscher, der u. a. Wetterdaten auswertete und Klimakarten publizierte. Er sammelte Daten, um aus diesen Zusammenhangshypothesen abzuleiten und empirisch abzusichern. Seine empirischen Arbeiten sind für mehrere Wissenschaftszweige als Pionierleistungen zu bewerten. Dies gilt insbesondere für die Statistik sowie für die *Biometrie*, die sich mit der Gewinnung und Auswertung von Daten an Lebewesen befasst und ein wichtiges Anwendungsfeld der Statistik darstellt.



Sir FRANCIS
GALTON

Galton hat das heute als *Galtonbrett* bezeichnete Demonstrationsmodell für das Zustandekommen bestimmter Wahrscheinlichkeitsverteilungen hervorgebracht und auch zur Entwicklung der Regressionsanalyse beigetragen. So widmete er sich z. B. der Untersuchung eines Zusammenhangs zwischen der Körpergröße X von Eltern – er verwendete für X den Mittelwert der Körpergrößen beider Elternteile – und der Größe Y ihrer Kinder im Erwachsenenalter. Galton stellte fest, dass die beobachteten Datenpaare $(x_1; y_1), \dots, (x_n; y_n)$ um eine Gerade mit positiver Steigung streuten. Auffällig war, dass die Ausprägungen des Merkmals Y , die sich auf den gleichen Wert des Merkmals X bezogen, annähernd normalverteilt waren.

Die Varianz der Normalverteilung schien aber für verschiedene Werte von X konstant zu bleiben. Hieraus folgerte Galton, dass man zwischen beiden Merkmalen einen linearen Zusammenhang unterstellen kann, der durch nicht-systematische Zufallseinflüsse überdeckt ist. Diese Studie wird als erste Regressionsanalyse der Statistik angesehen. Die Körpergröße von Kindern wurde hier zurückgeführt (Regression = Rückbildung; Rückführung) auf die Körpergröße der Eltern.

Grundbegriffe der Regressionsanalyse

Die **Regressionsanalyse** zielt darauf ab, die Werte einer Variablen Y anhand der Werte eines Merkmals X oder auch mehrerer Merkmale X_1, \dots, X_k zu erklären, wobei der Zusammenhang über eine Funktion f modelliert wird. Letztere wird **Regressionsfunktion** genannt. Im Falle nur *eines* erklärenden Merkmals ($k = 1$) spricht man vom **einfachen Regressionsmodell**, bei Verwendung *mehrerer* erklärender Merkmale ($k \geq 2$) vom **multiplen Regressionsmodell**. In allen Fällen wird angenommen, dass der funktionale Zusammenhang nicht exakt gilt, sondern durch nicht-systematische zufällige Störeinflüsse überlagert ist. Nicht-systematisch meint, dass sich die Störungen „im Mittel“ aufheben.

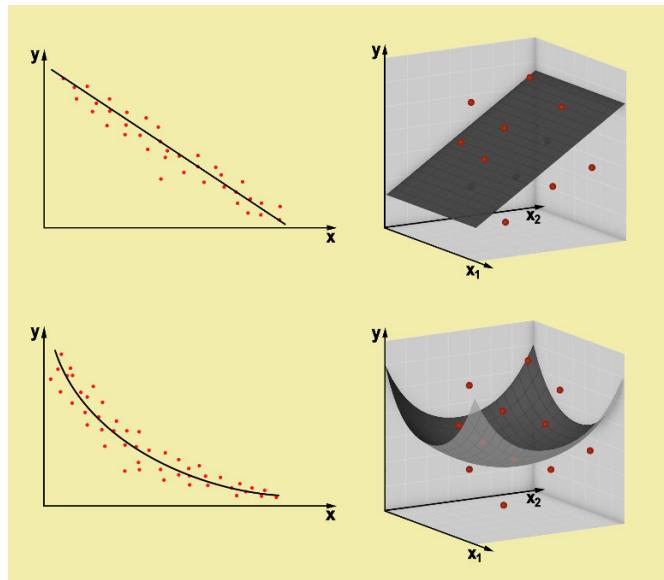


Abb. 17.1: Lineares und nicht-lineares Regressionsmodell. Oben: Lineares Regressionsmodell mit $k = 1$ bzw. mit $k = 2$ erklärenden Variablen. Unten: Nicht-lineares Regressionsmodell mit $k = 1$ bzw. mit $k = 2$ erklärenden Variablen

Abbildung 17.1 zeigt Datenpunkte in der Ebene bzw. im Raum und eine Funktion f , mit der ein funktionaler Zusammenhang zwischen einem Merkmal Y und k erklärenden Variablen modelliert wird. In der oberen Hälfte der vierteiligen Abbildung ist die Funktion f linear. Man spricht

hier man von einem **linearen Regressionsmodell**. In der unteren Hälfte der Grafik ist f hingegen nicht-linear. Die dort dargestellten Modelle sind **nicht-lineare Regressionsmodelle**.

17.1 Das einfache lineare Regressionsmodell

In diesem Manuskript wird nur das *lineare* Regressionsmodell thematisiert. Ausgangspunkt sei zunächst das **einfache lineare Regressionsmodell**. Die Regressionsfunktion f ist hier durch eine Gerade repräsentiert, die auch **Regressionsgerade** heißt. Deren Lage lässt sich anhand von Beobachtungen $(x_1, y_1), \dots, (x_n, y_n)$ für die beiden Merkmale X und Y festlegen. Wenn der lineare Zusammenhang zwischen erklärender und erklärter Variablen durch eine von Beobachtungsperiode zu Beobachtungsperiode variierende Störung überlagert wird, kann man letztere formal in jeder Periode durch eine nicht direkt beobachtbare Zufallsvariable (Störvariable) modellieren, für die sich die Ausprägung u_i einstellt. Man hat also auf der empirischen Ebene die Beziehung¹

$$y_i = \alpha + \beta x_i + u_i \quad i = 1, \dots, n. \quad (17.1)$$



Die die Lage der Geraden determinierenden Parameter α (Schnittpunkt mit der y-Achse) und β (Steigung der Geraden) heißen **Regressionskoeffizienten**. Für die Variablen X und Y werden in der Literatur verschiedene Begriffe synonym verwendet:

Modellvariable X	Modellvariable Y
erklärende Variable	erklärte Variable
unabhängige Variable	abhängige Variable
exogene Variable	endogene Variable
Regressor	Ressand

Video „Das einfache Regressionsmodell“



Tab. 17.1: Bezeichnungen für Variablen des einfachen Regressionsmodells

Falls das Merkmal X unter kontrollierten Bedingungen im Rahmen eines Experiments verändert wird, bezeichnet man es auch als Kontrollvariable, während das Merkmal Y als Ziel- oder Responsevariable angesprochen wird. Wenn Regressionsmodelle zu Prognosezwecken eingesetzt werden, nennt man die erklärende Variable auch gelegentlich **Prädiktor** oder Prädiktorvariable. In der *Psychologie* wird der Terminus „Prädiktor“ i. Allg. in der Bedeutung von „unabhängige Variable“ verwendet und für die abhängige Variable findet man hier auch den Terminus **Kriterium**.

¹Die Notation ist in der Literatur nicht ganz einheitlich. In manchen Statistik-Lehrbüchern werden für die Regressionskoeffizienten die Bezeichnungen a und b anstelle von α und β verwendet und für die Störvariable ϵ_i oder e_i statt u_i .

Modellannahmen Wenn man die Störeinflüsse u_i als Realisationen von Zufallsvariablen U_i modelliert, sind auch die Werte y_i der abhängigen Variablen Y als Ausprägungen von Zufallsvariablen Y_i zu interpretieren. Die Werte x_i der erklärenden Variablen X werden hingegen i. Allg. als determiniert modelliert, also als nicht-stochastische Größen. Mit diesen Annahmen lässt sich das einfache lineare Regressionsmodell in der Form

$$Y_i = \alpha + \beta x_i + U_i \quad i = 1, \dots, n \quad (17.2)$$

schreiben. Zu (17.2) gehören folgende Modellannahmen:

Annahmen bezüglich der Spezifikation der Regressionsfunktion:

- A1: Außer der Variablen X werden keine weiteren exogenen Variablen zur Erklärung von Y benötigt.
- A2: Die lineare Funktion, die den Zusammenhang zwischen der erklärenden Variablen X und der erklärten Variablen Y vermittelt, ist fest, d. h. die Parameter α und β sind konstant.

Annahmen bezüglich der Störvariablen:

- A3a: Die Störeinflüsse u_i sind Ausprägungen von Zufallsvariablen U_i mit Erwartungswert 0 und Varianz σ_u^2 , die im Folgenden mit σ^2 abgekürzt sei. Die Störungen sind also nicht-systematischer Natur und die Stärke der Zufallsschwankungen um die Regressionsgerade ändert sich nicht (Annahme sog. *Homoskedastizität*).
- A3b: Störvariablen U_i und U_j aus unterschiedlichen Beobachtungsperioden ($i \neq j$), sind unkorreliert (fehlende *Autokorrelation*). ²
- A3c: Die Störvariablen U_i sind normalverteilt.

Die Annahmen A3a - A3c beinhalten zusammen, dass die Störeinflüsse unabhängig identisch $N(0; \sigma^2)$ -verteilt sind:

- A3: Die Störvariablen U_i sind unabhängig identisch $N(0; \sigma^2)$ -verteilte Zufallsvariablen.

Annahmen bezüglich der unabhängigen Modellvariablen:

- A4: Die Werte der unabhängigen Variable X sind determiniert, d. h. die unabhängige Variable wird nicht als Zufallsvariable spezifiziert.
- A5: Die Variable X ist nicht konstant für $i = 1, \dots, n$ (Ausschluss eines trivialen Falls).

²Oft wird anstelle von Unkorreliertheit von Störvariablen aus verschiedenen Beobachtungsperioden die etwas stärkere Forderung stochastischer Unabhängigkeit gefordert, die nach (14.17) Unkorreliertheit impliziert.

17.2 KQ-Schätzung im einfachen Regressionsmodell

Ohne den Störterm u_i wäre die lineare Regression (17.1) eine exakte Linearbeziehung. Die Beobachtungsdaten (x_i, y_i) würden dann alle auf einer Geraden R liegen, die sich durch die Gleichung

$$y = \alpha + \beta x$$

beschreiben ließe. Diese „wahre“ Gerade ist unbekannt, d. h. die sie determinierenden Regressionskoeffizienten α und β müssen anhand der Daten geschätzt werden. Für die geschätzte Gerade wird die Notation \hat{R} verwendet und für die Geradengleichung

$$\hat{y} = \hat{\alpha} + \hat{\beta}x. \quad (17.3)$$

Zur Schätzung der Regressionskoeffizienten wird in der Praxis meist die **Methode der kleinsten Quadrate** herangezogen, kurz **KQ-Schätzung**. Bei dieser greift man auf die Abweichungen

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i \quad i = 1, \dots, n \quad (17.4)$$

zwischen dem Beobachtungswert y_i und dem Wert \hat{y}_i der Regressionsgeraden in der Beobachtungsperiode i zurück. Die Differenzen (17.4) werden **Residuen** genannt. Da diese sowohl positiv als auch negativ sein können, ist die Residuensumme kein geeignetes Kriterium für die Auswahl einer „gut“ angepassten Regressionsgeraden. Man wählt bei der KQ-Methode daher aus der Menge aller denkbaren Anpassungsgeraden diejenige Regressionsgerade \hat{R} aus, bei der die Summe der *quadrierten* Residuen \hat{u}_i^2 bezüglich der beiden Geradenparameter minimal ist:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \rightarrow \text{Min.} \quad (17.5)$$

Abbildung 17.2 veranschaulicht das Prinzip. Die Abbildung zeigt einen kleineren bivariaten Datensatz (x_i, y_i) und eine – zunächst noch nicht optimierte – Regressionsgerade. Für einen ausgewählten Datenpunkt (x_i, y_i) ist das Residuum $\hat{u}_i = y_i - \hat{y}_i$ visualisiert. Die KQ-Regressionsgerade ist dann dadurch charakterisiert, dass für sie die in (17.5) wiedergegebene Summe der Residuquadrat ein Minimum erreicht. Für die KQ-Gerade und deren Koeffizienten könnte man zur Kennzeichnung des Schätzverfahrens einen Index „KQ“ anbringen, also z. B. \hat{R}_{KQ} , $\hat{\alpha}_{KQ}$ und $\hat{\beta}_{KQ}$

Prinzip der
KQ-Schätzung

schreiben.³ Da in diesem Manuscript nur die KQ-Methode zur Schätzung von Regressionskoeffizienten verwendet wird, kann auf eine solche Kennzeichnung verzichtet werden.

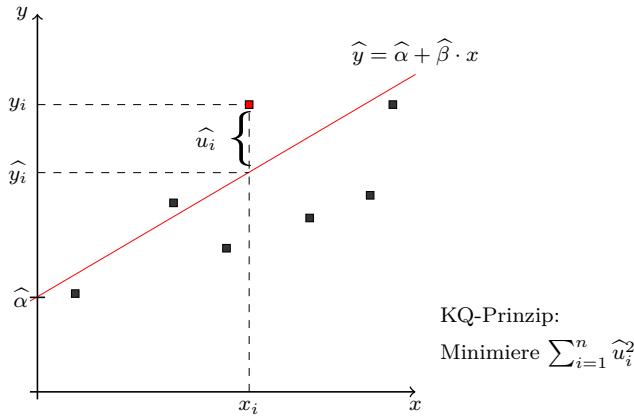


Abb. 17.2: Veranschaulichung von „Residuum“ und KQ-Methode

KQ-Schätzung
der Regressions-
koeffizienten

Um eine Formel für die KQ-Schätzungen α und β zu erhalten, muss man die Summe (17.5), deren Wert offenbar von den Geradenparametern abhängt, nach beiden Parametern einzeln differenzieren (sog. *partielle* Differentiation), anschließend die resultierenden Gleichungen Null setzen und nach α und β auflösen. Man erhält so bei Beachtung der Varianzzerlegungsformel (5.7) für die Regressionskoeffizienten β und α ⁴

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\overline{xy} - \bar{x} \bar{y}}{\overline{x^2} - \bar{x}^2} \quad (17.6)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}. \quad (17.7)$$

Eigenschaften der
KQ-Regression

Wenn man in die Gleichung (17.3) der Regressionsgeraden für x den Wert $x = \bar{x}$ einsetzt, resultiert für die abhängige Variable der Wert $\hat{y} = \hat{\alpha} + \hat{\beta} \cdot \bar{x}$, nach (17.7) also $\hat{y} = \bar{y}$. Dies bedeutet, dass die nach der KQ-Methode ermittelte Regressionsgerade stets durch den Schwerpunkt (\bar{x}, \bar{y}) des für die Schätzung herangezogenen Datensatzes $(x_1; y_1), \dots, (x_n; y_n)$ geht. Mit (17.7) kann man außerdem ableiten, dass die Summe der KQ-Residuen stets Null ist. Setzt man nämlich in (17.4) für $\hat{\alpha}$ gemäß (17.7) den Term $\bar{y} - \hat{\beta} \cdot \bar{x}$ ein, erhält man zunächst

$$\hat{u}_i = y_i - (\bar{y} - \hat{\beta} \cdot \bar{x}) - \hat{\beta} \cdot x_i = y_i - \bar{y} + \hat{\beta} \cdot \bar{x} - \hat{\beta} \cdot x_i \quad i = 1, \dots, n$$

³Wenn eine andere Schätzmethode im Spiel ist, z. B. die hier nicht thematisierte **Maximum-Likelihood-Methode**, ließe sich dies entsprechend kenntlich machen, etwa durch einen tiefgestellten Index „ML“.

⁴Bezüglich der Herleitung der beiden KQ-Schätzformeln sei auf FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 3.6.2) verwiesen.

und hieraus durch Aufsummieren der n Terme

$$\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n y_i - n\bar{y} + \hat{\beta} \cdot n\bar{x} - \hat{\beta} \cdot \sum_{i=1}^n x_i = n \cdot (\bar{y} - \bar{y} + \bar{x} - \bar{x}) = 0.$$

Dies bedeutet, dass die KQ-Schätzung fehlerausgleichend wirkt, in dem Sinne, dass sich die Abweichungen \hat{u}_i zwischen den Ordinatenwerten der Datenpunkte und denen der Regressionsgeraden herausmitteln.

Nicht nur die Regressionskoeffizienten β und α , sondern auch die in Annahme A3 eingehende Varianz $\sigma^2 := \sigma_u^2$ der Störvariablen lässt sich anhand der Beobachtungsdaten schätzen. Man verwendet hierfür die Summe der quadrierten Residuen \hat{u}_i^2 , die man noch durch $n - 2$ dividiert, weil diese Korrektur zu einer erwartungstreuen Schätzung führt. Man erhält mit (17.4) sowie mit $\hat{\beta}$ und $\hat{\alpha}$ aus (17.6) und (17.7)

$$\hat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2. \quad (17.8)$$

KQ-Schätzung der Varianz der Störvariablen

Beispiel 17.1: Berechnung von KQ-Schätzungen

Die Berechnung der KQ-Schätzwerte nach diesen Formeln sei aus didaktischen Gründen – leichte Berechenbarkeit nur mit Papier und Bleistift – anhand eines sehr kleinen Datensatzes illustriert. Der Beispieldatensatz ist einem Ökonometrielehrbuch von VON AUER (2007, dort Tabelle 3.1) entnommen und bezieht sich auf $n = 3$ Restaurantbesucher, für die die Merkmale „Rechnungsbetrag X in Euro“ und „gezahltes Trinkgeld Y in Euro“ erfasst wurden. Die Beobachtungspaare sind $(10; 2)$, $(30; 3)$ und $(50; 7)$, d. h. es ist $\bar{x} = 30$ und $\bar{y} = 4$. Es wird angenommen, dass der Modellsatz (17.1) hier anwendbar ist, die Höhe des Trinkgelds also eine durch Störeinflüsse überlagerte lineare Funktion des Rechnungsbetrags ist. Wenn man s_{xy} und s_x^2 zu Übungszwecken manuell berechnen will, empfiehlt es sich, eine Arbeitstabelle anzulegen:

i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-20	400	-2	40
2	0	0	-1	0
3	20	400	3	60
Summe:		800		100

Tab. 17.2: Arbeitstabelle für die manuelle KQ-Berechnung

Für die KQ-Schätzung $\hat{\beta}$ von β folgt wegen $s_{xy} = \frac{100}{3}$ und $s_x^2 = \frac{800}{3}$ gemäß (17.6) zunächst $\hat{\beta} = 0,125$. Hieraus resultiert mit $\bar{x} = 30$ und $\bar{y} = 4$ für $\hat{\alpha}$ gemäß (17.7) der Schätzwert $\hat{\alpha} = 0,25$. Man verifiziert durch Einsetzen von $\bar{x} = 30$, dass die nach der KQ-Methode geschätzte Regressionsgerade $\hat{y} = 0,25 + 0,125 \cdot x$ durch den Schwerpunkt $(\bar{x}; \bar{y}) = (30; 4)$ des Datensatzes verläuft.

Interpretation der Ergebnisse



Aufgabe 17.1

Der Schätzwert $\hat{\beta} = 0,125$ für den Regressionskoeffizienten β beinhaltet, dass mit jedem zusätzlichen Euro auf der Rechnung mit einer Erhöhung des Trinkgelds um 0,125 Euro zu rechnen ist. Bei einem Rechnungsbetrag in Höhe von z. B. $x = 16$ wäre der prognostizierte Wert für das Trinkgeld durch $\hat{y} = 0,25 + 0,125x = 2,25$ gegeben. Der Schätzwert $\hat{\alpha} = 0,25$ ist formal der Wert, den das Modell für $x = 0$ liefert, also bei Nicht-Konsum. Dieser Wert macht deutlich, dass das Modell, dessen Parameter auf der Basis von x -Werten zwischen $x = 10$ und $x = 50$ geschätzt wurden, nicht mehr zwangsläufig außerhalb des Stützbereichs anwendbar sein muss.

Bei größeren Datensätzen wird man stets einen Taschenrechner oder eine geeignete Statistik-Software heranziehen - z. B. SPSS, STATISTICA, JMP von SAS, STATA, EViews oder R, und so natürlich dieselben Ergebnisse erhalten. In Abbildung ?? sind ein SPSS- und darunter ein EViews-Screenshot für dieses Beispiel wiedergegeben. Die oben berechneten KQ-Schätzwerte $\hat{\beta}$ und $\hat{\alpha}$ sind bei beiden Screenshots in der zweiten Spalte zu finden. Auf die Informationen in den Folgespalten sei an dieser Stelle nicht eingegangen.

Modell	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		T	Signifikanz
	B	Standardfehler	Beta			
1 (Konstante)	,250	1,479			,169	,893
Rechnungsbetrag	,125	,043	,945		2,887	,212

Equation LINEARESMODELL Workfile: KQ-SCHÄTZUNG:Untitled

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: TRINKGELD
Method: Least Squares
Date: 05/12/17 Time: 09:51
Sample: 1 3
Included observations: 3
TRINKGELD= C(1)+C(2)*RECHNUNGSBETRAG

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.250000	1.479020	0.169031	0.8934
C(2)	0.125000	0.043301	2.886751	0.2123
R-squared	0.892857	Mean dependent var	4.000000	
Adjusted R-squared	0.785714	S.D. dependent var	2.645751	

Abb. 17.3: Computerausdruck (SPSS und EViews) zur KQ-Schätzung

Da in die KQ-Schätzungen Werte der abhängigen Variablen Y eingehen und letztere wegen des in Annahme 3a spezifizierten Zufallsvariablencharakters der Störvariablen ebenfalls als Zufallsvariable zu interpretieren ist, sind auch die aus Daten berechneten Schätzungen (17.6) und (17.7) als Ausprägungen von Zufallsvariablen zu verstehen. Will man zwischen beiden differenzieren, kann man die Zufallsvariablen als **Schätzer** oder **Schätzfunktionen** ansprechen und die aus Beobachtungsdaten errechneten Ausprägungen als **Schätzwerte**. In der Regel ist aber eine explizite

Unterscheidung nicht erforderlich, weil meist aus dem Kontext schon klar hervorgeht, welche Ebene gemeint ist.

Grundsätzlich ist die Regressionsanalyse auch im Rahmen der beschreibenden Statistik möglich, d. h. auf der empirischen Ebene ohne Rückgriff auf das Zufallsvariablenkonzept der schließenden Statistik. Nur die Einbettung der Regressionsanalyse in die schließende Statistik ermöglicht allerdings die Ableitung von Eigenschaften der Schätzungen für Parameter des Regressionsmodells. Für die KQ-Schätzfunktionen $\hat{\beta}$, $\hat{\alpha}$ und $\hat{\sigma}^2$, aus denen sich nach (17.6) - (17.8) Schätzwerte aus den Daten errechnen, lässt sich mit den hier getroffenen Annahmen ableiten, dass sie erwartungstreu sind, d. h. es gilt

$$E(\hat{\beta}) = \beta; \quad E(\hat{\alpha}) = \alpha; \quad E(\hat{\sigma}^2) = \sigma^2. \quad (17.9)$$

Es sei auf die Wiedergabe der z. T. nicht ganz einfachen und den Rahmen einer Statistik-Einführung sprengenden Beweise verzichtet und auf TOUTENBURG / HEUMANN (2008, Abschnitt 9.2.1) verwiesen.

Setzt man für die Störvariablen nach A3 Normalverteilung voraus und bezeichnet die Varianzen $V(\hat{\beta})$ und $V(\hat{\alpha})$ der Schätzer $\hat{\beta}$ resp. $\hat{\alpha}$ mit $\sigma_{\hat{\beta}}^2$ und $\sigma_{\hat{\alpha}}^2$, so gelten für die beiden Schätzer die Normalverteilungsaussagen

$$\hat{\beta} \sim N(\beta; \sigma_{\hat{\beta}}^2) \quad (17.10)$$

$$\hat{\alpha} \sim N(\alpha; \sigma_{\hat{\alpha}}^2). \quad (17.11)$$

Die Formeldarstellungen für die Varianzen seien ohne Beweis angeführt (s. hierzu z. B. FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Abschnitt 12.1)). Es gilt mit der nun mit s_x^2 bezeichneten unkorrigierten empirischen Varianz aus (5.6)

$$\sigma_{\hat{\beta}}^2 = \frac{1}{n \cdot s_x^2} \cdot \sigma^2 \quad (17.12)$$

$$\sigma_{\hat{\alpha}}^2 = \frac{\bar{x}^2}{n \cdot s_x^2} \cdot \sigma^2, \quad (17.13)$$

wobei σ^2 wieder die Varianz der Störvariablen U_i aus (17.2) bezeichnet.

17.3 Das Bestimmtheitsmaß

Hat man eine Regressionsgerade anhand eines Datensatzes $(x_1; y_1), \dots, (x_n; y_n)$ bestimmt, stellt sich die Frage, wie gut die Regressionsgerade die Variabilität der Daten erklärt. Die Summe der Residuenquadrate ist kein geeignetes Maß für die Anpassungsgüte, weil sie keine feste obere Schranke hat und zudem maßstabsabhängig ist. Man geht daher anders

Deskriptive vs.
induktive
Regressionsanalyse

Eigenschaften der
KQ-Schätzer

Eigenschaften bei
Normalverteilung

vor und zerlegt die Gesamtvarianz s_y^2 der abhängigen Variablen in die durch den Regressionsansatz erklärte Varianz $s_{\hat{y}}^2$ und eine nicht erklärte Restvarianz s_u^2 . Die Varianzen sind gemäß (5.6) definiert, also

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{s_{\hat{y}}^2} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (\hat{u}_i - \bar{u})^2}_{s_u^2}.$$

Da die Summe der n Residuen Null ist, kann man im letzten Summenterm $\bar{u} = 0$ setzen und im mittleren Summenterm $\bar{\hat{y}} = \bar{y}$. Setzt man noch anstelle von \hat{u}_i den äquivalenten Term $y_i - \hat{y}_i$ ein, folgt

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{s_{\hat{y}}^2} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2}_{s_u^2}. \quad (17.14)$$

Messung der Anpassungsgüte

Als Maß für die Anpassungsgüte eines bivariaten Datensatzes an eine Regressionsgerade wird das **Bestimmtheitsmaß** R^2 verwendet. Dieses auch gelegentlich als **Determinationskoeffizient** bezeichnete Maß vergleicht den durch die lineare Regression erklärten Varianzanteil $s_{\hat{y}}^2$ mit der Gesamtvariation s_y^2 der endogenen Variablen. Das Bestimmtheitsmaß ist also gegeben durch

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_u^2}{s_y^2}. \quad (17.15)$$

Diese Gleichung lässt sich noch in eine kürzere Form bringen. Wenn man (17.14) mit n erweitert, erhält man eine Zerlegung der Streuung in drei Summen, die mit SQ abgekürzt (Summe von Abweichungsquadrataten) und mit einem aussagekräftigen Index versehen werden:⁵

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SQ_{Total}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SQ_{Regression}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SQ_{Residual}}. \quad (17.16)$$

Anstelle von (17.15) hat man dann die äquivalente Darstellung

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}. \quad (17.17)$$

Aus der Nicht-Negativität aller Komponenten der Zerlegungen (17.14) und (17.16) folgt, dass R^2 zwischen Null und Eins liegt.

⁵In der Literatur findet man häufig auch die Abkürzungen SQT (engl: sum of squares total), SQE (sum of squares explained) und SQR (sum of squares residuals).

Abbildung 17.4 zeigt zwei Datensätze des Umfangs $n = 25$, die hieraus berechneten KQ-Regressionsgeraden und jeweils das Anpassungsgütemaß R^2 . Die Gerade in der ersten Teilgrafik liefert einen relativ hohen Erklärungsbeitrag zur Gesamtvariation der Daten (80 %), die in der zweiten Grafik hingegen nur einen schwachen Beitrag (50 %).

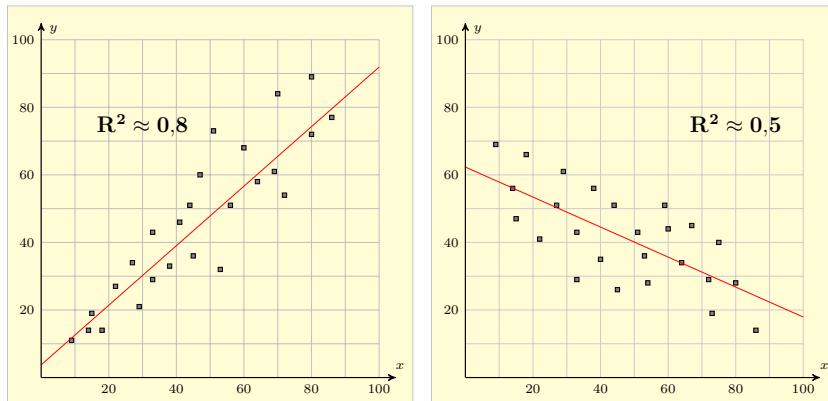


Abb. 17.4: Datensatz mit KQ-Schätzgeraden und Bestimmtheitsmaß R^2

Wenn $R^2 = 1$ gilt, ist das Modell perfekt ($s_u^2 = 0$). Im Falle $R^2 = 0$ liefert das lineare Modell keinen Erklärungsbeitrag ($s_y^2 = 0$), was aber keinesfalls ausschließt, dass zwischen den Variablen X und Y ein nichtlinearer Zusammenhang besteht.

Für die Berechnung von R^2 bietet sich eine Formel an, die direkt von den Daten ausgeht. Setzt man in (17.14) beim mittleren Summenterm für \hat{y}_i den Term $\hat{\alpha} + \hat{\beta} \cdot x_i$ und für \bar{y} den Term $\hat{\alpha} + \hat{\beta} \cdot \bar{x}$ ein, erhält man für die Varianzkomponente s_y^2 die Darstellung

$$s_y^2 = \frac{1}{n} \cdot \sum_{i=1}^n \left[(\hat{\alpha} + \hat{\beta} \cdot x_i) - (\hat{\alpha} + \hat{\beta} \cdot \bar{x}) \right]^2 = \hat{\beta}^2 s_x^2.$$

Hieraus folgt

$$R^2 = \frac{\hat{\beta} s_{xy}}{s_y^2} = \frac{(s_{xy})^2}{s_x^2 s_y^2} = r^2. \quad (17.18)$$

Im einfachen Regressionsmodell stimmt demnach das Bestimmtheitsmaß R^2 mit dem Quadrat des in (10.10) eingeführten empirischen Korrelationskoeffizienten r nach Bravais-Pearson überein.



Interaktives Objekt
„Anpassungsgüte“

Beispiel 17.2: Berechnung des Bestimmtheitsmaßes

Für den Datensatz aus Beispiel 17.1 (Trinkgeldbeträge von Restaurantbesuchern) wurden in Tabelle 17.2 für s_x^2 und s_{xy} die Werte $s_x^2 = \frac{800}{3}$ resp. $s_{xy} = \frac{100}{3}$ berechnet. Anhand der vorletzten Spalte von Tabelle 17.2 verifiziert man leicht, dass $s_y^2 = \frac{14}{3}$ ist. Mit (17.18) erhält man hieraus für das Bestimmtheitsmaß



$$R^2 = \frac{\left(\frac{100}{3}\right)^2}{\frac{800}{3} \cdot \frac{14}{3}} = \frac{25}{28} \approx 0,893.$$

Dieser Wert ist in Tabelle ?? mit etwas größerer Genauigkeit ausgewiesen.

Aufgabe 17.2

17.4 Das multiple lineare Regressionsmodell

Eine Verallgemeinerung des Modellansatzes (17.1) mit nur *einer* erklärenden Variablen ist das **multiple Regressionsmodell**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i \quad i = 1, \dots, n \quad (17.19)$$

mit k erklärenden Variablen. Man erkennt, dass (17.19) für $k = 1$ in das einfache lineare Regressionsmodell (17.1) übergeht, wenn man dort $\alpha =: \beta_0$ und $\beta =: \beta_1$ setzt.

Mit (17.19) ist ein aus n Gleichungen bestehendes Gleichungssystem gegeben – je eine Gleichung für jeden Beobachtungsindex i . Diese n Gleichungen und auch die Modellannahmen lassen sich knapper unter Verwendung der Vektor- und Matrixschreibweise formulieren.⁶ Dazu fasst man die n Werte der abhängigen Variablen und auch die n Werte der Störvariablen zu Spaltenvektoren \mathbf{y} resp. \mathbf{u} zusammen:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = (y_1, y_2, \dots, y_n)' \quad (17.20)$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = (u_1, u_2, \dots, u_n)'. \quad (17.21)$$

⁶Für die Lektüre der Abschnitte 17.4 und 17.5 werden Grundlagen der Vektor- und Matrixrechnung benötigt. Diese sind in komprimierter Form unter der Webadresse <https://mittag-statistik.de/matrizen> zu finden.

Auch die in (17.19) auftretenden $k + 1$ Koeffizienten $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ lassen sich zu einem Vektor zusammenfassen:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)' . \quad (17.22)$$

Die Werte der k Regressoren werden zu einer Matrix zusammengefasst. In die Matrix wird noch eine erste Spalte eingefügt, die nur aus Einsen besteht (Einsvektor). Dies ist ein Kunstgriff, der beinhaltet, dass man in (17.19) nach dem Koeffizienten β_0 eine Variable einsetzt, die für alle i den konstanten Wert 1 annimmt (Einfügung einer Schein- oder Dummyvariablen). Die resultierende Matrix \mathbf{X} ist eine $[n \times (k+1)]$ -Matrix, d. h. eine Matrix mit n Zeilen und $k + 1$ Spalten:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} . \quad (17.23)$$

Mit den Vektoren (17.20) - (17.22) und der Matrix (17.23) kann man die n Gleichungen (17.19) des multiplen linearen Regressionsmodells ausführlich in der Form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \quad (17.24)$$

schreiben oder kürzer als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} . \quad (17.25)$$

Fasst man die Störterme u_i aus (17.18) wieder als Realisationen von Zufallsvariablen U_i auf, so sind auch hier die Werte y_i der abhängigen Variablen Realisationen stochastischer Größen Y_i .

Modellannahmen

Spezifiziert man, wie in der Praxis üblich, die Werte $x_{i1}, x_{i2}, \dots, x_{ik}$ der unabhängigen Variablen als nicht-stochastisch, lässt sich das multiple lineare Regressionsmodell (17.19) wie folgt schreiben:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + U_i \quad i = 1, \dots, n . \quad (17.26)$$

Auch diese Gleichung lässt sich durch (17.25) kürzer darstellen, wenn man für Vektoren mit stochastischen Elementen unverändert Kleinbuchstaben verwendet, also bei der Notation auf eine Unterscheidung von Vektoren mit festen und zufälligen Elementen verzichtet.⁷

Modellannahmen Das multiple lineare Modell ist durch die folgenden Annahmen charakterisiert, die allerdings nicht immer erfüllt sein müssen und daher auf ihre Gültigkeit zu überprüfen sind:

Annahmen bezüglich der Spezifikation der Regressionsfunktion:

- MA1: Alle k erklärenden Variablen liefern einen relevanten Erklärungsbeitrag; es fehlen keine weiteren exogenen Variablen.
- MA2: Die den Zusammenhang zwischen den Regressoren X_1, X_2, \dots, X_k und der abhängigen Variablen Y vermittelnde lineare Funktion ist fest, d. h. die Parameter $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ sind konstant.

Annahmen bezüglich der Störvariablen des Regressionsmodells:

- MA3a: Die Störterme u_i des Modells sind Realisationen von Zufallsvariablen U_i mit Erwartungswert 0 und fester Varianz σ^2 , d. h. die Störeinflüsse sind nicht-systematisch und von gleich bleibender Stärke (*Homoskedastizität*).
- MA3b: Störvariablen U_i und U_j aus unterschiedlichen Beobachtungsperioden ($i \neq j$), sind unkorreliert, d. h. es gilt $Cov(U_i, U_j) = 0$ für $i \neq j$ (Fehlen von *Autokorrelation*).
- MA3c: Die Störvariablen U_i sind normalverteilt.

Die Annahmen MA3a - MA3c lassen sich wie folgt zusammenfassen:

- MA3: Die Störvariablen U_1, \dots, U_n , sind unabhängig identisch $N(0; \sigma^2)$ -verteilt.

Annahmen bezüglich der unabhängigen Modellvariablen:

- MA4: Die Werte der k unabhängigen Variablen X_1, X_2, \dots, X_k sind determiniert, d.h. die unabhängigen Variablen werden nicht als Zufallsvariablen modelliert.
- MA5: Zwischen den k Regressoren existieren keine linearen Abhängigkeiten, d. h. keine erklärende Variable lässt sich als Linearkombination anderer erklärender Variablen darstellen (Fehlen sog. *Multikollinearität*).

⁷Es werden dann z. B. sowohl der Vektor $(u_1, u_2, \dots, u_n)'$ aus (17.21) wie auch der Vektor $(U_1, U_2, \dots, U_n)'$ der in (17.26) eingehenden Zufallsvariablen mit \mathbf{u} abgekürzt. Würde man Zufallsvektoren mit Großbuchstaben kennzeichnen, hätte dies den Nachteil, dass sie fälschlich als Matrizen interpretiert werden könnten.

Wenn die Elemente der Matrix \mathbf{X} in (17.25) als nicht-stochastisch spezifiziert sind, gehen nur in \mathbf{u} und \mathbf{y} Zufallsgrößen ein, d. h. \mathbf{u} und \mathbf{y} sind Zufallsvektoren. Deren Erwartungswert wird gebildet, indem man den Erwartungswertoperator auf jedes Element des jeweiligen Vektors anwendet. Für den Erwartungswert $E(\mathbf{u})$ folgt z. B.

$$E(\mathbf{u}) = \mathbf{0} \quad (17.27)$$

gilt. Dabei bezeichnet $\mathbf{0}$ den Nullvektor, dessen Elemente nur Nullen sind. Für die Kovarianzmatrix $V(\mathbf{u})$ des Zufallsvektors \mathbf{u} erhält man mit (MA3a) und (M3b) die Darstellung

$$V(\mathbf{u}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \cdot \mathbf{I}_n. \quad (17.28)$$

Für $k = 1$ und mit $\beta_1 =: \beta$ sowie $x_1 =: x$ geht nicht nur (17.19) in (17.1) über, sondern natürlich auch (17.25). Dies gilt, weil (17.22) im Falle $k = 1$ nur aus den ersten beiden Elementen β_0 und β_1 und die Matrix (17.23) nur aus den ersten beiden Spalten besteht:

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \\ &= \begin{pmatrix} \alpha + \beta x_1 \\ \alpha + \beta x_2 \\ \vdots \\ \alpha + \beta x_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} \alpha + \beta x_1 + u_1 \\ \alpha + \beta x_2 + u_2 \\ \vdots \\ \alpha + \beta x_n + u_n \end{pmatrix}. \end{aligned} \quad (17.29)$$

Dies ist genau das Gleichungssystem (17.1). Für $k = 2$ ist (17.29) in naheliegender Weise zu modifizieren – der Vektor $\boldsymbol{\beta}$ hat dann drei Elemente $\beta_0, \beta_1, \beta_2$ und die Matrix \mathbf{X} drei Spalten.

Spezialfall $k = 1$

Exkurs 17.1: Datenimputation mittels multipler Regression

In der statistischen Praxis kommt es oft vor, dass einzelne Werte in Datensätzen fehlen oder nicht plausibel sind und deswegen nicht analysiert werden können. Manchmal ist es der einfachste Weg, solche Datensätze auszusondern. Allerdings kann dadurch wesentliche Information verloren gehen und dazu führen, dass für die Analyse nicht ausreichend Daten zur Verfügung stehen oder die verbliebenen Daten eine Verzerrung aufweisen. Letzteres kann z. B. bei Onlinebefragungen passieren, bei denen oft Antwortausfälle (engl.: *non-response*) bei Fragen auftreten, die sehr persönliche Informationen betreffen.

Ansätze zur Schließung von Datenlücken	Um die Information auch unvollständiger Datensätze möglichst gut auszuschöpfen, werden statistische Methoden zum Auffüllen von Datenlücken angewendet. Ein solches Auffüllen inkompletter Datenreihen heißt Imputation . Es gibt eine Vielzahl von Imputationsverfahren. Eine einfache Möglichkeit besteht darin, die Lücken zwischen Datenpunkten mit dem arithmetischen Mittel oder dem Median zu schließen. In manchen Situationen kann auch ein multiples lineares Regressionsmodell zur Datenimputation herangezogen werden, wie das folgende Beispiel mit Daten aus monatlichen Verkäufe von Dosengetränken im Zeitraum Januar 2016 – Juni 2019 bei belgischen Discountern zeigt. Auf der Basis der 42 Verkaufszahlen für den betrachteten Zeitraum soll eine Prognose für zukünftige Verkäufe erstellt werden. Allerdings liegen – wie im linken Teil von Abbildung 17.5 zu sehen – Lücken vor (Lücken im Sommer 2016 und im Frühling 2018).
--	--

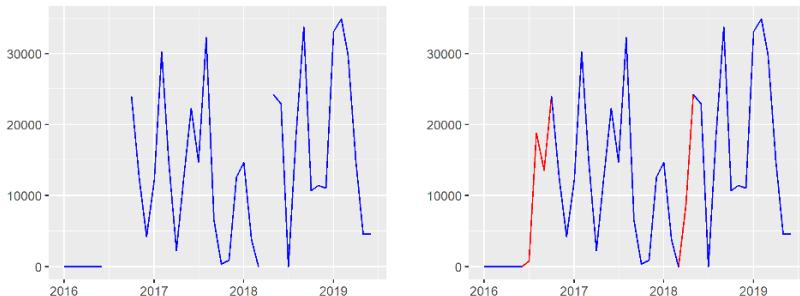


Abb. 17.5: Monatliche Verkaufszahlen von Dosengetränken in belgischen Discountern. Links: Graph der unvollständigen Zeitreihe. Rechts: Zeitreihengraph nach Imputation

Die Datenlücken können daher röhren, dass sich bei den Discountern die Zuordnungen der Produkte zu bestimmten Kategorien ändern, Hersteller aufgekauft werden oder vom Markt verschwinden. Zur Datenimputation kann hier ein lineares Regressionsmodell verwendet werden. Es lässt sich plausibel annehmen, dass die Verkäufe sowohl ein saisonales Muster als auch einen langfristigen Trend aufweisen. Ein spezielles Modell des Typs (17.19), das diese Annahmen

abbildet, besitzt die folgende Form:

$$\text{Anzahl der Verkäufe} = \beta_0 + \beta_1 \cdot \text{Datum} + \beta_2 \cdot \text{Januar} + \beta_3 \cdot \text{Februar} + \beta_4 \cdot \text{März} + \\ \beta_5 \cdot \text{April} + \beta_6 \cdot \text{Mai} + \beta_7 \cdot \text{Juni} + \beta_8 \cdot \text{Juli} + \beta_9 \cdot \text{August} + \\ \beta_{10} \cdot \text{September} + \beta_{11} \cdot \text{Oktober} + \beta_{12} \cdot \text{November}$$

Die Datumsvariable wird als numerischer Vektor verarbeitet und entspricht einer Nummerierung der 42 Beobachtungen. Mit ihr wird geschätzt, ob die Verkaufsmenge im Laufe der Zeit tendenziell zu- oder abnimmt (**Trendkomponente** des Modells). Im Gegensatz dazu wird mit den Monatsvariablen der Einfluss der Jahreszeiten auf den Absatz modelliert (**Saisonkomponente**). Mithilfe des obigen Modells werden die fehlenden Werte der obigen Zeitreihe geschätzt. Sie sind im rechten Teil von Abbildung 17.5 rot eingezzeichnet. Die bei der Schätzung verwendeten Rohdaten sind nachstehend wiedergegeben:

	volume_old	Date	Month_1	Month_2	Month_3	Month_4	Month_5	Month_6	Month_7	Month_8	Month_9	Month_10	Month_11
1:	0	2016-01-01	1	0	0	0	0	0	0	0	0	0	0
2:	0	2016-02-01	0	1	0	0	0	0	0	0	0	0	0
3:	0	2016-03-01	0	0	1	0	0	0	0	0	0	0	0
4:	0	2016-04-01	0	0	0	1	0	0	0	0	0	0	0
5:	0	2016-05-01	0	0	0	0	1	0	0	0	0	0	0
6:	0	2016-06-01	0	0	0	0	0	1	0	0	0	0	0
7:	NA	2016-07-01	0	0	0	0	0	0	1	0	0	0	0
8:	NA	2016-08-01	0	0	0	0	0	0	0	1	0	0	0
9:	NA	2016-09-01	0	0	0	0	0	0	0	0	1	0	0
10:	23990	2016-10-01	0	0	0	0	0	0	0	0	0	1	0
11:	12480	2016-11-01	0	0	0	0	0	0	0	0	0	0	1
12:	4260	2016-12-01	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13:	12200	2017-01-01	1	0	0	0	0	0	0	0	0	0	0
14:	30270	2017-02-01	0	1	0	0	0	0	0	0	0	0	0
15:	14870	2017-03-01	0	0	1	0	0	0	0	0	0	0	0
16:	2260	2017-04-01	0	0	0	1	0	0	0	0	0	0	0
17:	12520	2017-05-01	0	0	0	0	1	0	0	0	0	0	0
18:	22280	2017-06-01	0	0	0	0	0	1	0	0	0	0	0
19:	14770	2017-07-01	0	0	0	0	0	0	1	0	0	0	0
20:	32250	2017-08-01	0	0	0	0	0	0	0	1	0	0	0
21:	6590	2017-09-01	0	0	0	0	0	0	0	0	1	0	0
22:	360	2017-10-01	0	0	0	0	0	0	0	0	0	1	0
23:	960	2017-11-01	0	0	0	0	0	0	0	0	0	0	1
24:	12390	2017-12-01	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
25:	14740	2018-01-01	1	0	0	0	0	0	0	0	0	0	0
26:	3830	2018-02-01	0	1	0	0	0	0	0	0	0	0	0
27:	0	2018-03-01	0	0	1	0	0	0	0	0	0	0	0
28:	NA	2018-04-01	0	0	0	1	0	0	0	0	0	0	0
29:	24240	2018-05-01	0	0	0	0	0	1	0	0	0	0	0
30:	22930	2018-06-01	0	0	0	0	0	0	1	0	0	0	0
31:	0	2018-07-01	0	0	0	0	0	0	0	1	0	0	0
32:	18560	2018-08-01	0	0	0	0	0	0	0	0	1	0	0
33:	33760	2018-09-01	0	0	0	0	0	0	0	0	0	1	0
34:	10710	2018-10-01	0	0	0	0	0	0	0	0	0	1	0
35:	11420	2018-11-01	0	0	0	0	0	0	0	0	0	0	1
36:	11040	2018-12-01	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
37:	33090	2019-01-01	1	0	0	0	0	0	0	0	0	0	0
38:	34890	2019-02-01	0	1	0	0	0	0	0	0	0	0	0
39:	29780	2019-03-01	0	0	1	0	0	0	0	0	0	0	0
40:	15050	2019-04-01	0	0	0	1	0	0	0	0	0	0	0
41:	4590	2019-05-01	0	0	0	0	1	0	0	0	0	0	0
42:	4610	2019-06-01	0	0	0	0	0	1	0	0	0	0	0

Abb. 17.6: Rohdaten mit Verkaufsmenge, Datum und Monatsdummies

Man beachte, dass für den Dezember kein eigener Koeffizient geschätzt wird. Die Monatsvariablen sind Dummy-Variablen, die nur die Werte 0 oder 1 annehmen (1 falls der jeweilige Monat vorliegt). Wenn alle Monatsindikatorvariablen von Januar bis November den Wert 0 annehmen, muss es sich um eine Beobachtung aus dem Dezember handeln, so dass eine eigene Dezember-Variable redundante Information enthielte – das Modell wäre dann nicht mehr eindeutig schätzbar. Wollte man Verkaufszahlen für Produkte prognostizieren, bei denen der Saisoneinfluss keinen Einfluss besitzt, käme für die Datenimputation ein lineares Regressionsmodell in Betracht, das lediglich den Trend abbildet. Das obige Modell würde sich zu (17.1) mit $\alpha =: \beta_0$ und $\beta =: \beta_1$ vereinfachen.

17.5 KQ-Schätzung im multiplen Regressionsmodell

Wie beim einfachen linearen Regressionsmodell, will man auch im multiplen Fall die Regressionskoeffizienten und die Varianz der Störvariablen aus Beobachtungswerten schätzen. Während die Daten beim einfachen Regressionsmodell durch Punkte $(x_1, y_1), \dots, (x_n, y_n)$ in der Ebene repräsentiert sind, sind sie nun durch Punkte $(x_{11}, \dots, x_{1k}; y_1), \dots, (x_{n1}, \dots, x_{nk}; y_n)$ im dreidimensionalen Raum ($k = 2$) oder einem Raum höherer Ordnung gegeben ($k > 2$). Auch hier kann man die **Methode der kleinsten Quadrate**, kurz **KQ-Schätzung**, zur Schätzung von Modellparametern anwenden, wobei es nun nicht mehr um die Bestimmung einer den Daten optimal angepassten *Geraden* geht, sondern um die Bestimmung einer optimalen *Ebene* ($k = 2$) bzw. *Hyperebene* ($k > 2$).

Prinzip der KQ-Schätzung Die Grundidee der KQ-Schätzung bleibt aber unverändert. Man wählt bei der KQ-Schätzung im multiplen Regressionsmodell aus der Menge aller denkbaren Anpassungshyperebenen (bzw. Ebenen im Falle $k = 2$) diejenige aus, bei der die Summe der *quadrierten* Residuen \hat{u}_i^2 bezüglich der Regressionskoeffizienten $\beta_0, \beta_1, \dots, \beta_k$ minimal ist. Die **Residuen** für eine beliebige Regressionshyperebene sind analog zu (17.4) durch

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik} \quad (17.30)$$

definiert. Abbildung 17.7 visualisiert die Residuen für Datenpunkte im dreidimensionalen Raum ($k = 2$).

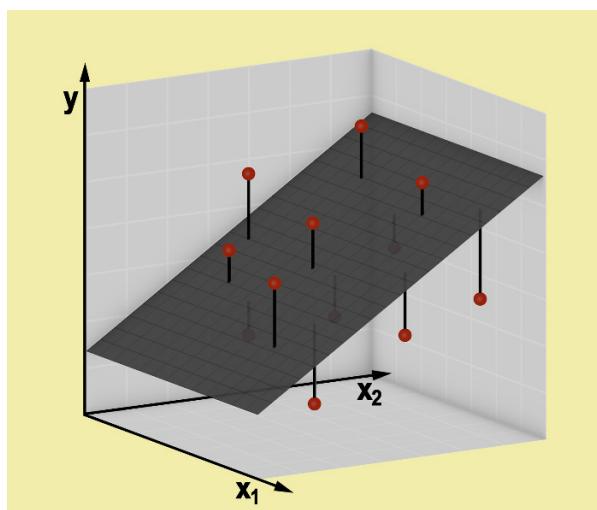


Abb. 17.7: Veranschaulichung der Residuen und der KQ-Schätzung im Modell mit zwei erklärenden Variablen

Die (17.5) entsprechende Minimierungsaufgabe lautet

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik})^2 \rightarrow \text{Min.} \quad (17.31)$$

und die nach der KQ-Methode optimale **Regressionshyperebene** ist durch

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)' \quad (17.32)$$

definiert. Fasst man die n Residuen aus (17.30) zum **Residuenvektor**

$$\hat{\mathbf{u}} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \quad (17.33)$$

zusammen, kann man (17.31) äquivalent als

$$\sum_{i=1}^n \hat{u}_i^2 = \hat{\mathbf{u}}' \hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \rightarrow \text{Min} \quad (17.34)$$

schreiben, wobei die Minimierung bezüglich aller denkbaren Vektoren $\hat{\boldsymbol{\beta}}$ von Regressionskoeffizienten erfolgt. Zur Lösung dieses Minimierungsproblems wird nach $\hat{\boldsymbol{\beta}}$ differenziert, Null gesetzt und nach $\hat{\boldsymbol{\beta}}$ aufgelöst.⁸ Dies führt zur Darstellung

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (17.35)$$

für die gesuchte KQ-Schätzung von $\boldsymbol{\beta}$. Der Vektor $\hat{\boldsymbol{\beta}}$ minimiert (17.34). Die Invertierbarkeit der Matrix $\mathbf{X}'\mathbf{X}$ ist durch die Annahme (MA5) des multiplen Regressionsmodells gesichert.

Im Spezialfall $k = 1$ und mit den Setzungen $x_1 =: x$ sowie $\hat{\beta}_0 =: \alpha$ und $\hat{\beta}_1 =: \beta$ hat \mathbf{X} die in (17.29) schon aufgeführte spezielle Gestalt einer $(n \times 2)$ -Matrix und der Vektor $\hat{\boldsymbol{\beta}}$ geht über in den zweielementigen Spaltenvektor $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\beta})'$. Wenn man diese spezielle Ausprägungen für \mathbf{X} und $\hat{\boldsymbol{\beta}}$ in (17.35) einsetzt, resultieren nach elementaren Umformungen für $\hat{\boldsymbol{\beta}}$ und $\hat{\alpha}$ die KQ-Schätzformeln (17.6) und (17.7).

Aufgabe 17.3



Spezialfall $k = 1$

⁸Eine Herleitung findet man bei TOUTENBURG / HEUMANN (2008, Abschnitt 9.3.1)

KQ-Schätzung der Varianz der Störvariablen

Die KQ-Residuen (17.31) werden wie im einfachen Regressionsmodell auch für die Schätzung der Varianz der Störvariablen U_i herangezogen. Man verwendet wieder die Summe der quadrierten Residuen \hat{u}_i^2 , die man nun noch durch $n - (k + 1)$ dividiert, um eine unverzerrte Schätzung zu erhalten. Man erhält in Verallgemeinerung von (17.8)

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - k - 1} \cdot \sum_{i=1}^n \hat{u}_i^2 \\ &= \frac{1}{n - k - 1} \cdot \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_k x_{ik})^2.\end{aligned}\quad (17.36)$$

Dabei sind $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ die Elemente des KQ-Schätzvektors aus (17.35).

Messung der Anpassungsgüte

Als Maß für die Güte der Anpassung der nach der KQ-Methode bestimmten Hyperebene an die Daten lässt sich erneut das **Bestimmtheitsmaß R^2** verwenden. Dieses ist wieder durch (17.15) bzw. (17.17) erklärt, vergleicht also die durch die KQ-Hyperebene erklärte empirische Varianz mit der Gesamtvarianz des Datensatzes. Bei perfekter Anpassung gilt $R^2 = 1$; alle Datenpunkte liegen dann auf der Hyperebene. Im Falle $R^2 = 0$ liefert das lineare Regressionsmodell keinen Erklärungsbeitrag zur Variabilität der Daten – das eventuelle Vorhandensein eines nicht-linearen Zusammenhangs zwischen den erklärenden Variablen und der erklärten Variablen ist damit nicht ausgeschlossen.

17.6 Ausblick auf verallgemeinerte Regressionsmodelle

Bei dem bisher vorgestellten einfachen oder multiplen Regressionsmodell, das einen linearen Zusammenhang zwischen einer oder mehreren erklärenden Variablen und einer erklärten Variable Y herstellt, ist die Responsevariable Y aufgrund der Modellannahmen (A3) resp. (MA3) *normalverteilt*. In der Praxis ist die Annahme einer normalverteilten erklärten Variablen aber nicht immer passend. Es gibt Beispiele für abhängige Merkmale, die zwar *stetig*, aber nicht normalverteilt sind. Denkbar ist etwa der Fall, dass Y nur positive Werte annehmen kann und anhand einer linkssteilen Verteilung zu modellieren ist (vgl. Abbildung 4.8). Hier ist das bisher behandelte Regressionsmodell nicht verwendbar. Gleiches gilt, wenn die erklärte Variable *diskret* ist, z. B. den Charakter einer Binärvariablen oder den einer Zählvariablen hat. Als Beispiel für eine diskrete Responsevariable sei das Merkmal „Beschäftigtenstatus“ genannt mit den beiden Kategorien „arbeitslos“ und „nicht-arbeitslos“ oder „Erfolg“ mit den Ausprägungen „ja / tritt ein“ und „nein / tritt nicht ein“. Codiert man die Ausprägungen der dichotomen Variablen zu „1“ und „0“ um und

Beispiele für nicht-normalverteilte Responsevariablen

bezeichnet die Eintrittswahrscheinlichkeit für die Ausprägung 1 bei der i -ten Beobachtung mit $p_i = P(Y_i = 1)$, hat man eine bernoulli-verteilte Responsevariable in Gestalt einer Null-Eins-Verteilung.

Regressionsmodelle, bei denen die Verteilung der Zielvariablen Y zu einer umfassenderen, als **Exponentialfamilie** bezeichneten Gruppe von stetigen oder diskreten Verteilungen gehört, nennt man **verallgemeinerte lineare Modelle** (engl: *generalized linear models, kurz GLMs*). Die Exponentialfamilie ist eine Verteilungsklasse, zu der neben der *Normalverteilung* u. a. die *Bernoulli-Verteilung*, die *Binomialverteilung* und die zur Modellierung seltener Ereignisse verwendete *Poisson-Verteilung* gehören. Diese Verteilungsfamilie wird in einer Monographie von FAHRMEIR / KNEIP / LANG (2009, Abschnitt 4.4), welche neben verallgemeinerten linearen Modellen auch nicht-parametrische Regressionsansätze behandelt, detaillierter charakterisiert.



Das
Regressionsmodell
mit binärer
Responsevariablen

Die folgenden Ausführungen beschränken sich auf den Fall einer bernoulli-verteilten Responsevariablen. Diese sei in jeder Beobachtungsperiode durch eine Zufallsvariable Y_i repräsentiert und deren Realisation mit y_i bezeichnet ($i = 1, 2, \dots, n$). Der Modellansatz (17.19) des „klassischen“ Regressionsmodells ist nun nicht mehr unmittelbar verwendbar. Man erkennt dies sofort, wenn man den Erwartungswert auf beiden Seiten der Gleichung (17.19) bildet. Der Erwartungswert $E(Y_i) = p_i$ der bernoulli-verteilten Variablen Y_i liegt im Intervall $[0; 1]$. Der Erwartungswert $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$ der rechten Seite von (17.19) hat hingegen einen Wert, der auch außerhalb des Intervalls $[0; 1]$ liegen kann.

Man muss daher bei binärer Responsevariablen den Modellansatz (17.19) modifizieren. Man unterzieht dazu $p_i = P(Y_i = 1)$ einer Transformation

$$p_i = h(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}) \quad (17.37)$$

mit einer streng monoton wachsenden Funktion $h(..)$, deren Werte im Intervall $[0; 1]$ liegen. Geeignet ist z. B. die Funktion

$$h(x) = \frac{\exp x}{1 + \exp x}, \quad (17.38)$$

die auch **logistische Funktion** genannt wird. Sie vermittelt den Zusammenhang zwischen dem Erwartungswert der Responsevariablen und den erklärenden Variablen. Aus (17.37) und (17.38) leitet man die folgende Modelldarstellung ab:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad (17.39)$$

Der Logarithmus des Verhältnisses der Wahrscheinlichkeit $p_i = P(Y_i = 1)$ und der Gegenwahrscheinlichkeit $1 - p_i = P(Y_i = 0)$, heißt **Logit**.

Link. Das lineare Modell (17.39) für $\log \frac{p_i}{1-p_i}$ wird **Logit-Modell** oder **logistische Regression** genannt.

Man kann auch Regressionsmodelle betrachten, bei der die abhängige Variable *mehr als zwei* Ausprägungen aufweist. Beispiele für solche Zielvariablen sind das nominalskalierte Merkmal „Transportmittel“ mit den Kategorien „Fahrrad“, „Bus“, „Bahn“ sowie „PKW“ oder das ordinalskalierte Merkmal „Gesundheitsstatus“, etwa mit den Realisationen „gesund“, „Erkrankung vom Typ 1“ und „Erkrankung vom Typ 2“. Für ein Modell mit einer derartigen Zielvariablen findet man die Bezeichnung **kategoriales Regressionsmodell**.

Exkurs 17.2: Trendvorhersage für den S&P 500-Index

Der S&P 500 ist ein Aktienindex, der die Aktienkurse der 500 größten börsennotierten US-amerikanischen Unternehmen abbildet. Wir betrachten die Tagesrenditen des Indexes im Zeitraum 2001 bis 2005. Der Datensatz heißt Smarket und ist im R-Paket ISLR zu finden.

```
> library(ISLR)
> library(data.table)
> library(magrittr)
> sm <- Smarket %>% as.data.table
> sm
   Year Lag1 Lag2 Lag3 Lag4 Lag5 volume Today Direction
1: 2001  0.381 -0.192 -2.624 -1.055  5.010 1.19130  0.959      Up
2: 2001   0.959   0.381 -0.192 -2.624 -1.055 1.29650  1.032      Up
3: 2001   1.032   0.959   0.381 -0.192 -2.624 1.41120 -0.623     Down
4: 2001  -0.623   1.032   0.959   0.381 -0.192 1.27600   0.614      Up
5: 2001   0.614  -0.623   1.032   0.959   0.381 1.20570   0.213      Up
---
1246: 2005   0.422   0.252 -0.024 -0.584 -0.285 1.88850   0.043      Up
1247: 2005   0.043   0.422   0.252 -0.024 -0.584 1.28581 -0.955     Down
1248: 2005  -0.955   0.043   0.422   0.252 -0.024 1.54047   0.130      Up
1249: 2005   0.130  -0.955   0.043   0.422   0.252 1.42236 -0.298     Down
1250: 2005  -0.298   0.130  -0.955   0.043   0.422 1.38254 -0.489     Down
> |
```

Abb. 17.8: R-Output der Daten

Der Datensatz umfasst 1 250 Beobachtungen und die nachstehenden 9 Variablen:

- **Year:** Jahr der Beobachtung.
- **Lag1, ..., Lag5:** Renditen der letzten 5 Tage (in %)
- **Volume:** Anzahl der gehandelten Aktien am Tag (in Mrd. US-Dollar)
- **Today:** Rendite des jeweiligen Tages (in %)
- **Direction:** dichotomer Richtungsindikator für den jeweiligen Tag („Up“ bzw. „Down“ bei positiver resp. negativer Tagesrendite).

Wir wollen ein Modell (17.39) mit der binären Responsevariablen „Direction“ bilden, das den Trend von S&P 500 vorhersagt. Dabei deuten wir auch an, wie man die Modellgüte prüfen kann.

Vorab visualisieren wir die Daten unter Verwendung der R-Funktion `ggpairs`. Um zu sehen, in welchem Bereich die Werte der 9 Variablen liegen, kann man Boxplots heranziehen, für die Entdeckung von Strukturen Streudiagramme und für die Veranschaulichung der Werteverteilung Histogramme. Lineare Zusammenhänge werden anhand der Korrelationskoeffizienten erfasst. Die Grafiken und die Korrelationskoeffizienten lassen sich anhand einer (9x9)-Matrix präsentieren.

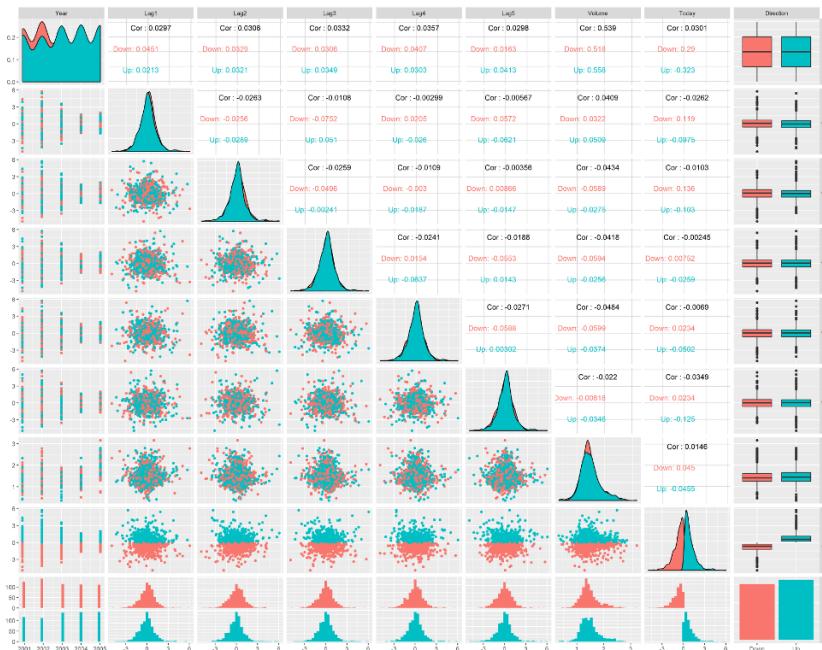


Abb. 17.9: Boxplots, Streudiagramme, Histogramme und Korrelationskoeffizienten für Daten zum S&P 500-Index. Rot: Negative Renditen („Down“); Blau: Positive Renditen („Up“)

Die Boxplots zeigen, dass die Renditen alle im Bereich von -6 bis 6 liegen. Dass die Boxplots für die Variablen $\text{Lag}1, \dots, \text{Lag}5$ sehr ähnlich aussehen, ist nicht überraschend, denn von einem zum nächsten Tag sollte es im Regelfall keine nennenswerten Veränderungen bei den Renditen geben. Die Punktwolken lassen erkennen, dass es keine Abhängigkeit zwischen den verschiedenen Renditevariablen und dem Aktienvolumen gibt. Dass keine stärkeren linearen Abhängigkeiten vorliegen, zeigen die Werte für die Korrelationskoeffizienten. Lediglich der Korrelationskoeffizient zwischen dem Aktienvolumen und den Jahren ist mit 0,539 deutlich von Null verschieden.

Die Zielvariable Y der logistischen Regression, also die Binärvariable „Direction“ mit den Ausprägungen „Up“ und „Down“, verknüpfen wir über den Logit-Link mit den erklärenden Variablen $X_1 = \text{Lag}1, \dots, X_5 = \text{Lag}5$ und $X_6 = \text{Volume}$. Um die Vorhersagegüte dieses Modells zu überprüfen, unterteilen wir die Daten in einen **Trainingsdatensatz** und in einen **Testdatensatz**. Als Testdatensatz verwenden wir die Daten des Jahres 2005, während die älteren

Beobachtungen den Trainingsdatensatz konstituieren. Anschließend schätzen wir die Koeffizienten $\beta_0, \beta_1, \dots, \beta_6$ des Modells anhand der Trainingsdaten:

```
> log.regr.train <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
+                         family = binomial,
+                         data = sm,
+                         subset = training.indices)
> summary(log.regr.train)

call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    volume, family = binomial, data = sm, subset = training.indices)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.302 -1.190  -0.079   1.160   1.350 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 0.191213  0.333690  0.573  0.567    
Lag1        -0.054178  0.051785 -1.046  0.295    
Lag2        -0.045805  0.051797 -0.884  0.377    
Lag3         0.007200  0.051644  0.139  0.889    
Lag4         0.006441  0.051706  0.125  0.901    
Lag5        -0.004223  0.051138 -0.083  0.934    
Volume     -0.116257  0.239618 -0.485  0.628    

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1383.3 on 997 degrees of freedom
Residual deviance: 1381.1 on 991 degrees of freedom
AIC: 1395.1

Number of Fisher Scoring iterations: 3
```

Abb. 17.10: Schätzung der Koeffizienten der logistischen Regression

Man erkennt, dass die Schätzwerte für die Lag-Variablen sehr nahe an Null liegen, diese Variablen somit zur Trendprognose kaum etwas beitragen. Dass das verwendete Modell nicht optimal ist, kann man auch anhand des Testdatensatzes beurteilen. Der nachstehende R-Output zeigt eine Vierfeldertafel, in der analog zu den Tabellen 11.3 – 11.5 die Häufigkeiten für korrekte Prognosen von „Up“ resp. „Down“ auf der Hauptdiagonalen und Fehlprognosen auf der Nebendiagonalen wiedergegeben sind. Für die insgesamt 252 Börsentage des Jahres 2005 gab es demnach nur $77 + 44 = 121$ korrekte Vorhersagen, was einer inakzeptablen Fehlerquote von etwa 52 % entspricht.

```
> Direction2005 <- testdata[, Direction]
> table(log.regr.fc.test, Direction2005)
  Direction2005
log.regr.fc.test Down Up
      Down    77 97
      Up      34 44
> mean(log.regr.fc.test == Direction2005)
[1] 0.4801587
```

Abb. 17.11: Häufigkeiten für korrekte und falsche Trendprognosen für S&P 500-Tagesrenditen im Jahr 2005

Es empfiehlt sich daher, ein anderes Modell unter Verwendung von Trainingsdaten zu testen, etwa eine logistische Regression mit einer reduzierten Anzahl von Regressoren, und zu bewerten, ob dieses besser abschneidet. Bei der Modellspezifikation gilt es, einen Kompromiss zu finden zwischen hoher Anpassungsgüte der Daten und zu großer Modellkomplexität. Es gibt verschiedene Kriterien zur Bewertung der Qualität einer Modellspezifikation, u. a. das in Abbildung 17.10 wiedergegebene Gütemaß **AIC** (engl.: Akaike's Information Criterion), das möglichst niedrige Werte annehmen soll. Wir verweisen diesbezüglich auf FAHRMEIR / KNEIP / LANG (2009, Abschnitt 3.6.2).



18 Grundzüge der Varianzanalyse



Vorschau auf
das Kapitel

Bei der Varianzanalyse geht man wie beim linearen Regressionsmodell von einem linearen Zusammenhang zwischen einem oder mehreren unabhängigen oder erklärenden Merkmalen und einer durch diese erklärten Variablen aus. Für die unabhängigen Variablen, die hier Faktoren genannt werden, werden aber nur wenige Ausprägungen betrachtet (Faktorstufen), d. h. die erklärenden Variablen sind als diskret spezifiziert.

Es wird zunächst ein Modell mit nur einem unabhängigem Merkmal betrachtet (*einfaktorielles Modell* der Varianzanalyse). Anhand eines Tests mit *F*-verteilter Prüfgröße (*F*-Test) wird hier untersucht, ob die Veränderung einer Faktorstufe einen Effekt auf den Erwartungswert der erklärten Variablen hat. Danach wird noch kurz das varianzanalytische Modell mit zwei erklärenden Variablen vorgestellt (*zweifaktorielles Modell*).

In Abschnitt 16.7 wurde der **Zweistichproben-t-Test** vorgestellt. Mit diesem lassen sich für zwei normalverteilte Stichproben die in (16.26) formulierten Hypothesen überprüfen, ob es bei den beiden Gruppen Unterschiede bezüglich der Erwartungswerte gibt. Für die Stichproben wurde in Abschnitt 15.5 vorausgesetzt, dass sie unabhängig sind.

Oft gilt es in der Praxis, *mehr als zwei* Gruppenmittelwerte zu vergleichen. Man denke an Studien in der *Medizin* oder der *Psychologie*, bei denen verschiedene Personengruppen unterschiedlichen Behandlungen ausgesetzt werden, etwa unterschiedlichen Medikamenten oder unterschiedlichen verhaltensbeeinflussenden Reizen. Der Vergleich von Gruppenmittelwerten ist auch eine in *Industrie* und *Technik* häufig vorkommende Aufgabe, die sich hier aber i. Allg. auf unbelebte Materie bezieht, z. B. auf Werkstoffe oder Lebensmittel, und der Optimierung von Produkten und Prozessen dient. Bei der Planung neuer Modelle im Automobilbau experimentiert man mit verschiedenen Werkstoffen, die man in planmäßig angelegten Versuchen Belastungen unterschiedlicher Intensität aussetzt. In der Werbeindustrie wird die Varianzanalyse eingesetzt zur Abschätzung des Effekts unterschiedlicher Werbeträger (u. a. Anzeigen in Printmedien, Radiospots, Werbung im Fernsehen oder im Internet) auf den Konsum.

Eine Methode, mit der sich Mittelwertvergleiche für mehr als zwei Gruppen durchführen lassen, ist die **Varianzanalyse**. Sie wurde von Sir Ronald Aylmer FISHER (1890 - 1962) begründet, der zu den führenden Statistikern des 20. Jahrhunderts zählt. Fisher trat u. a. durch Beiträge zur Schätztheorie hervor und gab der Versuchsplanung (engl.: *design of experiments*) wichtige Impulse. Weniger bekannt ist, dass *F*-Verteilung und *F*-Test nach ihm, genauer nach dem Anfangsbuchstaben seines Namens,



Sir RONALD
A. FISHER

benannt sind. Fisher war zeitweise an einer landwirtschaftlichen Versuchsstation tätig und wandte hier erstmals Modelle der Varianzanalyse an, um den Effekt von Düngemitteln auf den Ernteertrag zu untersuchen mit dem Ziel der Optimierung des Düngemitteleinsatzes. Die Varianzanalyse hatte also ihren Ausgangspunkt in den *Agrarwissenschaften*, ist aber heute fester Bestandteil des Methodenarsenals aller Wissenschaften, in denen Experimente zur Datengewinnung eingesetzt werden.

Grundbegriffe der Varianzanalyse

Die Varianzanalyse geht wie das lineare Regressionsmodell (17.2) oder (17.26) von einem linearen Zusammenhang zwischen einer Einflussgröße X oder mehreren Einflussgrößen X_1, X_2, \dots, X_k und einer zu erklärenden Variablen Y aus. Die abhängige Variable Y (Responsevariable) wird auch in der Varianzanalyse als stetig modelliert, nicht aber die Einflussgrößen. Letztere müssen in varianzanalytischen Modellen *diskret* vorliegen, d. h. es werden entweder nur bestimmte Ausprägungen einer quantitativen Variablen betrachtet oder die Einflussgrößen sind qualitative Merkmale und damit von vorneherein auf wenige Ausprägungen beschränkt. Man nennt die Einflussgrößen bei einer Varianzanalyse **Faktoren** und deren Ausprägungen **Faktorstufen**. Wenn die bei der Durchführung einer Varianzanalyse zu berücksichtigenden Faktorstufen von vorneherein festgelegt sind, spricht man von einem **Modell der Varianzanalyse mit festen Effekten**, bei einer zufallsgesteuerten Auswahl von einem **Modell der Varianzanalyse mit zufälligen Effekten**. Im Folgenden wird nur die praxisrelevantere Varianzanalyse mit festen Effekten behandelt.

Es wird zwischen **einfaktorieller Varianzanalyse** und **mehr faktorieller Varianzanalyse** unterschieden, je nachdem, ob nur *eine* Einflussgröße oder *mehrere* Einflussgrößen betrachtet werden. In der einschlägigen Literatur hat sich für die Analyse varianzanalytischer Modelle mit einer abhängigen Variablen die Abkürzung **ANOVA** (engl: *analysis of variance*) etabliert. Es gibt auch allgemeinere – im Folgenden aber nicht behandelte – Modelle der Varianzanalyse mit mehreren abhängigen Variablen. Deren Analyse ist mit der Abkürzung **MANOVA** (engl: *multivariate analysis of variance*) belegt.

Beispiel 18.1: Faktoren und Faktorstufen

Faktoren, die man bei der Analyse von Einkommensdaten heranziehen könnte, wären u. a. das nominalskalierte Merkmal „Geschlecht“ und die rangskalierte Variable „Bildungsstand“. Letztere lässt sich anhand des höchsten erreichten Bildungsabschlusses operationalisieren; die Faktorstufen sind hier durch die Bildungsabschlüsse repräsentiert.

In der Psychologie kann die einfaktorielle Varianzanalyse etwa eingesetzt werden, um Informationen zum Einfluss von Stress auf die Konzentrationsfähigkeit zu gewinnen. Stress könnte im Experiment auf unterschiedliche Weise induziert

werden, etwa über ein Dauergeräusch, durch Hitze oder durch eine andere Störquelle. Ein Beispiel aus der Psychologie zur zweifaktoriellen Varianzanalyse ist die Untersuchung der beruflichen Zufriedenheit von Lehrern (als stetig modelliert und gemessen anhand des Ergebnisses einer schriftlichen Befragung) in Abhängigkeit vom Schultyp und vom Geschlecht des Unterrichtenden. Die Einflussgrößen sind hier qualitativ.

Bei einem der agrarwissenschaftlichen Experimente von Fisher ging es um die Analyse von Düngemitteleffekten auf den Ernteertrag beim Anbau von Kartoffeln. Als Düngemittel wurden Ammonium- und Kaliumsulfat eingesetzt (zweifaktorielle Varianzanalyse), wobei jedes Düngemittel in vier unterschiedlichen Konzentrationen zum Einsatz kam (4^2 Kombinationen von Faktorstufen). Es wurden also nur vier Stufen für jede der beiden Einflussgrößen betrachtet, obwohl die Düngemittelkonzentration eigentlich eine stetige Variable darstellt. Die auch bei industriellen Anwendungen der Varianzanalyse bei quantitativen Merkmalen übliche Beschränkung auf wenige ausgewählte Faktorstufen ist zweckmäßig, weil Versuche Kosten verursachen und sich wesentliche Erkenntnisse i. Allg. schon anhand weniger Faktorstufen erreichen lassen.

18.1 Das Modell der einfaktoriellen Varianzanalyse

Es sei eine größere Grundgesamtheit betrachtet, z. B. alle Personen in Deutschland mit Bluthochdruck. Bei der einfaktoriellen Varianzanalyse geht es darum zu untersuchen, wie sich in der Grundgesamtheit die Variation *einer* Einflussgröße X auf eine Zielvariable auswirkt – bei dem genannten Beispiel etwa die Wirkung der Verabreichung eines Medikaments (Faktor) in unterschiedlichen Dosierungen (Faktorstufen) auf den Blutdruck. Für die Untersuchung wird i. Allg. schon aus Kostengründen nicht die komplette Grundgesamtheit herangezogen (Vollerhebung), sondern eine Zufallsstichprobe des Umfangs n . Diese zerlegt man in s Teilmengen des Umfangs n_i ($i = 1, 2, \dots, s$) und setzt die Elemente jeder Teilmenge einer anderen Intensität (Faktorstufe) des Einflussfaktors X aus. Von Interesse ist es dann zu untersuchen, wie sich die unterschiedliche Behandlung auf die Zielvariable Y auswirkt, hier also auf den Blutdruck.

Das univariate Modell der Varianzanalyse geht davon aus, dass die Responsevariable Y innerhalb der betrachteten Grundgesamtheit normalverteilt ist mit einer unbekannten, aber in allen Teilgesamtheiten gleichen Varianz σ^2 . Es wird insbesondere angenommen, dass die Werte von Y bei den Merkmalsträgern der Grundgesamtheit unabhängig voneinander sind (Unabhängigkeitssannahme). Die Unabhängigkeitssannahme ist z. B. verletzt, wenn an ein und demselben Merkmalsträger Messungen zu

verschiedenen Zeitpunkten durchgeführt werden.¹ Für den Erwartungswert des abhängigen Merkmals Y wird angenommen, dass er nur von der gewählten Stufe des Einflussfaktors X abhängt, also innerhalb der Teilgruppen einen festen Wert μ_i hat.

Tabelle 18.1 verdeutlicht das Design und die Basisannahmen:

Grundgesamtheit $Y \sim N(\mu; \sigma^2)$	Ziehung von Zufallsstichproben (Gesamtumfang aller Stichproben: n)
Teilgesamtheit 1	→ Stichprobe 1; Umfang n_1 : $Y \sim N(\mu_1; \sigma^2)$ mit $\mu_1 = \mu + \alpha_1$
Teilgesamtheit 2	→ Stichprobe 2; Umfang n_2 : $Y \sim N(\mu_2; \sigma^2)$ mit $\mu_2 = \mu + \alpha_2$
⋮	⋮
Teilgesamtheit s	→ Stichprobe s ; Umfang n_s : $Y \sim N(\mu_s; \sigma^2)$ mit $\mu_s = \mu + \alpha_s$

Tab. 18.1: Design einer einfaktoriellen Varianzanalyse

Die Schwankungen der Responsevariablen innerhalb der Gruppen werden wie beim Regressionsmodell durch eine Störvariable U mit $E(U) = 0$ repräsentiert. Das **Modell der einfaktoriellen Varianzanalyse** lässt sich dann in der Form

$$Y_{ik} = \mu_i + U_{ik} \quad i = 1, \dots, s; \quad k = 1, \dots, n_i \quad (18.1)$$

schreiben mit $E(U_{ik}) = 0$ und $n_1 + n_2 + \dots + n_s = n$.² Die Modelldarstellung impliziert, dass die Responsevariable in der i -ten Gruppe eine Ausprägung hat, die sich vom gruppenspezifischen Erwartungswert μ_i nur durch einen Störterm unterscheidet, der vom jeweiligen Element der Gruppe abhängt, im Mittel aber den Wert 0 aufweist.

Wenn man den Erwartungswert μ_i innerhalb der i -ten Gruppe noch additiv in eine für alle Gruppen identische Basiskomponente μ und eine gruppenspezifische Komponente α_i zerlegt, geht (18.1) über in

$$Y_{ik} = \mu + \alpha_i + U_{ik} \quad i = 1, \dots, s; \quad k = 1, \dots, n_i. \quad (18.2)$$

¹In der *Psychologie* sind bei Varianzanalysen wiederholte Messungen an Personen sehr verbreitet, etwa um Langzeitwirkungen intervenierender Maßnahmen zu untersuchen. Die hierbei verwendeten Modelle, für die die Annahme unabhängiger Merkmalswerte nicht mehr gilt, werden **varianzanalytische Modelle mit Messwiederholungen** genannt (vgl. hierzu u. a. RASCH / KUBINGER / YANAGIDA (2011, Kapitel 10)).

²Sind die Stichprobenumfänge n_i alle gleich groß, spricht man von einem **varianzanalytischen Modell mit balanciertem Design**.

Dabei ist $n_1 \cdot \alpha_1 + n_2 \cdot \alpha_2 + \dots + n_s \cdot \alpha_s = 0$, weil sich die Effekte der Faktorstufen im Mittel ausgleichen. Die Modellvariante (18.2) wird auch als **Modell der einfaktoriellen Varianzanalyse in Effektdarstellung** angesprochen. Der Term μ ist der – gelegentlich auch als *Grand Mean* bezeichnete – **globale Erwartungswert** der Responsevariablen, während α_i den Effekt der i -ten Faktorstufe auf Y widerspiegelt.

Die Varianzanalyse stellt nicht nur ein Modell zur Beurteilung der Wirkung einer oder mehrerer Faktoren auf eine metrische Responsevariable bereit. Vielmehr ermöglicht sie anhand eines Tests auch eine Entscheidung darüber, ob die Veränderung von Faktorstufen einen signifikanten Einfluss auf den Erwartungswert der Responsevariablen hat. Die Nullhypothese H_0 des Tests beinhaltet, dass die Faktorstufe keinen Effekt auf die Ausprägung der erklärten Variablen Y hat. Da aufgrund der Modellannahmen die Stichproben unabhängig normalverteilt sind mit gleicher Varianz, ist das Fehlen eines Effekts der Veränderung von Faktorstufen damit äquivalent, dass die Erwartungswerte $\mu_1, \mu_2, \dots, \mu_s$ übereinstimmen. Die Alternativhypothese H_1 , die die eigentliche Forschungshypothese repräsentiert, sagt aus, dass es mindestens eine Faktorstufenkombination (μ_i, μ_j) gibt, für die $\mu_i \neq \mu_j$ gilt. Man testet also im Falle von (18.1)

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 = \dots = \mu_s \quad \text{gegen} \\ H_1 : \mu_i &\neq \mu_j \quad \text{für mindestens ein } (i, j) \end{aligned} \tag{18.3}$$

und analog bei Zugrundelegung des Modells (18.2)

$$\begin{aligned} H_0 : \alpha_1 &= \alpha_2 = \dots = \alpha_s = 0 \quad \text{gegen} \\ H_1 : \alpha_i &\neq 0 \quad \text{und} \quad \alpha_j \neq 0 \quad \text{für mind. ein } (i, j). \end{aligned} \tag{18.4}$$

Was leistet die Varianzanalyse?

18.2 Durchführung einer einfaktoriellen Varianzanalyse

Um die Hypothesen (18.3) resp. (18.4) zu testen, benötigt man die Daten der s Zufallsstichproben. Diese kann man übersichtlich in tabellarischer Form zusammenstellen. In Tabelle 18.2 sind die Daten der Stichproben auf gerastertem Hintergrund präsentiert – jede Zeile entspricht einer Stichprobe. Die Länge der Zeilen ist nur dann gleich, wenn die s Stichproben alle denselben Umfang aufweisen. Hinter den Zeilen mit den Daten y_{ij} ist in den beiden Folgespalten noch die Summe y_i sowie der Mittelwert \bar{y}_i der Elemente der i -ten Stichprobe wiedergegeben ($i = 1, \dots, s$).

	Element-Nr der Stichprobe						Summen der Zeilen	Mittelwerte der Zeilen	
	1	2	...	k	...	n_i			
Stichprobe (Gruppe)	1	y_{11}	y_{12}	...	y_{1k}	...	y_{1,n_1}	$y_{1..}$	$\bar{y}_{1..}$
	2	y_{21}	y_{22}	...	y_{2k}	...	y_{2,n_2}	$y_{2..}$	$\bar{y}_{2..}$
	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	i	y_{i1}	y_{i2}	...	y_{ik}	...	y_{i,n_i}	$y_{i..}$	$\bar{y}_{i..}$
	...	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	s	y_{s1}	y_{s2}	...	y_{sk}	...	y_{s,n_s}	$y_{s..}$	$\bar{y}_{s..}$

Tab. 18.2: Daten bei einer einfaktoriellen Varianzanalyse

Zerlegung der Gesamtstreuung in zwei Komponenten

Bei der Herleitung einer Prüfgröße für einen Test der genannten Hypothesen wird ausgenutzt, dass sich die Streuung der n Beobachtungen aus allen s Stichproben (Gesamtstreuung) analog zur Zerlegung (17.16) in zwei Komponenten zerlegen lässt, nämlich in eine Komponente $SQ_{zwischen}$, die die Variabilität zwischen den Gruppen widerspiegelt, und eine Restkomponente $SQ_{Residual}$, die die Variation innerhalb der Stichproben repräsentiert. Die erstgenannte Komponente gibt den Streuungsanteil an, der durch das Modell erklärt wird, also durch die Veränderung von Faktorstufen hervorgerufen wird, während die zweite Komponente eine durch das Modell nicht erklärte Reststreuung darstellt. Für die beiden genannten Komponenten findet man in der Fachliteratur uneinheitliche Abkürzungen. Dieser Umstand und auch die etwas sperrige Notation (Doppelindizes für die Beobachtungsdaten bei der einfaktoriellen, Dreifachindizes bei der zweifaktoriellen Varianzanalyse) erschweren den Zugang zur Thematik.

Die Gesamtstreuung der n Werte im grau hinterlegten Inneren von Tabelle 18.2 lässt sich anhand der Summe

$$SQ_{\text{Total}} := \sum_{i=1}^s \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{..})^2 \quad (18.5)$$

aller quadrierten Abweichungen der Beobachtungswerte y_{ik} vom Gesamtmittelwert $\bar{y}_{..}$ erfassen.³ Die Quadrierung der Differenzen $y_{ik} - \bar{y}_{..}$ verhindert, dass sich positive und negative Abweichungen vom Gesamtmittel kompensieren. Der Wert $\bar{y}_{..}$ lässt sich errechnen, indem man die Summe der s Elemente der vorletzten Spalte von Tabelle 18.2 durch die Gesamtzahl n aller Beobachtungen dividiert.⁴

³ SQ steht wieder für „Summe der Abweichungsquadrate“ oder „sum of squares“.

⁴ Alternativ kann man die Werte der letzten Spalte von Tabelle 18.2 mit den jeweiligen Stichprobenumfängen gewichten und dann die Summe der gewichteten Stichprobenmittelwerte durch n teilen.

In der letzten Spalte von Tabelle 18.2 wird jede Stichprobe zum Stichprobenmittelwert verdichtet, also auf eine einzige Kenngröße heruntergebrochen. Die Information zur Streuung innerhalb der Stichproben geht dabei verloren. Für die Messung der Variation *zwischen* den Stichproben – unter Ausblendung der Variation innerhalb der Gruppen – bietet es sich daher an, von den Abweichungen $\bar{y}_{i\cdot} - \bar{y}_{..}$ vom Gesamtmittelwert auszugehen, diese zu quadrieren, die Quadrate mit den jeweiligen Stichprobenumfängen zu gewichten und aufzusummieren:

$$SQ_{\text{zwischen}} := \sum_{i=1}^s n_i \cdot (\bar{y}_{i\cdot} - \bar{y}_{..})^2. \quad (18.6)$$

Die nicht durch die Variation von Faktorstufen erklärte Reststreuung kann für jede Stichprobe $y_{i1}, y_{i2}, \dots, y_{in_i}$ durch die Abweichungen $y_{ik} - \bar{y}_{i\cdot}$ der Stichprobenelemente vom Stichprobenmittelwert $\bar{y}_{i\cdot}$ beschrieben werden. Für die s Stichproben hat man also

$$SQ_{\text{Residual}} := \sum_{i=1}^s \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{i\cdot})^2. \quad (18.7)$$

Mit diesen Bezeichnungen gilt die zu (17.16) analoge Streuungszerlegung

$$SQ_{\text{Total}} = SQ_{\text{zwischen}} + SQ_{\text{Residual}}. \quad (18.8)$$

Auf eine Herleitung wird verzichtet; man findet diese z. B. bei TOUTENBURG / HEUMANN (2009, Abschnitt 10.2.2). Ebenfalls ohne Beweis sei angeführt, dass die Streuungskomponenten SQ_{zwischen} und SQ_{Residual} unter der hier getroffenen Normalverteilungsannahme χ^2 -verteilt sind mit $s - 1$ resp. $n - s$ Freiheitsgraden.

Die aus den Daten der Tabelle 18.2 errechneten Stichprobenmittelwerte $\bar{y}_{i\cdot}$ und der Gesamtmittelwert $\bar{y}_{..}$ lassen sich als Realisationen von Zufallsvariablen $\bar{Y}_{i\cdot}$ resp. $\bar{Y}_{..}$ auffassen und zur unverzerrten Schätzung der in Tabelle 18.1 eingehenden Erwartungswerte μ_i und μ sowie der Effektstärken α_i einsetzen. Verwendet man für die Zufallsvariablen erneut Großbuchstaben, sind erwartungstreue Schätzer für μ_i , μ resp. α_i durch

$$\hat{\mu}_i = \bar{Y}_{i\cdot}; \quad \hat{\mu} = \bar{Y}_{..}; \quad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} \quad (18.9)$$

gegeben. Auch die Varianz σ^2 des in Tabelle 18.1 veranschaulichten Modells lässt sich erwartungstreue schätzen, wobei man entweder von der Streuungskomponente SQ_{zwischen} aus (18.6) oder aber von der Komponente SQ_{Residual} aus (18.7) ausgehen kann. Im erstgenannten Fall erhält man – vgl. (15.9) – eine unverzerrte Schätzung, indem man die aus s

Erwartungstreue
Schätzung der
Varianz

Einzeltermen bestehende Summe SQ_{zwischen} nicht durch s , sondern durch $s - 1$ dividiert:

$$\hat{\sigma}^2 = \frac{1}{s-1} \cdot SQ_{\text{zwischen}}. \quad (18.10)$$

Geht man bei der Herleitung einer erwartungstreuen Schätzung für σ^2 von der Restkomponente SQ_{Residual} aus, in die alle n Beobachtungswerte mit Differenzierung nach s Stichproben einfließen, so ist die aus n Einzeltermen bestehende Summe SQ_{Residual} durch $n - s$ zu teilen:

$$\hat{\sigma}^2 = \frac{1}{n-s} \cdot SQ_{\text{Residual}}. \quad (18.11)$$

Um nun zu testen, ob die Variation von Faktorstufen einen signifikanten Einfluss auf den Erwartungswert der Responsevariablen hat, vergleicht man nicht das Verhältnis der Streuungskomponenten SQ_{zwischen} und SQ_{Residual} , sondern bildet den Quotienten aus den korrigierten empirischen Streuungsmaßen (18.10) und (18.11), verwendet also die hier mit F bezeichnete Teststatistik

$$F := \frac{\frac{1}{s-1} \cdot SQ_{\text{zwischen}}}{\frac{1}{n-s} \cdot SQ_{\text{Residual}}} = \frac{n-s}{s-1} \cdot \frac{SQ_{\text{zwischen}}}{SQ_{\text{Residual}}}. \quad (18.12)$$

Dieser Quotient hat den Vorteil, dass er unter der Nullhypothese H_0 aus (18.3) bzw. (18.4) einer bekannten Verteilung folgt, nämlich einer **F-Verteilung** mit $s - 1$ und $n - s$ Freiheitsgraden. Unter H_0 gilt also $F \sim F_{s-1; n-s}$ (lies: F ist F -verteilt mit $s - 1$ und $n - s$ Freiheitsgraden). Die Alternativhypothese H_1 in (18.3) resp. (18.4) wird dann als statistisch gesichert angesehen mit einer vorab spezifizierten Irrtumswahrscheinlichkeit α , wenn der genannte Quotient „hinreichend“ groß ist. Letzteres wird als gegeben angesehen, wenn der für die Teststatistik errechnete Wert das $(1 - \alpha)$ -Quantil $F_{s-1; n-s; 1-\alpha}$ der F-Verteilung mit $s - 1$ und $n - s$ Freiheitsgraden überschreitet (vgl. hierzu Abbildung 13.9). Man führt also einen **F-Test** zum Signifikanzniveau α durch.

Beispiel 18.2: Wirkung alternativer Unterrichtsformen

Eine Population von 29 Schülern einer Altersstufe wird während einer Unterrichtseinheit zur Geometrie, die sich der Satzgruppe des Pythagoras widmet, im Rahmen eines Experiments über einen Zufallsalgorithmus in drei Gruppen aufgeteilt. In der ersten Teilpopulation des Umfangs $n_1 = 10$ erfahren die Schüler einen lehrerzentrierten Unterricht (Gruppe 1). In der zweiten Teilpopulation mit gleichem Umfang $n_2 = 10$ wird überwiegend in Zweiergruppen mit Aufgabenblättern gearbeitet (Gruppe 2). Die dritte Unterrichtsform, die auf $n_3 = 9$ Schüler bezogen wird, unterscheidet sich von der zweiten dadurch, dass hier

bei der Bearbeitung der Aufgaben leistungsfähige Computer mit interaktiver Geometriesoftware benutzt werden (Gruppe 3). In allen drei Gruppen ist die Lehrkraft im Einsatz.

Am Ende der Unterrichtseinheit werden alle 29 Schüler zur Messung ihrer individuellen Leistung einem Test unterzogen, bei dem maximal 100 Punkte zu erzielen sind. Es wird angenommen, dass sich die Punktzahl Y , die von den Schülern erreicht wird, approximativ durch eine Normalverteilung beschreiben lässt, deren Varianz in allen drei Gruppen gleich ist. Es soll zum Signifikanzniveau $\alpha = 0,05$ getestet werden, ob sich die verschiedenen Unterrichtsformen im Mittel auf die Leistungen bei der Abschlussprüfung auswirken. Zu testen sind also die Hypothesen aus (18.3), die sich hier auf $s = 3$ Erwartungswerte beziehen. In der Praxis wird man bei Durchführung eines solchen Tests ein Statistiksoftwarepaket heranziehen. Es ist aber durchaus verständnisfördernd, die einzelnen Zwischenschritte bis zum Wert der Prüfstatistik (18.12) einmal ohne Softwareunterstützung ausgeführt zu haben.

Bei der Leistungsmessung am Ende der Geometrieunterrichtseinheit gab es für die drei Stufen des Faktors „Unterrichtsform“ folgende Einzelergebnisse:

Gruppe	Element-Nr der Stichprobe										\sum	Mittelwert
	1	59	48	65	38	74	43	62	42	62	58	
2	57	77	64	49	48	74	50	51	46	58	574	57,4
3	78	81	63	79	67	76	75	52	59		630	70,0

Tab. 18.3: Punktzahlen beim Geometrieabschlusstest

Aus den Daten errechnet man den Gesamtmittelwert $\bar{y}_{..}$ als Summe der drei Elemente der vorletzten Spalte, wenn man diese noch durch die Anzahl $n = 29$ der Beobachtungen teilt. Man erhält so

$$\bar{y}_{..} = \frac{551 + 574 + 630}{29} = \frac{1755}{29} \approx 60,517.$$

Für den Anteil SQ_{zwischen} der Streuung zwischen den drei Gruppen an der Gesamtstreuung SQ_{Total} folgt nach (18.6)

$$SQ_{\text{zwischen}} = 10 \cdot (55,1 - \bar{y}_{..})^2 + 10 \cdot (57,4 - \bar{y}_{..})^2 + 9 \cdot (70,0 - \bar{y}_{..})^2 \approx 1199,94.$$

Für die Reststreuung SQ_{Residual} errechnet man dann, z. B. unter Einsatz eines Tabellenkalkulationsprogramms, mit (18.7)

$$SQ_{\text{Residual}} = \sum_{i=1}^3 \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{i.})^2 \approx 3153,30.$$

Für die Testgröße (18.12) ergibt sich schließlich mit $s = 3$ und $n - s = 26$

$$F = \frac{26}{2} \cdot \frac{SQ_{\text{zwischen}}}{SQ_{\text{Residual}}} \approx 4,95.$$

Der aus den Daten errechnete Wert $F \approx 4,95$ ist noch mit dem 0,95-Quantil der F -Verteilung mit 2 und 26 Freiheitsgraden zu vergleichen. Da dieses Quantil nach Tabelle 19.6 den Wert $F_{2;26;0,95} = 3,37$ hat, ist die Nullhypothese H_0 wegen $F > 3,37$ abzulehnen, d. h. es ist von einem statistisch signifikanten Einfluss der Unterrichtsform auf die Leistungen im Geometrieunterricht auszugehen. Hätte man den Test z. B. mit $\alpha = 0,01$ durchgeführt, also nur eine deutlich geringere Irrtumswahrscheinlichkeit α in Kauf genommen, wäre wegen $F_{2;26;0,99} = 5,53$ keine Ablehnung von H_0 erfolgt.

Eine grafische Darstellung der Beobachtungsdaten, etwa anhand eines Boxplots pro Gruppe, kann den F -Test ergänzen und zusätzliche Informationen vermitteln. Abbildung 18.1 zeigt dies für den hier betrachteten Beispieldatensatz. Die Grafik liefert nicht nur Informationen über Lageparameter der Teilpopulationen, sondern z. B. auch solche, die die Streuung innerhalb der Gruppen betreffen.

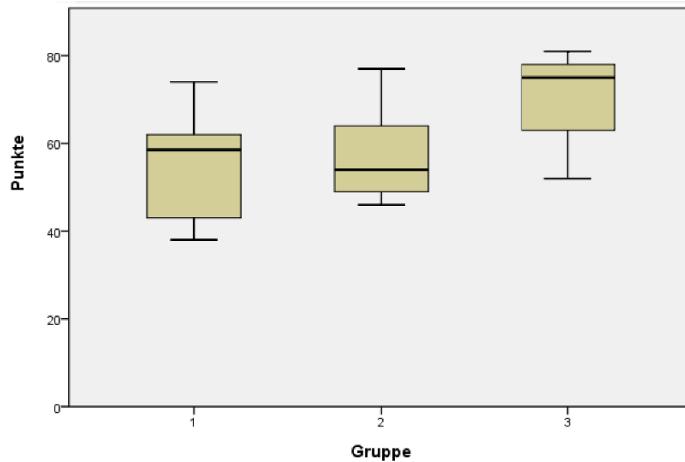


Abb. 18.1: Boxplots für die Punktzahlen beim Geometrieabschlusstest

Grenzen der
Varianzanalyse

Kommt man mit Anwendung des F -Tests zu einer Verwerfung der Nullhypothese, weiß man nur, dass zwischen mindestens zwei Gruppen ein statistisch signifikanter Unterschied bezüglich der Erwartungswerte besteht. Welche Gruppen dies sind, beantwortet die Varianzanalyse noch *nicht*. Man ist dann auf ergänzende Verfahren angewiesen (z. B. paarweiser Gruppenvergleich), auf die in dieser Einführung ebenso wenig eingegangen werden kann wie auf die Vorgehensweise bei Verletzung der Normalverteilungsannahme. Im letztgenannten Falle bietet sich die Anwendung nicht-parametrischer Tests anstelle des F -Tests an, also von Tests, die nicht die Annahme einer bestimmten Verteilung (hier: Normalverteilung) voraussetzen.

Sehr nützlich ist auch die Visualisierung der Beobachtungen für die einzelnen Gruppen, etwa – wie in Abbildung 18.1 beispielhaft illustriert –

anhand von Boxplots. Auf diese Weise erhält man schon einen guten Eindruck von der Verteilung der erklärten Variablen innerhalb der Gruppen und kann Unterschiede bezüglich der empirischen Verteilungen oft schon aus der Grafik erkennen. Instrumente der beschreibenden Statistik können jedenfalls häufig Einsichten vermitteln, die die Ergebnisse von Verfahren der schließenden Statistik, z. B. eines F -Tests, sinnvoll ergänzen.

18.3 Zweifaktorielle Varianzanalyse

Wenn man den Einfluss von *zwei* Einflussgrößen X_1 und X_2 mit s resp. r Faktorstufen auf eine Responsevariable Y betrachtet, erhält man anstelle von (18.1) eine Darstellung, die sich auf $s \cdot r$ Faktorstufenkombinationen bezieht:

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + U_{ijk} & i = 1, \dots, s; \\ && j = 1, \dots, r; \\ && k = 1, \dots, n_{ij}, \end{aligned} \quad (18.13)$$

wobei die Störvariablen als unabhängig identisch $N(0; \sigma^2)$ -verteilt spezifiziert sind. Zerlegt man die Erwartungswerte μ_{ij} der Responsevariablen in den $s \cdot r$ Gruppen wieder additiv in einen für alle Gruppen identischen Basisanteil μ und in faktorstufenspezifische Komponenten α_i (Effekt der i -ten Stufe des Faktors X_1) sowie β_j (Effekt der j -ten Stufe des Faktors X_2) und berücksichtigt man noch einen mit $(\alpha\beta)_{ij}$ bezeichneten möglichen Wechselwirkungseffekt zwischen der i -ten Stufe von X_1 und der j -ten Stufe von X_2 , erhält man das **Modell der zweifaktoriellen Varianzanalyse in Effektdarstellung**:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + U_{ijk} & i = 1, \dots, s; \\ && j = 1, \dots, r; \\ && k = 1, \dots, n_{ij}. \end{aligned} \quad (18.14)$$

Wechselwirkung oder **Interaktion** beinhaltet, dass der Effekt einer bestimmten Faktorstufe eines Faktors auf die erklärte Variable Y davon abhängt, welche Faktorstufe bei dem anderen Faktor vorliegt.

In Tabelle 18.1 ist also jeder Wert y_{ik} im grau markierten Tabelleninneren durch r Werte $y_{ij1}, y_{ij2}, \dots, y_{ijr}$ zu ersetzen. Die Streuungszerlegung (18.8) gilt zwar unverändert, die Komponente SQ_{zwischen} lässt sich jetzt aber aufteilen in einen Streuungsanteil, der nur auf die Variation des Faktors X_1 zurückgeht, einen weiteren, der durch die Veränderung von Faktorstufen bei X_2 bedingt ist und einen dritten, der auf Interaktionseffekten zwischen den beiden Faktoren beruht. Abbildung 18.2 veranschaulicht dies.

Zerlegung der Streuung zwischen den Gruppen

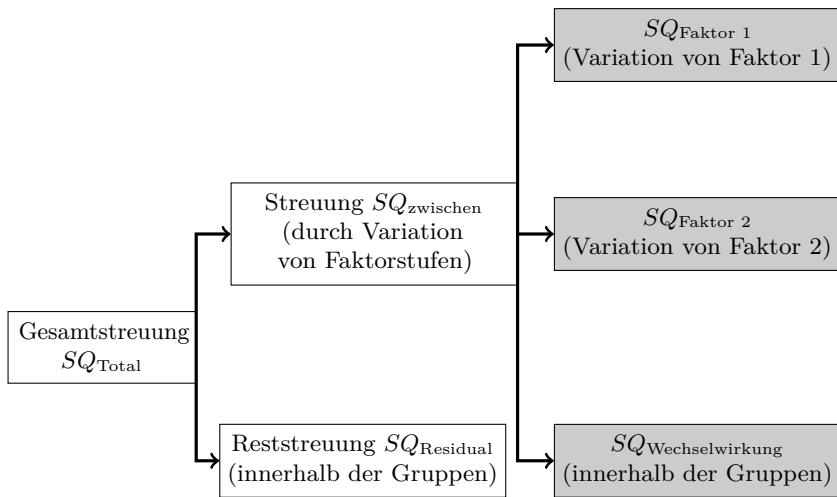


Abb. 18.2: *Streuungszerlegung bei der zweifaktoriellen Varianzanalyse
(Zerlegung im einfaktoriellen Fall: ohne grau gerasterte Komponenten)*

Effekte auf die Responsevariable Y , die durch die Veränderung von Stufen von Faktor X_1 oder von Faktor X_2 hervorgerufen werden, heißen **Haupteffekte**. Wirkungen auf Y , die durch Interaktion der beiden Faktoren induziert werden, nennt man **Wechselwirkungseffekte** oder auch **Interaktionseffekte**. Entsprechend hat man bei der zweifaktoriellen Varianzanalyse *drei* F-Tests durchzuführen – zwei zur Überprüfung von Haupt- und einen zur Feststellung von Wechselwirkungseffekten. Eine ausführlichere Behandlung der ein- und zweifaktoriellen Varianzanalyse findet man bei FAHRMEIR / HEUMANN / KÜNSTLER / PIGEOT / TUTZ (2016, Kapitel 13) sowie bei SCHLITTGEN (2012, Kapitel 17).



Mit dem Modell der zweifaktoriellen Varianzanalyse und dem linearen multiplen Regressionsmodell wurden zwei multivariate statistische Modelle vorgestellt. Es gibt eine Reihe weiterer multivariater Modelle und Verfahren, deren detaillierte Behandlung den Rahmen einer Einführung in die Statistik sprengen würden. Nur erwähnt sei hier noch die vor allem in der *Psychologie* breit eingesetzte **Faktorenanalyse**, deren Ziel es ist, Beobachtungen an mehreren manifesten Variablen durch möglichst wenige latente Variablen zu erklären. Diese und weitere multivariate Verfahren sind bei HANDL / KUHLENKASPER (2017) und FAHRMEIR / HAMERLE / TUTZ (1996) detailliert beschrieben.

Teil III

Anhänge



Lernziele zu Teil III

Der letzte Teil dieses Manuskripts enthält vor allem Aufgaben mit ausführlichen Lösungen zur Verständnissicherung und Lernerfolgskontrolle. Danach folgen Tabellen mit Werten von Verteilungsfunktionen der Binomial- und der Standardnormalverteilung sowie Quantile dieser und einiger weiterer Verteilungen. Das anschließende Literaturverzeichnis wird durch eine Sammlung interessanter Internetadressen ergänzt. Es werden auch die wichtigsten in der Statistik gängigen Symbole und Notationen übersichtsartig präsentiert.

Nach Bearbeitung des dritten Teils dieses Kurses sollten Sie

- in der Lage sein, Quantile und Werte von Verteilungsfunktionen aus den zur Verfügung gestellten Tabellen abzulesen und die inhaltliche Bedeutung der tabellierten Werte zu verstehen;
- in der Lage sein, die Aufgaben zu Teil I – II dieses Buches erfolgreich zu bearbeiten;
- sich bei der Lösung der Übungsaufgaben auch den unter der Webadresse <https://statistiklehrbuch.statup.solutions> eingestellten R-Code zu den Lösungen angeschaut und selbst ausprobier haben;
- einige interessante Internet-Seiten zur Datenvisualisierung aufgesucht und damit neuere Entwicklungen auf diesem Sektor kennengelernt haben;
- in einige Diskussionsforen sowie Online-Lehrmaterial-Sammlungen zur Statistik hineingeschaut und – hoffentlich – auf diese Weise die Statistik als lebensnahe Disziplin erfahren haben;
- sich mit den im Symbolverzeichnis zusammengestellten mathematischen Symbolen und Schreibweisen, die in der Statistik häufiger verwendet werden, vertraut gemacht haben.

19 Tabellenanhang



In diesem Kapitel findet man Werte der *Verteilungsfunktion* der Binomialverteilung und der Standardnormalverteilung. Ferner sind für die Standardnormalverteilung, die χ^2 -, die t - und die F -Verteilung *Quantile* tabelliert.

Die Tabellen geben nur ausgewählte Werte der genannten Verteilungsfunktionen bzw. Quantile wieder. Es wurden aber interaktive Visualisierungen entwickelt, die einen Zugang zu weiteren Werten vermitteln und eine inhaltliche Interpretation der tabellierten Werte liefern. Außerdem ist R-Code aufgeführt, mit dem sich alle Werte leicht nachrechnen lassen.

Vorschau auf das Kapitel

19.1 Verteilungsfunktion der Binomialverteilung

Es sei $X \sim B(n; p)$ eine mit Parametern n und p binomialverteilte Zufallsvariable. Die Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ lautet

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots, n$$



Interaktives Objekt „Binomialverteilung“

$$F(x) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k} \quad x = 0, 1, \dots, n.$$

In Tabelle 19.1 sind Werte $F(x)$ der Verteilungsfunktion einer $B(n; p)$ -verteilten Zufallsvariablen X für $n = 1, 2, \dots, 20$ und $p = 0,05$ sowie für ausgewählte Werte p zusammengestellt. Man entnimmt der Tabelle z. B., dass $F(x)$ für $x = 3$, $n = 10$ und $p = 0,50$ den Wert $F(3) = 0,1719$ annimmt. Dieser Wert entspricht der Summe $f(0), f(1), f(2), f(3)$ aller Werte der Wahrscheinlichkeitsfunktion bis zur Stelle $x = 3$. Der Wert $f(3)$ der Wahrscheinlichkeitsfunktion der genannten Binomialverteilung ergibt sich auch als Differenz $f(3) = F(3) - F(2) = 0,1172$.

Mit R lassen sich Werte $f(x)$ und $F(x)$ der Wahrscheinlichkeitsfunktion und der Verteilungsfunktion einer $B(n; p)$ -verteilten Zufallsvariablen anhand der Funktionen `dbinom(x, size, prob)` resp. `pbinom(x, size, prob)` leicht berechnen. Im Falle $x = 3$ sowie $n = 10$ und $p = 0,5$ erhält man für den Wert $F(3)$ der Verteilungsfunktion



Interaktives Objekt „Rechnen mit der Binomialverteilung“

```
> pbinom(3, size = 10, prob = .5)
[1] 0.171875
```

und für den Wert $f(3)$ der Wahrscheinlichkeitsfunktion

```
> dbinom(3, size = 10, prob = .5)
[1] 0.1171875
```

Die Werte können auch, wie in Tabelle 19.1 realisiert, mit einer geringeren Anzahl von Dezimalstellen ausgegeben werden.

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

n	x	p=0,05	p=0,10	p=0,20	p=0,30	p=0,40	p=0,50
1	0	0,9500	0,9000	0,8000	0,7000	0,6000	0,5000
	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	0	0,9025	0,8100	0,6400	0,4900	0,3600	0,2500
	1	0,9975	0,9900	0,9600	0,9100	0,8400	0,7500
2	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	0	0,8574	0,7290	0,5120	0,3430	0,2160	0,1250
	1	0,9928	0,9720	0,8960	0,7840	0,6480	0,5000
	2	0,9999	0,9990	0,9920	0,9730	0,9360	0,8750
	3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	0	0,8145	0,6561	0,4096	0,2401	0,1296	0,0625
	1	0,9860	0,9477	0,8192	0,6517	0,4752	0,3125
	2	0,9995	0,9963	0,9728	0,9163	0,8208	0,6875
	3	1,0000	0,9999	0,9984	0,9919	0,9744	0,9375
	4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	0	0,7738	0,5905	0,3277	0,1681	0,0778	0,0313
	1	0,9774	0,9185	0,7373	0,5282	0,3370	0,1875
	2	0,9988	0,9914	0,9421	0,8369	0,6826	0,5000
	3	1,0000	0,9995	0,9933	0,9692	0,9130	0,8125
	4	1,0000	1,0000	0,9997	0,9976	0,9898	0,9688
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	0	0,7351	0,5314	0,2621	0,1176	0,0467	0,0156
	1	0,9672	0,8857	0,6554	0,4202	0,2333	0,1094
	2	0,9978	0,9842	0,9011	0,7443	0,5443	0,3438
	3	0,9999	0,9987	0,9830	0,9295	0,8208	0,6563
	4	1,0000	0,9999	0,9984	0,9891	0,9590	0,8906
	5	1,0000	1,0000	0,9999	0,9993	0,9959	0,9844
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	0	0,6983	0,4783	0,2097	0,0824	0,0280	0,0078
	1	0,9556	0,8503	0,5767	0,3294	0,1586	0,0625
	2	0,9962	0,9743	0,8520	0,6471	0,4199	0,2266

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

n	x	p=0,05	p=0,10	p=0,20	p=0,30	p=0,40	p=0,50
7	3	0,9998	0,9973	0,9667	0,8740	0,7102	0,5000
7	4	1,0000	0,9998	0,9953	0,9712	0,9037	0,7734
7	5	1,0000	1,0000	0,9996	0,9962	0,9812	0,9375
7	6	1,0000	1,0000	1,0000	0,9998	0,9984	0,9922
7	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
8	0	0,6634	0,4305	0,1678	0,0576	0,0168	0,0039
8	1	0,9428	0,8131	0,5033	0,2553	0,1064	0,0352
8	2	0,9942	0,9619	0,7969	0,5518	0,3154	0,1445
8	3	0,9996	0,9950	0,9437	0,8059	0,5941	0,3633
8	4	1,0000	0,9996	0,9896	0,9420	0,8263	0,6367
8	5	1,0000	1,0000	0,9988	0,9887	0,9502	0,8555
8	6	1,0000	1,0000	0,9999	0,9987	0,9915	0,9648
8	7	1,0000	1,0000	1,0000	0,9999	0,9993	0,9961
8	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
9	0	0,6302	0,3874	0,1342	0,0404	0,0101	0,0020
9	1	0,9288	0,7748	0,4362	0,1960	0,0705	0,0195
9	2	0,9916	0,9470	0,7382	0,4628	0,2318	0,0898
9	3	0,9994	0,9917	0,9144	0,7297	0,4826	0,2539
9	4	1,0000	0,9991	0,9804	0,9012	0,7334	0,5000
9	5	1,0000	0,9999	0,9969	0,9747	0,9006	0,7461
9	6	1,0000	1,0000	0,9997	0,9957	0,9750	0,9102
9	7	1,0000	1,0000	1,0000	0,9996	0,9962	0,9805
9	8	1,0000	1,0000	1,0000	1,0000	0,9997	0,9980
9	9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10	0	0,5987	0,3487	0,1074	0,0282	0,0060	0,0010
10	1	0,9139	0,7361	0,3758	0,1493	0,0464	0,0107
10	2	0,9885	0,9298	0,6778	0,3828	0,1673	0,0547
10	3	0,9990	0,9872	0,8791	0,6496	0,3823	0,1719
10	4	0,9999	0,9984	0,9672	0,8497	0,6331	0,3770
10	5	1,0000	0,9999	0,9936	0,9527	0,8338	0,6230
10	6	1,0000	1,0000	0,9991	0,9894	0,9452	0,8281
10	7	1,0000	1,0000	0,9999	0,9984	0,9877	0,9453
10	8	1,0000	1,0000	1,0000	0,9999	0,9983	0,9893
10	9	1,0000	1,0000	1,0000	1,0000	0,9999	0,9990
10	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
11	0	0,5688	0,3138	0,0859	0,0198	0,0036	0,0005
11	1	0,8981	0,6974	0,3221	0,1130	0,0302	0,0059
11	2	0,9848	0,9104	0,6174	0,3127	0,1189	0,0327
11	3	0,9984	0,9815	0,8389	0,5696	0,2963	0,1133

Fortsetzung nächste Seite

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

n	x	p=0,05	p=0,10	p=0,20	p=0,30	p=0,40	p=0,50
11	4	0,9999	0,9972	0,9496	0,7897	0,5328	0,2744
11	5	1,0000	0,9997	0,9883	0,9218	0,7535	0,5000
11	6	1,0000	1,0000	0,9980	0,9784	0,9006	0,7256
11	7	1,0000	1,0000	0,9998	0,9957	0,9707	0,8867
11	8	1,0000	1,0000	1,0000	0,9994	0,9941	0,9673
11	9	1,0000	1,0000	1,0000	1,0000	0,9993	0,9941
11	10	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995
11	11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
12	0	0,5404	0,2824	0,0687	0,0138	0,0022	0,0002
12	1	0,8816	0,6590	0,2749	0,0850	0,0196	0,0032
12	2	0,9804	0,8891	0,5583	0,2528	0,0834	0,0193
12	3	0,9978	0,9744	0,7946	0,4925	0,2253	0,0730
12	4	0,9998	0,9957	0,9274	0,7237	0,4382	0,1938
12	5	1,0000	0,9995	0,9806	0,8822	0,6652	0,3872
12	6	1,0000	0,9999	0,9961	0,9614	0,8418	0,6128
12	7	1,0000	1,0000	0,9994	0,9905	0,9427	0,8062
12	8	1,0000	1,0000	0,9999	0,9983	0,9847	0,9270
12	9	1,0000	1,0000	1,0000	0,9998	0,9972	0,9807
12	10	1,0000	1,0000	1,0000	1,0000	0,9997	0,9968
12	11	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
12	12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
13	0	0,5133	0,2542	0,0550	0,0097	0,0013	0,0001
13	1	0,8646	0,6213	0,2336	0,0637	0,0126	0,0017
13	2	0,9755	0,8661	0,5017	0,2025	0,0579	0,0112
13	3	0,9969	0,9658	0,7473	0,4206	0,1686	0,0461
13	4	0,9997	0,9935	0,9009	0,6543	0,3530	0,1334
13	5	1,0000	0,9991	0,9700	0,8346	0,5744	0,2905
13	6	1,0000	0,9999	0,9930	0,9376	0,7712	0,5000
13	7	1,0000	1,0000	0,9988	0,9818	0,9023	0,7095
13	8	1,0000	1,0000	0,9998	0,9960	0,9679	0,8666
13	9	1,0000	1,0000	1,0000	0,9993	0,9922	0,9539
13	10	1,0000	1,0000	1,0000	0,9999	0,9987	0,9888
13	11	1,0000	1,0000	1,0000	1,0000	0,9999	0,9983
13	12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
13	13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
14	0	0,4877	0,2288	0,0440	0,0068	0,0008	0,0001
14	1	0,8470	0,5846	0,1979	0,0475	0,0081	0,0009
14	2	0,9699	0,8416	0,4481	0,1608	0,0398	0,0065
14	3	0,9958	0,9559	0,6982	0,3552	0,1243	0,0287
14	4	0,9996	0,9908	0,8702	0,5842	0,2793	0,0898

Fortsetzung nächste Seite

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

n	x	p=0,05	p=0,10	p=0,20	p=0,30	p=0,40	p=0,50
14	5	1,0000	0,9985	0,9561	0,7805	0,4859	0,2120
14	6	1,0000	0,9998	0,9884	0,9067	0,6925	0,3953
14	7	1,0000	1,0000	0,9976	0,9685	0,8499	0,6047
14	8	1,0000	1,0000	0,9996	0,9917	0,9417	0,7880
14	9	1,0000	1,0000	1,0000	0,9983	0,9825	0,9102
14	10	1,0000	1,0000	1,0000	0,9998	0,9961	0,9713
14	11	1,0000	1,0000	1,0000	1,0000	0,9994	0,9935
14	12	1,0000	1,0000	1,0000	1,0000	0,9999	0,9991
14	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
14	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
15	0	0,4633	0,2059	0,0352	0,0047	0,0005	0,0000
15	1	0,8290	0,5490	0,1671	0,0353	0,0052	0,0005
15	2	0,9638	0,8159	0,3980	0,1268	0,0271	0,0037
15	3	0,9945	0,9444	0,6482	0,2969	0,0905	0,0176
15	4	0,9994	0,9873	0,8358	0,5155	0,2173	0,0592
15	5	0,9999	0,9978	0,9389	0,7216	0,4032	0,1509
15	6	1,0000	0,9997	0,9819	0,8689	0,6098	0,3036
15	7	1,0000	1,0000	0,9958	0,9500	0,7869	0,5000
15	8	1,0000	1,0000	0,9992	0,9848	0,9050	0,6964
15	9	1,0000	1,0000	0,9999	0,9963	0,9662	0,8491
15	10	1,0000	1,0000	1,0000	0,9993	0,9907	0,9408
15	11	1,0000	1,0000	1,0000	0,9999	0,9981	0,9824
15	12	1,0000	1,0000	1,0000	1,0000	0,9997	0,9963
15	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995
15	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
16	0	0,4401	0,1853	0,0281	0,0033	0,0003	0,0000
16	1	0,8108	0,5147	0,1407	0,0261	0,0033	0,0003
16	2	0,9571	0,7892	0,3518	0,0994	0,0183	0,0021
16	3	0,9930	0,9316	0,5981	0,2459	0,0651	0,0106
16	4	0,9991	0,9830	0,7982	0,4499	0,1666	0,0384
16	5	0,9999	0,9967	0,9183	0,6598	0,3288	0,1051
16	6	1,0000	0,9995	0,9733	0,8247	0,5272	0,2272
16	7	1,0000	0,9999	0,9930	0,9256	0,7161	0,4018
16	8	1,0000	1,0000	0,9985	0,9743	0,8577	0,5982
16	9	1,0000	1,0000	0,9998	0,9929	0,9417	0,7728
16	10	1,0000	1,0000	1,0000	0,9984	0,9809	0,8949
16	11	1,0000	1,0000	1,0000	0,9997	0,9951	0,9616
16	12	1,0000	1,0000	1,0000	1,0000	0,9991	0,9894
16	13	1,0000	1,0000	1,0000	1,0000	0,9999	0,9979
16	14	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997
16	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Fortsetzung nächste Seite

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

n	x	p=0,05	p=0,10	p=0,20	p=0,30	p=0,40	p=0,50
16	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
17	0	0,4181	0,1668	0,0225	0,0023	0,0002	0,0000
17	1	0,7922	0,4818	0,1182	0,0193	0,0021	0,0001
17	2	0,9497	0,7618	0,3096	0,0774	0,0123	0,0012
17	3	0,9912	0,9174	0,5489	0,2019	0,0464	0,0064
17	4	0,9988	0,9779	0,7582	0,3887	0,1260	0,0245
17	5	0,9999	0,9953	0,8943	0,5968	0,2639	0,0717
17	6	1,0000	0,9992	0,9623	0,7752	0,4478	0,1662
17	7	1,0000	0,9999	0,9891	0,8954	0,6405	0,3145
17	8	1,0000	1,0000	0,9974	0,9597	0,8011	0,5000
17	9	1,0000	1,0000	0,9995	0,9873	0,9081	0,6855
17	10	1,0000	1,0000	0,9999	0,9968	0,9652	0,8338
17	11	1,0000	1,0000	1,0000	0,9993	0,9894	0,9283
17	12	1,0000	1,0000	1,0000	0,9999	0,9975	0,9755
17	13	1,0000	1,0000	1,0000	1,0000	0,9995	0,9936
17	14	1,0000	1,0000	1,0000	1,0000	0,9999	0,9988
17	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
17	16	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
17	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	0	0,3972	0,1501	0,0180	0,0016	0,0001	0,0000
18	1	0,7735	0,4503	0,0991	0,0142	0,0013	0,0001
18	2	0,9419	0,7338	0,2713	0,0600	0,0082	0,0007
18	3	0,9891	0,9018	0,5010	0,1646	0,0328	0,0038
18	4	0,9985	0,9718	0,7164	0,3327	0,0942	0,0154
18	5	0,9998	0,9936	0,8671	0,5344	0,2088	0,0481
18	6	1,0000	0,9988	0,9487	0,7217	0,3743	0,1189
18	7	1,0000	0,9998	0,9837	0,8593	0,5634	0,2403
18	8	1,0000	1,0000	0,9957	0,9404	0,7368	0,4073
18	9	1,0000	1,0000	0,9991	0,9790	0,8653	0,5927
18	10	1,0000	1,0000	0,9998	0,9939	0,9424	0,7597
18	11	1,0000	1,0000	1,0000	0,9986	0,9797	0,8811
18	12	1,0000	1,0000	1,0000	0,9997	0,9942	0,9519
18	13	1,0000	1,0000	1,0000	1,0000	0,9987	0,9846
18	14	1,0000	1,0000	1,0000	1,0000	0,9998	0,9962
18	15	1,0000	1,0000	1,0000	1,0000	1,0000	0,9993
18	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
18	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
18	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	0	0,3774	0,1351	0,0144	0,0011	0,0001	0,0000
19	1	0,7547	0,4203	0,0829	0,0104	0,0008	0,0000

Fortsetzung nächste Seite

Tab. 19.1: Verteilungsfunktion $F(x)$ der Binomialverteilung

19.2 Verteilungsfunktion der Standardnormalverteilung



Eine mit Erwartungswert μ und Varianz σ^2 normalverteilte Zufallsvariable X lässt sich stets anhand der Transformation $Z := \frac{X-\mu}{\sigma}$ in die **Standardnormalverteilung** überführen. Diese hat die Dichtefunktion



Interaktives Objekt
„Standardnormal-
verteilung“

und die Verteilungsfunktion

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{t^2}{2}\right) dt.$$

In Tabelle 19.2 sind für den Bereich von $z = 0,00$ bis $z = 3,99$ Werte der Verteilungsfunktion $\Phi(z)$ auf vier Dezimalstellen genau wiedergegeben. Im Kopf der Tabelle ist die zweite Dezimalstelle der Argumentvariablen z ausgewiesen. Die Beschränkung auf nicht-negative Werte von z ist aufgrund der Symmetriebeziehung $\Phi(z) = 1 - \Phi(-z)$ gerechtfertigt.



Interaktives Objekt
„Rechnen mit der
Standard-
normalverteilung“



Werte der Verteilungsfunktion $\Phi(z)$ sind auch über das nebenstehende interaktive Lernobjekt zugänglich. Bei Verwendung dieses Elements kann $\Phi(z)$ sogar für negatives z direkt abgelesen werden. Die Fläche unter der Dichtekurve $\phi(z)$ zwischen zwei Punkten der z -Achse lässt sich als Differenz von Werten der Funktion $\Phi(z)$ ausdrücken. Bei der Berechnung solcher Differenzen kann das zweite auf dieser Seite präsentierte interaktive Lernobjekt herangezogen werden. In Abbildung 13.4 sind beispielhaft die Werte $\Phi(1) = 0,8413$ und die Differenz $\Phi(2) - \Phi(-1) = 0,8185$ visualisiert.

Mit R lassen sich Werte $\Phi(z)$ der Verteilungsfunktion einer standardnormalverteilten Zufallsvariablen Z anhand der Funktion `dnorm(z)` besonders einfach bestimmen. Für $\Phi(1)$ erhält man so

```
> pnorm(1)
[1] 0.8413447
```

und für $\Phi(2) - \Phi(-1)$

```
> pnorm(2) - pnorm(-1)
[1] 0.8185946
```

Die Werte stimmen mit den oben angegebenen Ergebnissen überein, wenn man auf nur vier Dezimalstellen runden. Die Rundung auf eine spezifizierte Anzahl von Dezimalstellen kann in R mittels der Funktion `round(..)` erfolgen.

Tab. 19.2: Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung
 (Ziffern im Tabellenkopf: Zweite Dezimalstelle von z)

19.3 Quantile der Standardnormalverteilung

Aus Tabelle 19.2 lassen sich auch **Quantile** ablesen. Ein p -Quantil z_p der Standardnormalverteilung ist durch

$$\Phi(z_p) = p \quad (0 < p < 1)$$

definiert und markiert den Punkt auf der z-Achse, bis zu dem die Fläche unter der Dichte gerade p ist. Im rechten Teil von Abbildung 13.3 ist beispielhaft das 0,95-Quantil $z_{0,95} = 1,6449$ der Standardnormalverteilung visualisiert.

In der folgenden Tabelle 19.3 sind einige Quantile z_p der zusammengestellt. Die Tabellierung beschränkt sich auf p -Quantile mit $p \geq 0,5$. Weitere Quantile ergeben sich aus der Beziehung

$$z_p = -z_{1-p},$$



die in (13.25) wiedergegeben ist und sich aus der Symmetrie von Dichtefunktion und Verteilungsfunktion bezüglich $z = 0$ ergibt. Mit $z_{0,95} = 1,6449$ gilt z. B. $z_{0,05} = -1,6449$. Quantile der Standardnormalverteilung sind auch über das nebenstehende interaktive Lernobjekt zugänglich. Dieses gestattet es sogar Quantile z_p mit $p < 0,5$ direkt abzulesen.

Interaktives Objekt

„Quantile der Standardnormalverteilung“

p	0,500	0,600	0,700	0,800	0,900
z_p	0,0000	0,2533	0,5244	0,8416	1,2816
p	0,950	0,975	0,990	0,995	0,999
z_p	1,6449	1,9600	2,3263	2,5758	3,0902

Tab. 19.3: Quantile z_p der Standardnormalverteilung

Mit R lassen sich Quantile einer standardnormalverteilten Zufallsvariablen Z anhand der Funktion `qnorm(z)` berechnen. Für das 0,95-Quantil ergibt sich

```
> qnorm(.95)
[1] 1.644854
```

Das Ergebnis stimmt mit dem in Tabelle 19.3 angegebenen Wert überein, wenn man auf nur vier Dezimalstellen runden.

19.4 Quantile der χ^2 -Verteilung

In der folgenden Tabelle sind Quantile $\chi_{n;p}^2$ der χ^2 -Verteilung mit n Freiheitsgraden für $n = 1$ bis $n = 25$ und ausgewählte Werte p zusammengestellt. Man entnimmt der Tabelle z. B., dass das im linken Teil von Abbildung 13.6 visualisierte 0,95-Quantil der χ^2 -Verteilung mit $n = 4$ Freiheitsgraden den Wert $\chi_{4;0,95}^2 = 9,488$ besitzt. Weitere Quantile der χ^2 -Verteilung sind über das nebenstehende interaktive Lernobjekt zugänglich oder anhand der Funktion $qchisq(p, df)$ von R. Dabei bezeichnet df die Anzahl der Freiheitsgrade (engl: degrees of freedom). Man erhält für das Quantil $\chi_{4;0,95}^2$

```
> qchisq(.95, df = 4)
[1] 9.487729
```



Interaktives Objekt
„Quantile der
 χ^2 -Verteilung“

Bei Rundung auf drei Dezimalstellen entspricht dies dem tabellierten Wert.

Tab. 19.4: Quantile der χ^2 -Verteilung

n	$p = 0,01$	$p = 0,025$	$p = 0,05$	$p = 0,95$	$p = 0,975$	$p = 0,99$
1	0,000	0,001	0,004	3,841	5,024	6,635
2	0,020	0,051	0,103	5,991	7,378	9,210
3	0,115	0,216	0,352	7,815	9,348	11,345
4	0,297	0,484	0,711	9,488	11,143	13,277
5	0,554	0,831	1,145	11,070	12,833	15,086
6	0,872	1,237	1,635	12,592	14,449	16,812
7	1,239	1,690	2,167	14,067	16,013	18,475
8	1,647	2,180	2,733	15,507	17,535	20,090
9	2,088	2,700	3,325	16,919	19,023	21,666
10	2,558	3,247	3,940	18,307	20,483	23,209
11	3,053	3,816	4,575	19,675	21,920	24,725
12	3,571	4,404	5,226	21,026	23,337	26,217
13	4,107	5,009	5,892	22,362	24,736	27,688
14	4,660	5,629	6,571	23,685	26,119	29,141
15	5,229	6,262	7,261	24,996	27,488	30,578
16	5,812	6,908	7,962	26,296	28,845	32,000
17	6,408	7,564	8,672	27,587	30,191	33,409
18	7,015	8,231	9,390	28,869	31,526	34,805
19	7,633	8,907	10,117	30,144	32,852	36,191
20	8,260	9,591	10,851	31,410	34,170	37,566
21	8,897	10,283	11,591	32,671	35,479	38,932
22	9,542	10,982	12,338	33,924	36,781	40,289
23	10,196	11,689	13,091	35,172	38,076	41,638
24	10,856	12,401	13,848	36,415	39,364	42,980
25	11,524	13,120	14,611	37,652	40,646	44,314

19.5 Quantile der t -Verteilung

Nachstehend sind **Quantile** $t_{n;p}$ der t -Verteilung für $n = 1$ bis $n = 40$ Freiheitsgrade und ausgewählte Werte p zusammengestellt. Aus der Tabelle geht z. B. hervor, dass das im rechten Teil von Abbildung 13.7 dargestellte 0,975-Quantil der t -Verteilung mit $n = 10$ Freiheitsgraden den Wert $t_{10;0,975} = 2,228$ besitzt. Bei bekanntem p -Quantil ergibt sich das $(1 - p)$ -Quantil aus der Beziehung



Interaktives Objekt
„Quantile der
 t -Verteilung“

$$t_{n;p} = -t_{n;1-p},$$

die in (13.29) wiedergegeben ist und sich aus der Symmetrie von Dichte- und Verteilungsfunktion bezüglich $x = 0$ ableitet. Weitere Quantile der t -Verteilung lassen sich anhand des nebenstehenden interaktiven Lernobjekts anzeigen oder über die Funktion $qt(p, df)$ von R berechnen, wobei $df = n$ die Anzahl der Freiheitsgrade der t -Verteilung bezeichnet. Mit R erhält man z. B.

```
> qt(.975, df = 10)
[1] 2.228139
```

Ab etwa $n = 30$ lassen sich die Quantile der t -Verteilung gut durch die entsprechenden Quantile z_p der Standardnormalverteilung approximieren.

Tab. 19.5: Quantile der t -Verteilung

n	p							
	0,800	0,850	0,900	0,950	0,975	0,990	0,995	
1	1,376	1,963	3,078	6,314	12,706	31,821	63,657	
2	1,061	1,386	1,886	2,920	4,303	6,965	9,925	
3	0,979	1,250	1,638	2,353	3,182	4,541	5,841	
4	0,941	1,190	1,533	2,132	2,776	3,747	4,604	
5	0,920	1,156	1,476	2,015	2,571	3,365	4,032	
6	0,906	1,134	1,440	1,943	2,447	3,143	3,707	
7	0,896	1,119	1,415	1,895	2,365	2,998	3,499	
8	0,889	1,108	1,397	1,860	2,306	2,896	3,355	
9	0,883	1,100	1,383	1,833	2,262	2,821	3,250	
10	0,879	1,093	1,372	1,812	2,228	2,764	3,169	
11	0,876	1,088	1,363	1,796	2,201	2,718	3,106	
12	0,873	1,083	1,356	1,782	2,179	2,681	3,055	
13	0,870	1,080	1,350	1,771	2,160	2,650	3,012	
14	0,868	1,076	1,345	1,761	2,145	2,624	2,977	
15	0,866	1,074	1,341	1,753	2,131	2,602	2,947	

Fortsetzung nächste Seite

Tab. 19.5: Quantile der t -Verteilung

n	p						
	0,800	0,850	0,900	0,950	0,975	0,990	0,995
16	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,854	1,055	1,310	1,697	2,042	2,457	2,750
31	0,853	1,054	1,310	1,696	2,040	2,453	2,744
32	0,853	1,054	1,309	1,694	2,037	2,449	2,739
33	0,853	1,053	1,308	1,692	2,035	2,445	2,733
34	0,852	1,053	1,307	1,691	2,032	2,441	2,728
35	0,852	1,052	1,306	1,690	2,030	2,438	2,724
36	0,852	1,052	1,306	1,688	2,028	2,435	2,720
37	0,851	1,051	1,305	1,687	2,026	2,431	2,715
38	0,851	1,051	1,304	1,686	2,024	2,429	2,712
39	0,851	1,050	1,304	1,685	2,023	2,426	2,708
40	0,851	1,050	1,303	1,684	2,021	2,423	2,705

19.6 Quantile der F-Verteilung

Die folgende Tabelle weist **Quantile** $F_{m;n;p}$ einer F -Verteilung mit m und n Freiheitsgraden für ausgewählte Werte von m und n und $p = 0,95$ sowie $p = 0,99$ aus. Der Tabelle entnimmt man z. B., dass das im linken Teil von Abbildung 13.9 visualisierte 0,95-Quantil der F -Verteilung mit $m = 10$ und $n = 15$ Freiheitsgraden den Wert $F_{10;15;0,95} = 2,54$ hat. Weitere Quantile der F -Verteilung sind über das nebenstehende interaktive Lernobjekt sowie über die Funktion $qf(p, df1, df2)$ von R zugänglich. Dabei bezeichnen $df1 = m$ und $df2 = n$ die beiden Freiheitsgrade der F -Verteilung. Mit R erhält man so



Interaktives Objekt

```
> qf(.95, 10, 15)
[1] 2.543719
```

„Quantile der
 F -Verteilung“

Bei Rundung auf nur zwei Dezimalstellen deckt sich das Ergebnis mit dem tabellierten Wert.

Tab. 19.6: Quantile der F -Verteilung ($p = 0,95$, $m = 1$ bis $m = 10$)

n	m									
	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,14	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35

Fortsetzung nächste Seite

Tab. 19.6: Quantile der F -Verteilung ($p = 0,95$, $m = 1$ bis $m = 10$)

n	m									
	1	2	3	4	5	6	7	8	9	10
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03

Tab. 19.6: Quantile der F -Verteilung ($p = 0,95$, $m > 11$)

n	m									
	11	12	13	14	15	20	30	40	50	100
1	243	244	245	245	246	248	250	251	252	253
2	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5
3	8,76	8,74	8,73	8,71	8,70	8,66	8,62	8,59	8,58	8,55
4	5,94	5,91	5,89	5,87	5,86	5,80	5,75	5,72	5,70	5,66
5	4,70	4,68	4,66	4,64	4,62	4,56	4,50	4,46	4,44	4,41
6	4,03	4,00	3,98	3,96	3,94	3,87	3,81	3,77	3,75	3,71
7	3,60	3,57	3,55	3,53	3,51	3,44	3,38	3,34	3,32	3,27
8	3,31	3,28	3,26	3,24	3,22	3,15	3,08	3,04	3,02	2,97
9	3,10	3,07	3,05	3,03	3,01	2,94	2,86	2,83	2,80	2,76
10	2,94	2,91	2,89	2,86	2,85	2,77	2,70	2,66	2,64	2,59
11	2,82	2,79	2,76	2,74	2,72	2,65	2,57	2,53	2,51	2,46
12	2,72	2,69	2,66	2,64	2,62	2,54	2,47	2,43	2,40	2,35
13	2,63	2,60	2,58	2,55	2,53	2,46	2,38	2,34	2,31	2,26
14	2,57	2,53	2,51	2,48	2,46	2,39	2,31	2,27	2,24	2,19
15	2,51	2,48	2,45	2,42	2,40	2,33	2,25	2,20	2,18	2,12
16	2,46	2,42	2,40	2,37	2,35	2,28	2,19	2,15	2,12	2,07
17	2,41	2,38	2,35	2,33	2,31	2,23	2,15	2,10	2,08	2,02

Fortsetzung nächste Seite

Tab. 19.6: Quantile der F -Verteilung ($p = 0,95, m > 11$)

n	m									
	11	12	13	14	15	20	30	40	50	100
18	2,37	2,34	2,31	2,29	2,27	2,19	2,11	2,06	2,04	1,98
19	2,34	2,31	2,28	2,26	2,23	2,16	2,07	2,03	2,00	1,94
20	2,31	2,28	2,25	2,22	2,20	2,12	2,04	1,99	1,97	1,91
21	2,28	2,25	2,22	2,20	2,18	2,10	2,01	1,96	1,94	1,88
22	2,26	2,23	2,20	2,17	2,15	2,07	1,98	1,94	1,91	1,85
23	2,24	2,20	2,18	2,15	2,13	2,05	1,96	1,91	1,88	1,82
24	2,22	2,18	2,15	2,13	2,11	2,03	1,94	1,89	1,86	1,80
25	2,20	2,16	2,14	2,11	2,09	2,01	1,92	1,87	1,84	1,78
26	2,18	2,15	2,12	2,09	2,07	1,99	1,90	1,85	1,82	1,76
27	2,17	2,13	2,10	2,08	2,06	1,97	1,88	1,84	1,81	1,74
28	2,15	2,12	2,09	2,06	2,04	1,96	1,87	1,82	1,79	1,73
29	2,14	2,10	2,08	2,05	2,03	1,94	1,85	1,81	1,77	1,71
30	2,13	2,09	2,06	2,04	2,01	1,93	1,84	1,79	1,76	1,70
40	2,04	2,00	1,97	1,95	1,92	1,84	1,74	1,69	1,66	1,59
50	1,99	1,95	1,92	1,89	1,87	1,78	1,69	1,63	1,60	1,52

Tab. 19.6: Quantile der F -Verteilung ($p = 0,99, m = 1$ bis $m = 10$)

n	m									
	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056
2	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94

Fortsetzung nächste Seite

Tab. 19.6: Quantile der F -Verteilung ($p = 0,99$, $m = 1$ bis $m = 10$)

n	m									
	1	2	3	4	5	6	7	8	9	10
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70

Tab. 19.6: Quantile der F -Verteilung ($p = 0,99$, $m > 11$)

n	m									
	11	12	13	14	15	20	30	40	50	100
1	6083	6107	6126	6143	6157	6209	6260	6286	6302	6334
2	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5
3	27,1	27,1	27,0	26,9	26,9	26,7	26,5	26,4	26,4	26,2
4	14,5	14,4	14,3	14,2	14,2	14,0	13,8	13,7	13,7	13,6
5	9,96	9,89	9,82	9,77	9,72	9,55	9,38	9,29	9,24	9,13
6	7,79	7,72	7,66	7,60	7,56	7,40	7,23	7,14	7,09	6,99
7	6,54	6,47	6,41	6,36	6,31	6,16	5,99	5,91	5,86	5,75
8	5,73	5,67	5,61	5,56	5,52	5,36	5,20	5,12	5,07	4,96
9	5,18	5,11	5,05	5,01	4,96	4,81	4,65	4,57	4,52	4,41
10	4,77	4,71	4,65	4,60	4,56	4,41	4,25	4,17	4,12	4,01
11	4,46	4,40	4,34	4,29	4,25	4,10	3,94	3,86	3,81	3,71

Fortsetzung nächste Seite

Tab. 19.6: Quantile der F -Verteilung ($p = 0,99, m > 11$)

n	m									
	11	12	13	14	15	20	30	40	50	100
12	4,22	4,16	4,10	4,05	4,01	3,86	3,70	3,62	3,57	3,47
13	4,02	3,96	3,91	3,86	3,82	3,66	3,51	3,43	3,38	3,27
14	3,86	3,80	3,75	3,70	3,66	3,51	3,35	3,27	3,22	3,11
15	3,73	3,67	3,61	3,56	3,52	3,37	3,21	3,13	3,08	2,98
16	3,62	3,55	3,50	3,45	3,41	3,26	3,10	3,02	2,97	2,86
17	3,52	3,46	3,40	3,35	3,31	3,16	3,00	2,92	2,87	2,76
18	3,43	3,37	3,32	3,27	3,23	3,08	2,92	2,84	2,78	2,68
19	3,36	3,30	3,24	3,19	3,15	3,00	2,84	2,76	2,71	2,60
20	3,29	3,23	3,18	3,13	3,09	2,94	2,78	2,69	2,64	2,54
21	3,24	3,17	3,12	3,07	3,03	2,88	2,72	2,64	2,58	2,48
22	3,18	3,12	3,07	3,02	2,98	2,83	2,67	2,58	2,53	2,42
23	3,14	3,07	3,02	2,97	2,93	2,78	2,62	2,54	2,48	2,37
24	3,09	3,03	2,98	2,93	2,89	2,74	2,58	2,49	2,44	2,33
25	3,06	2,99	2,94	2,89	2,85	2,70	2,54	2,45	2,40	2,29
26	3,02	2,96	2,90	2,86	2,81	2,66	2,50	2,42	2,36	2,25
27	2,99	2,93	2,87	2,82	2,78	2,63	2,47	2,38	2,33	2,22
28	2,96	2,90	2,84	2,79	2,75	2,60	2,44	2,35	2,30	2,19
29	2,93	2,87	2,81	2,77	2,73	2,57	2,41	2,33	2,27	2,16
30	2,91	2,84	2,79	2,74	2,70	2,55	2,39	2,30	2,25	2,13
40	2,73	2,66	2,61	2,56	2,52	2,37	2,20	2,11	2,06	1,94
50	2,63	2,56	2,51	2,46	2,42	2,27	2,10	2,01	1,95	1,82



20 Übungsaufgaben und Lösungen



In diesem Kapitel sind etwa 50 Übungsaufgaben mit ausführlichen Lösungen zusammengestellt. Die Aufgaben sollen die Inhalte von Teil I - II dieses Buches vertiefen und zur Verständnissicherung beitragen.

Vorschau auf das Kapitel

Zu allen Aufgaben, die auch mit der freien Software R gelöst werden können, ist unter der Webadresse <https://statistiklehrbuch.statup.solutions> der R-Code für die Lösungen eingestellt.

20.1 Übungsaufgaben zu Teil I



Kapitel 1

Aufgabe 1.1 (Statistical Literacy)

Die Allbright-Studie präsentiert jährlich Fakten zu Frauen und Diversität in den Führungspositionen der Wirtschaft. Der [Studie](#) für das Jahr 2018 mit dem Titel „Die Macht der Monokultur“ zu entnehmen, dass sich der Frauenanteil in den Vorständen der 160 deutschen börsennotierten Unternehmen gegenüber dem Vorjahr um 0,7 Prozentpunkte erhöht hatte.

Sie erhalten folgende zusätzlichen Informationen:

- Information 1: Der Frauenanteil ist um rund 9,75 Prozent gestiegen.
- Information 2: Der Frauenanteil betrug zuvor rund 7,3 Prozent.

Frage 1: Wie groß ist der Frauenanteil in den Vorständen heute?

Frage 2: Wie viele Frauen sind heute in den Vorständen?

Welche der nachstehenden Aussagen treffen zu:

- a) Information 1 reicht aus, um Frage 1 zu beantworten.
- b) Information 2 reicht aus, um Frage 1 zu beantworten.
- c) 1 und 2 werden zusammen benötigt, um beide Fragen zu beantworten.
- d) 1 oder 2 alleine reichen jeweils aus, um Frage 2 zu beantworten.
- e) 1 und 2 zusammen reichen nicht aus, um Frage 2 zu beantworten.

Aufgabe 1.2 (Statistical Literacy)

Unter Bezugnahme auf die in Aufgabe 1.1 erwähnte Allbright-Studie „Die Macht der Monokultur“ titelte das [Manager-Magazin](#) am 30. September 2018: „Der Frauenanteil in den Vorständen ist gestiegen (um 0,7 Prozent)“.

- a) Ist die Aussage der Schlagzeile korrekt? Begründen Sie Ihre Antwort.
- b) Handelt es sich Ihrer Meinung nach um eine starke oder eine schwache Veränderung? Begründen Sie Ihre Antwort.

- c) Würde sich ihre Antwort zu Aufgabenteil b verändern, wenn die Schlagzeile gelautet hätte: "Der Frauenanteil in den Vorständen steigt (um 9,75 Prozent)."
- d) Wie würden Sie die Schlagzeile formulieren, um zu signalisieren, dass Sie die Frauenquote befürworten?
- e) Wie würden Sie die Schlagzeile formulieren, um zu signalisieren, dass Sie die Frauenquote ablehnen?



Aufgabe 2.1 (Grundbegriffe)

Kapitel 2 Ein Marktforschungsinstitut untersucht das Fernsehverhalten von Schulkindern in Deutschland. Die Untersuchung soll u. a. Aufschluss darüber geben, wie lange und zu welchen Tageszeiten Kinder durchschnittlich Fernsehen gucken und welche Sender sie bevorzugen.

Was sind hier Grundgesamtheit, statistische Einheit, Merkmal und Merkmalsausprägung? Wie könnte man bezüglich der Grundgesamtheit durch Bildung von Teilgrundgesamtheiten differenzieren? Welche Teilmengen der Grundgesamtheit könnten für die Untersuchung noch von Interesse sein?

Aufgabe 2.2 (Skalenarten)

Nachstehend sind vier Merkmale aufgeführt. Geben Sie bei jedem Merkmal an, welcher der Skalentypen „Nominalskala“, „Ordinalskala“ bzw. „Metrische Skala“ zutrifft. Der Begriff „Metrische Skala“ wird als Oberbegriff für Intervallskala, Verhältnisskala und Absolutskala verwendet.

- Höchster erreichter Schulabschluss (Ausprägungen: ohne Abschluss, Hauptschule, mittlere Reife, Fachhochschulreife, Abitur)
- Gewählte Partei bei einer Kommunalwahl (Ausprägungen: zwei freie Wählervereinigungen und alle im Landtag vertretenen Parteien)
- Bonität von Kunden einer Sparkasse (Kategorien: uneingeschränkte, eingeschränkte und fehlende Kreditwürdigkeit)
- Verfallsdatum bei einer Konfitürensorte (Tag der Herstellung + 18 Monate; auf der Ware angegeben).



Aufgabe 3.1 (Zeitreihen in den Medien)

Kapitel 3 Geben Sie einige Beispiele für Zeitreihen an, die regelmäßig in den Medien zu finden sind.

Aufgabe 3.2 (geschichtete Zufallsauswahl)

Von 600 Studierenden, die sich in einem erst 3 Semester laufenden Bachelor-Studiengang eingeschrieben haben, sollen 120 zufällig für eine Befragung ausgewählt werden. Als Schichtungskriterium wird die Semesterzahl verwendet. Es sind 270 Studierende im 1. Semester, 180 im 2. Semester und 150 im 3. Semester. Welchen Umfang haben die drei Schichten bei proportionaler Schichtung?


Aufgabe 4.1 (Ergebnisse der Nationalen Verzehrstudie II für Männer)

In der sog. **Nationalen Verzehrstudie II** wurde vom Max Rubner-Institut von Ende 2005 bis Anfang 2007 eine etwa 20 000 umfassende Stichprobe der deutschen Bevölkerung nach ihrem Ernährungsverhalten befragt. Dabei wurde anhand des Body-Mass-Index BMI (vgl. hierzu Beispiel 8.1) u. a. der Anteil der Unter- oder Normalgewichtigen ($BMI < 25$), der Übergewichtigen ($25 \leq BMI < 30$) und der Fettleibigen ($BMI \geq 30$) ermittelt. Die nachstehende Tabelle bezieht sich auf das diskrete Merkmal „Gewichtsstatus“ mit den drei Ausprägungen a_1 , a_2 und a_3 (a_1 : Unter- oder Normalgewicht, a_2 : Übergewicht, a_3 : Fettleibigkeit). Die nachstehende Tabelle zeigt die absoluten und relativen Häufigkeiten dieser Ausprägungen für die an der Studie beteiligten Männer.

Bundesland (männliche Teilnehmer)	Absolute und relative Häufigkeiten					
	$h(a_1)$	$f(a_1)$	$h(a_2)$	$f(a_2)$	$h(a_3)$	$f(a_3)$
Baden-Württemberg (846)	264	0,312	408	0,482	174	0,206
Bayern (1018)	345	0,339	455	0,447	218	0,214
Berlin (218)	74	0,339	104	0,477	40	0,184
Brandenburg (164)	51	0,311	71	0,433	42	0,256
Bremen (62)	24	0,387	29	0,468	9	0,145
Hamburg (91)	35	0,385	42	0,462	14	0,154
Hessen (456)	140	0,307	220	0,483	96	0,211
Mecklenburg-Vorp. (87)	28	0,322	38	0,437	21	0,241
Niedersachsen (750)	242	0,323	338	0,451	170	0,227
Nordrhein-Westf. (1237)	405	0,327	583	0,471	249	0,201
Rheinland-Pfalz (315)	101	0,321	155	0,492	59	0,187
Saarland (71)	24	0,338	37	0,521	10	0,141
Sachsen (302)	96	0,318	136	0,450	70	0,232
Sachsen-Anhalt (136)	42	0,309	65	0,478	29	0,213
Schleswig-Holstein (202)	64	0,317	89	0,441	49	0,243
Thüringen (162)	50	0,309	74	0,457	38	0,235
Summe: 6117	1985		2844		1288	

Stellen Sie die relativen Häufigkeiten in Form gestapelter Balkendiagramme dar. Unterdrücken Sie dabei die Wiedergabe der Häufigkeiten $f(a_1)$ und ordnen Sie die Bundesländer nach zunehmender Größe der Summe

$$f(a_2) + f(a_3) = 1 - f(a_1).$$

Aufgabe 4.2 (Gruppierung von Daten und Histogrammerstellung)

Für 80 Arbeitnehmer in Singapur wurden für das Referenzjahr 2019 folgende Bruttostundenverdienste ermittelt (in Euro und auf eine Dezimalstelle gerundet), hier nach aufsteigender Größe sortiert:

3,8	4,0	4,6	5,0	5,1	5,2	5,2	5,7	5,9	6,2
6,4	6,8	6,8	7,0	7,1	7,2	7,3	7,4	7,5	7,5
7,8	7,9	8,1	8,3	8,4	8,7	8,9	9,0	9,3	9,4
9,4	9,5	9,6	9,6	9,8	9,9	10,8	11,9	12,0	12,5
12,7	12,9	13,0	13,2	13,4	13,5	13,9	14,0	14,2	14,6
14,9	15,4	15,8	16,4	17,6	17,9	17,9	18,2	18,3	19,1
19,9	20,5	21,8	23,0	23,7	24,1	24,6	26,9	27,1	28,9
29,8	32,0	33,8	34,8	36,7	39,1	43,2	45,4	50,3	60,7

- a) Was sind hier Merkmalsträger und Merkmal?
- b) Ordnen Sie die obigen Individualdaten 15 Einkommensklassen zu, in dem Sie die Daten zu Intervallen von 5 Euro gruppieren – analog zu Abbildung 4.8, die sich allerdings auf Bruttojahresverdienste bezog und daher Intervalle von 5 000 Euro vorsah. Ermitteln Sie dann für das Merkmal „Bruttostundenverdienste“ die absoluten und die relativen Häufigkeiten für die Besetzung der Einkommensklassen, letztere in Prozent. Fertigen Sie eine Tabelle an, die in jeder Zeile eine Klasse sowie die zugehörige absolute und relative Häufigkeit für die Besetzung dieser Klasse ausweist.
- c) Visualisieren Sie auf der Basis obiger Klasseneinteilung die relativen Klassenbesetzungshäufigkeiten anhand eines Histogramms.

Aufgabe 4.3 (empirische Verteilungsfunktion)

Mit drei Würfeln wird 100-mal gewürfelt und jeweils die Augensumme ermittelt. Der Ausgang des Würfelexperiments lässt sich anhand der relativen Häufigkeiten für die beobachteten Ausprägungen des Merkmals „Augensumme“ charakterisieren oder – analog zum unteren Teil von Abbildung 4.14 – anhand der empirischen Verteilungsfunktion.

- a) Welche Ausprägungen kann das Merkmal „Augensumme“ annehmen?
- b) Wieviele Sprünge kann die empirische Verteilungsfunktion maximal aufweisen?



Aufgabe 5.1 (Häufigkeitsverteilungen; Kenngrößen)

Nachstehend ist das Ergebnis eines Würfelexperiments wiedergegeben, bei dem Kapitel 5 12 Mal nacheinander mit einem Würfel gewürfelt wurde:



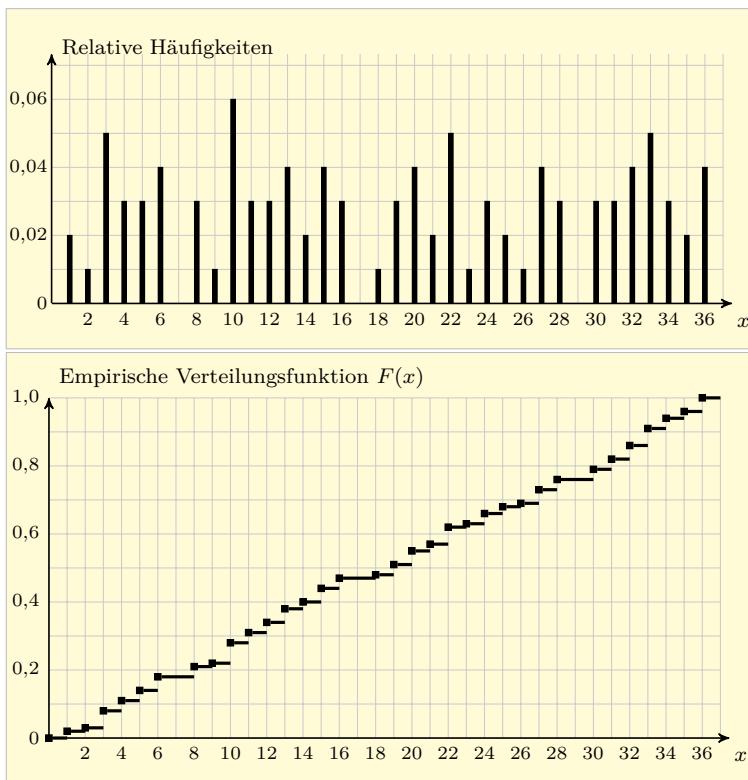
- a) Geben Sie für die 6 Merkmalsausprägungen die absoluten und die relativen Häufigkeiten an. Runden Sie die relativen Häufigkeiten auf 3 Stellen nach dem Komma oder verwenden Sie Brüche.
- b) Berechnen Sie für die durch die obigen 12 Augenzahlen definierte Urliste den Median und, auf 2 Nachkommastellen genau, den Mittelwert.
- c) Berechnen Sie für den obigen Datensatz mit 12 Elementen die Spannweite, die Varianz und die Standardabweichung. Die Ergebnisse sind auf 3 Stellen nach dem Dezimalkomma genau anzugeben.

Aufgabe 5.2 (Quantile und Boxplots)

- Bestimmen Sie für den in Aufgabe 5.1 veranschaulichten Datensatz mit 12 Werten (Würfelexperiment) die Quartile $x_{0,25}$ und $x_{0,75}$.
- Die 12 Werte lassen sich anhand eines Boxplots visualisieren. Geben Sie die 5 Größen an, durch die der Boxplot definiert ist. Wie groß ist der Interquartilsabstand Q , der die Länge der Box festlegt?
- Wenn noch einmal gewürfelt wird und die Augenzahl 3 erscheint, hat man einen Datensatz der Länge $n = 13$. Wie groß ist nun Q ?

Aufgabe 5.3 (Quantile und Boxplots)

Nachstehend ist die Häufigkeitsverteilung für die Ausgänge bei einer Serie von 100 Roulettespielen dargestellt:



- Bestimmen Sie den Median und Interquartilsabstand des Datensatzes.
- Welchen Wert nimmt die empirische Verteilungsfunktion für $x = 2$ und für $x = 35$ an?
- Visualisieren Sie den Datensatz anhand eines Boxplots.

**Aufgabe 6.1** (Kaplan-Meier-Verfahren)

Die folgende Tabelle zeigt die Ergebnisse einer fiktiven Studie, in der das progressionsfreie Überleben von Krebspatientinnen erhoben wurde, also die Zeit, in der die Erkrankung nicht fortschreitet. Für jede Patientin sind ihre ID, die Zeit bis zum Eintreten der Progression und der Status „1“ (Progression oder Tod) bzw. „0“ (keine Progression) erhoben.

Pat.-ID	progressionsfreie Überlebenszeit	Status
1	5 Monate	1
2	9 Monate	1
3	10 Monate	1
4	16 Monate	1
5	20 Monate	0

- Veranschaulichen Sie die Überlebenszeiten anhand eines zu Abbildung 6.1 analogen Diagramms.
- Berechnen Sie die Werte der Kaplan-Meier-Kurve an den Sprungstellen und zeichnen Sie die Kurve.
- Bestimmen Sie den Median der progressionsfreien Überlebenszeit.
- Um wieviel Prozent reduziert sich pro Ereigniszeitpunkt die noch unter Risiko stehende Patientinnengruppe ?

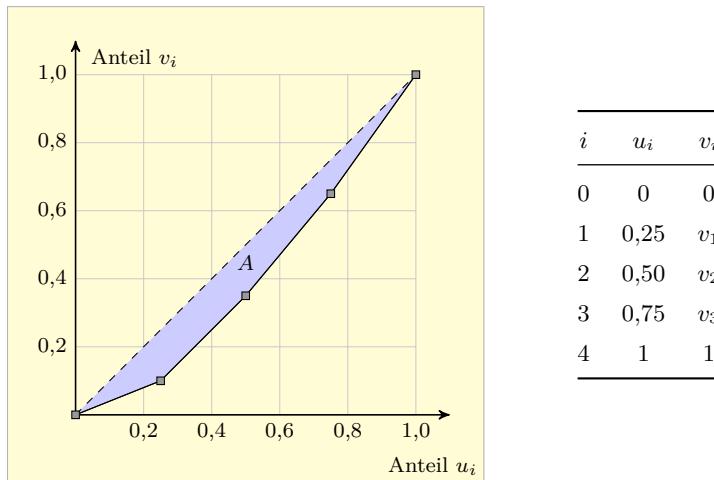
Aufgabe 6.2 (Kaplan-Meier-Verfahren)

An einer medizinischen Studie, bei der es um die Überlebensdauer nach Lungentransplantationen geht, nehmen 5 Personen teil. Die beobachteten Überlebenszeiten t_1, \dots, t_5 sind nach Größe geordnet.

- Bestimmen Sie die Werte $\hat{S}(t_1)$, $\hat{S}(t_2)$ und $\hat{S}(t_3)$ der Kaplan-Meier-Kurve an den ersten drei Ereigniszeitpunkten unter der Annahme, dass jedes Mal genau ein Ereignis eintritt ($d_1 = d_2 = d_3$).
- Wie lautet der Wert $\hat{S}(t_5)$, wenn eine Person unmittelbar davor, also in t_4 , aus der Untersuchung ausgeschieden ist (Zensierung in t_4)?

**Aufgabe 7.1** (Gini-Koeffizient)

Es seien $x_1 = 20$, $x_2 = 50$, $x_3 = 60$ und $x_4 = 70$ die Umsätze von vier Energieversorgungsunternehmen (in Millionen Euro) im letzten Geschäftsjahr. Die folgende Abbildung zeigt die auf der Basis dieser Daten errechnete Lorenzkurve, bei der die Stützpunkte (u_i, v_i) betont sind. In der Tabelle neben der Grafik sind die Abszissenwerte u_i schon eingetragen.



- a) Errechnen Sie die Ordinatenwerte v_1 , v_2 und v_3 sowie den Gini-Koeffizienten G aus (7.5) und den normierten Gini-Koeffizienten G^* aus (7.7).
- b) Welchen Inhalt hat die farbig markierte Fläche A ?

Aufgabe 7.2 (Herfindahl-Index)

- a) Berechnen Sie mit den Daten aus Aufgabe 7.1 den Herfindahl-Index.
- b) Wie groß ist hier die untere Schranke für den Index?

Aufgabe 8.1 (Militärausgaben 2014 im Ländervergleich)

Die nachstehende Tabelle zeigt Daten des Stockholmer Friedensforschungsinstituts *SIPRI* zu Militärausgaben in 12 Ländern für das Referenzjahr 2014.

Kapitel 8



Rang	Nation	Militärausgaben		
		absolut (Mrd. US-Dollar)	in % des BIP	pro Kopf
1.	USA (US)	609,9	3,5	1891
2.	China (CN)	216,4	2,1	155
3.	Russland (RU)	84,5	4,5	593
4.	Saudi-Arabien (SA)	80,8	10,4	2747
5.	Frankreich (FR)	62,3	2,2	964
6.	Großbritannien (UK)	60,5	2,2	952
7.	Indien (IN)	50,0	2,4	39
8.	Deutschland (DE)	46,5	1,2	562
9.	Japan (JP)	45,8	1,0	360
10.	Brasilien (BR)	31,7	1,5	157
11.	Israel (IL)	15,9	5,2	2040
12.	Singapur (SG)	9,8	3,7	1789

In der Tabelle sind die absoluten Ausgaben ausgewiesen (in Milliarden US-Dollar), der Anteil dieser Werte am BIP (in %) sowie die Ausgaben pro Kopf

(in US-Dollar). Die Länder sind nach absteigender Größe der absoluten Werte geordnet. In jeder Datenspalte sind die Extremwerte betont.

Veranschaulichen Sie die Werte in den letzten drei Spalten anhand je eines Säulendiagramms. Ordnen Sie die Werte jeder Spalte zuvor nach absteigender Größe und markieren Sie in den Grafiken jeweils den Balken für China. Verwenden Sie für die Ländernamen die angegebenen internationalen Codes.

Aufgabe 8.2 (Zusammengesetzte Indexzahlen – Medaillenspiegel)

Tabelle 8.1 zeigte die ersten zehn Länder beim Medaillenspiegel für die Olympiade 2008. Die beiden wiedergegebenen alternativen Rangfolgen unterschieden sich hinsichtlich der Gewichtung von Gold, Silber und Bronze. Beim ersten Ranking wurde nur Gold berücksichtigt (Gewichte 1 – 0 – 0), beim zweiten alle Medaillen mit gleichem Gewicht (1 – 1 – 1).

- Wie sähe für die zehn Länder der Tabelle 8.2 die Rangfolge aus, wenn man alle Medaillenarten berücksichtigte, aber mit unterschiedlichen Gewichten (5 – 3 – 2), also jede Goldmedaille mit 5 Punkten, jede Silbermedaille mit 3 Punkten und jede Bronzemedaille mit 2 Punkten bewertete?
- Wie sähe die Rangfolge für die zehn Länder aus, wenn man zwar den Ansatz 5 – 3 – 2 verwendete, die Punktzahlen aber auf die *Anzahl der Punkte pro 1 Million Einwohner* bezöge? Gehen Sie dabei von folgenden Bevölkerungszahlen aus (in Millionen; Daten des US Census Bureau für 2008): China – 1330,0; USA – 303,8; Russland – 140,7; Japan – 127,3; Deutschland – 82,4; Frankreich – 64,1; Italien – 58,1; Südkorea – 48,4; Australien – 21,0; Großbritannien – 60,9.

Aufgabe 8.3 (Preisindex)

Aktivieren Sie den **Inflationsrechner** des Statistischen Bundesamts. Wählen Sie über die Schaltfläche „Güterauswahl“ die Gütergruppe „Pauschalreisen“ aus. Welche Auffälligkeiten beobachten Sie?



Aufgabe 9.1 (Randverteilungen)

Bei einer medizinischen Studie wurde für $n = 360$ Personen erfasst, ob sie regelmäßig einen erhöhten Alkoholkonsum hatten und ob sie Leberfunktionsstörungen aufwiesen (adaptiert aus TOUTENBURG / SCHOMAKER / WISSMANN (2009)). Es sei X das Merkmal „Alkoholkonsum“ mit den Ausprägungen a_1 (Konsum oberhalb eines definierten Schwellenwerts) und a_2 (nicht oberhalb des Schwellenwerts) und Y das Merkmal „Leberstatus“ mit den Ausprägungen b_1 (Funktionsstörungen vorhanden) und b_2 (keine Funktionsstörungen).

	b_1	b_2
a_1	62	96
a_2	14	188

Ergänzen Sie diese Vierfeldertafel um die beiden Randverteilungen.

Aufgabe 9.2 (Bedingte Häufigkeitsverteilungen)

Interpretieren Sie die Werte für die in Beispiel 9.3 errechneten bedingten Häufigkeiten $f_X(a_5|b_1) \approx 0,107$ und $f_Y(b_1|a_2) \approx 0,620$.



Kapitel 10

Aufgabe 10.1 (Zusammenhangsmessung bei Nominalskalierung)

- Berechnen Sie den χ^2 -Koeffizienten auf der Basis der Daten aus Aufgabe 9.1. Runden Sie das Ergebnis auf drei Dezimalstellen.
- Bestimmen Sie dann auch den Phi-Koeffizienten Φ aus (10.3) und den in (10.5) eingeführten Kontingenzkoeffizienten V nach Cramér.

Aufgabe 10.2 (Zusammenhangsmessung bei metrischer Skalierung)

Das folgende Beispiel stammt aus BAMBERG / BAUR / KRAPP (2017):

Für 10 Staaten i , deren Namen codiert sind (z. B. „AT“ für „Austria“), sind für ein bestimmtes Referenzjahr Wertepaare (x_i, y_i) bekannt, wobei x_i Ausprägungen des Merkmals X (= Preisanstieg in %) und y_i Ausprägungen des Merkmals Y (= Erwerbslosenquote in %) bezeichnen:

Land i	x_i	y_i
BE	4,1	10,1
DE	2,4	4,0
UK	8,4	5,7
IE	8,2	10,2
IT	11,9	7,5
JP	4,6	2,1
CA	9,4	8,0
AT	3,6	1,3
SE	10,6	2,2
US	7,9	6,3

Mittelwerte: $\bar{x} = 7,11$ $\bar{y} = 5,74$

Berechnen Sie den Korrelationskoeffizienten r . Für eine manuelle Berechnung können Sie eine zu Tabelle 17.2 analoge Arbeitstabelle verwenden.

Aufgabe 10.3 (Zusammenhangsmessung bei ordinaler Skalierung)

Das folgende Beispiel findet man bei TOUTENBURG / HEUMANN (2009):

Fünf mit A, B, …, E bezeichnete Mannschaften bestreiten ein Handballturnier im Winter in der Halle und im Sommer im Freien. Nachstehend sind die Platzierungen bei beiden Turnieren wiedergegeben. Untersuchen Sie anhand des Rangkorrelationskoeffizienten von Spearman, ob zwischen dem Abschneiden der Mannschaften in der Halle und im Freien ein Zusammenhang besteht.

Mannschaft	Platzierung (Hallenturnier)	Platzierung (Frei- luftturnier)
A	1	2
B	2	3
C	3	1
D	4	5
E	5	4

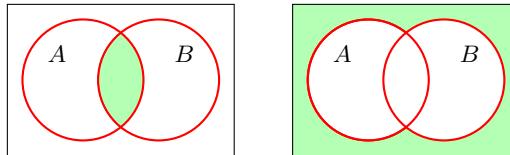
20.2 Übungsaufgaben zu Teil II



Aufgabe 11.1 (Venn-Diagramme)

Nachstehend sind zwei Venn-Diagramme abgebildet, die sich auf zwei Ereignisse oder Mengen A und B beziehen. Welche der folgenden Aussagen sind richtig?

Es bezeichnen \bar{A} und \bar{B} die Komplementärmengen von A und B , $A \cap B$ deren Schnittmenge und $A \cup B$ die Vereinigungsmenge.



- a) Im ersten Venn-Diagramm ist anhand der dunkler gefärbten Fläche die Schnittmenge $A \cap B$ von A und B dargestellt.
- b) Im zweiten Venn-Diagramm ist die Schnittmenge $A \cap \bar{B}$ aus A und der Komplementärmenge von B dargestellt.
- c) Die Schnittmengen $A \cap B$ und $A \cap \bar{B}$ sind disjunkt, d. h. ihre Darstellungen in Venn-Diagrammen überschneiden sich nicht.
- d) Die Vereinigung von $A \cap B$ und $A \cap \bar{B}$ liefert A .
- e) Die Wahrscheinlichkeit für den Eintritt eines aus zwei disjunkten Ereignissen zusammensetzten Ereignisses A ergibt sich als Summe der Wahrscheinlichkeiten der beiden Ereignisse.

Aufgabe 11.2 (Ereignisse und Ereignisraum)

Eine Münze wird dreimal nacheinander geworfen. Der Ausgang des dreifachen Münzwurfs ist durch ein Tripel definiert, das aus den Symbolen „Z“ (Zahl) und „K“ (Kopf) gebildet wird.

- a) Wie lautet die Ergebnismenge Ω für den dreifachen Münzwurf?
- b) Wieviele Elementarereignisse umfasst das Ereignis

$$A = \{\text{Bei mindestens zwei Würfen tritt „K“ auf}\}?$$

Aufgabe 11.3 (Wahrscheinlichkeiten bei Laplace-Experimenten)

- a) Bei dem Spiel *Mensch ärgere Dich nicht* kann jeder Spieler zu Beginn drei Mal würfeln. Sobald eine Sechs gewürfelt wird, darf eine Spielfigur starten. Spieler A beginnt. Wie groß ist die Wahrscheinlichkeit P dafür, dass bei Spieler A nach dem dritten Wurf noch keine Sechs gefallen ist? Setzen Sie voraus, dass der verwendete Würfel fair ist, also gleiche Eintrittswahrscheinlichkeiten für alle Augenzahlen aufweist.
- b) Berechnen Sie für das statistische Experiment „Dreifacher Münzwurf“ die Wahrscheinlichkeiten für die Ereignisse

$$A = \{\text{Bei mindestens zwei Würfen tritt „K“ auf}\}$$

$$\bar{A} = \{\text{Bei höchstens einem Wurf tritt „K“ auf}\}.$$

Setzen Sie eine faire Münze voraus, also gleiche Eintrittswahrscheinlichkeiten für die Ausgänge „Z“ (Zahl) und „K“ (Kopf).

Aufgabe 11.4 (Kombinatorik)

Eine Hochschule ordnet allen Studierenden eine mehrstellige Zahl zu (Matrikel-Nummer), die zu Identifikationszwecken verwendet wird. Wieviele Studierende könnte man maximal unterscheiden, wenn für jede Person anstelle einer Zahl 5 Großbuchstaben aus der Buchstabenfolge von A bis J verwendet würde, also z. B. BCBJD oder AFGGC?

Aufgabe 11.5 (Bedingte Wahrscheinlichkeiten und Vierfeldertafel)

An einer Bildungseinrichtung sind 160 Beschäftigte mit Hochschulabschluss in der Lehre tätig. Von diesem Personenkreis sind 64 vollzeitbeschäftigt (Ereignis A), 60 Personen sind promoviert (Ereignis B). Dabei kann für eine Person auch beides zutreffen. In der Tat sind von den 160 in der Lehre tätigen Beschäftigten mit Hochschulabschluss 40 Personen, für die beide Voraussetzungen zutreffen (Vollzeitbeschäftigung und Promotion). Es werde per Zufallsauswahl aus der Gruppe der 160 Lehrenden mit Hochschulabschluss eine Person ausgewählt.

- a) Wie groß ist dann die Wahrscheinlichkeit, dass die ausgewählte Person keine Vollzeitbeschäftigung hat?
- b) Berechnen Sie die Wahrscheinlichkeit dafür, dass sie sowohl eine Vollzeitbeschäftigung als auch eine abgeschlossene Promotion hat.
- c) Wie groß ist die Wahrscheinlichkeit, dass eine aus dem vollzeitbeschäftigten Lehrpersonal zufällig ausgewählte Person promoviert ist?
- d) Stellen Sie fest, ob die Ereignisse A und B unabhängig sind.
- e) Leiten Sie aus den obigen Vorgaben eine Vierfeldertafel für absolute Häufigkeiten ab. Die Tafel soll auch die Randverteilungen der binären Merkmale „Beschäftigtenstatus“ (Ausprägungen „Vollzeit“ A und „Teilzeit“ \bar{A}) sowie „Erreichter akademischer Grad“ (Ausprägungen „Hochschulabschluss mit Promotion“ B und „Hochschulabschluss ohne Promotion“ \bar{B}) ausweisen. Berechnen Sie die Wahrscheinlichkeiten aus den Aufgabenteilen a - c auch anhand dieser Tafel.

Aufgabe 11.6 (Bedingte Wahrscheinlichkeiten und Vierfeldertafel)

Es sei unterstellt, dass in einer größeren Grundgesamtheit von N Frauen etwa 0,8 % Brustkrebs haben (Gesundheitsstatus A) und der Krebs in 90 % der Fälle bei einer Mammographie entdeckt wird. Ferner sei angenommen, dass in der Teilpopulation ohne Erkrankung (Frauen mit Gesundheitsstatus \bar{A}) beim Screening im Mittel in 7 % aller Fälle ein falsch-positiver Befund resultiert.

- Welche Gestalt hat hier Tabelle 11.3, wenn man die vier dort grau hinterlegen Werte im Tabelleninneren als Vielfaches von N ausdrückt?
- Geben Sie Sensitivität und die Spezifität des Screeningsverfahrens an.
- Berechnen Sie die Wahrscheinlichkeiten $P(B)$ und $P(\bar{B})$ dafür, dass eine zufällig ausgewählte Frau einen positiven resp. einen negativen Befund erhält.
- Bestimmen Sie die Wahrscheinlichkeit $P(\bar{A}|B)$ dafür, dass eine Frau mit positivem Screeningbefund gesund ist, der Test also zu einem Fehlalarm führt (falsch-positiver Befund).
- Wie groß wären die Wahrscheinlichkeiten $P(B)$ und $P(\bar{A}|B)$, wenn man für die Erkrankungswahrscheinlichkeit $P(A)$ anstelle von 0,008 den Wert 0,006 voraussetzte?

Aufgabe 11.7 (Bedingte Wahrscheinlichkeiten und Baumdiagramm)

- Wie häufig man bei dem Screeningbeispiel aus Aufgabe 11.6 mit falsch-positiven Befunden (Fehlalarme) oder falsch-negativen Befunden (unterbliebene Alarne) zu rechnen hat, lässt sich auch anhand eines Baumdiagramms für absolute Häufigkeiten visualisieren. Zeichnen Sie unter Annahme von $N = 100\,000$ ein zu Abbildung 9.2 analoges Baumdiagramm, das sich auf absolute Häufigkeiten bezieht und als Zusatzinformation auch die relativen Häufigkeiten wiedergibt. Unterteilen Sie die Grundgesamtheit zunächst nach der Binärvariablen „Gesundheitsstatus“ mit den Ausprägungen „erkrankt“ (A) und „gesund“ (\bar{A}) und danach beide Teilmengen jeweils noch nach der Binärvariablen „Screeningbefund“ mit den Ausprägungen „positiv“ (B) und „negativ“ (\bar{B}).
- Wie sieht die Vierfeldertafel für absolute Häufigkeiten beim Screeningbeispiel aus Aufgabe 11.6 bei Wahl von $N = 100\,000$ aus?

**Aufgabe 12.1** (Würfeln mit zwei Würfeln)

Kapitel 12 Es werde mit zwei „fairen“ Würfeln gewürfelt, also solchen mit gleicher Eintrittswahrscheinlichkeit für jede Augenzahl, und die Summe X der beiden Augenzahlen festgestellt.

- Welche Ausprägungen sind für die Zufallsvariable X möglich? Welche Eintrittswahrscheinlichkeiten $f(x)$ besitzen die Ausprägungen?
- Welchen Wert hat die Verteilungsfunktion $F(x)$ der Augensumme X an den Stellen $x = 0,5, x = 3, x = 3,5$ und $x = 6$?
- Berechnen Sie auch den Erwartungswert von X .

Aufgabe 12.2 (Binomialverteilung)

In der Fußgängerzone einer Stadt ist ein Glücksrad installiert. Dieses ist in vier gleich große Teile unterteilt, die farblich unterschieden sind. Interessierte Passanten dürfen das Rad einmal drehen und erhalten in Abhängigkeit von der Farbe des am Ende oben stehenden Sektors einen Preis. Wenn der Sektor „Rot“ oben steht, gibt es einen Luftballon, bei „Gelb“ einen Kugelschreiber, bei „Blau“ ein Freiexemplar der aktuellen Ausgabe einer Tageszeitung und bei „Grün“ eine kostenlose Zustellung der Zeitung für eine ganze Woche.

Eine 4-köpfige Familie bleibt vor dem Glücksrad stehen. Jedes Familienmitglied betätigt es einmal. Wie groß ist die Wahrscheinlichkeit, dass bei den 4 Versuchen

- mindestens zwei Kugelschreiber gewonnen werden?
- genau einmal „Grün“ auftritt, also ein einwöchiges Freiabonnement gewährt wird?

Aufgabe 12.3 (Hypergeometrische Verteilung)

Aus einer Urne mit 10 Kugeln, die sich nur hinsichtlich der Farbe unterscheiden, werden nacheinander $n = 5$ Kugeln *ohne Zurücklegen* gezogen. Von den Kugeln sind $M = 7$ rot, die restlichen $N - M = 3$ Kugeln schwarz. Von Interesse sei die Anzahl X der insgesamt gezogenen roten Kugeln.

- Welche Ausprägungen kann X annehmen?
- Wie groß ist die Wahrscheinlichkeit dafür, dass insgesamt *genau* 4 rote Kugeln gezogen werden?
- Mit welcher Wahrscheinlichkeit sind *höchstens* 4 Kugeln rot?

Geben Sie in den Aufgabenteilen b – c die Wahrscheinlichkeiten auf vier Dezimalstellen genau an.

Aufgabe 12.4 (Hypergeometrische Verteilung)

In Österreich und der Schweiz wird das Lottospiel „6 aus 45“ gespielt, nicht „6 aus 49“ wie in Deutschland.

- Berechnen Sie den Erwartungswert für die Anzahl X der Richtigen.
- Wie groß ist hier die Wahrscheinlichkeit des Ereignisses „6 Richtige“?

Aufgabe 12.5 (Hypergeometrische Verteilung)

Aus einer Gruppe von 6 Personen, die aus 2 Männern und 4 Frauen besteht, werden im Rahmen eines Gewinnspiels zwei Gewinner ermittelt. Dazu wird jeder Person eine der Zahlen 1, 2, ..., 6 zugeordnet, die jeweilige Zahl auf einem Zettel notiert und die Zettel in identischen Briefumschlägen abgelegt. Nach Durchmischen der Umschläge werden nacheinander und ohne Zurücklegen zwei Umschläge zufällig ausgewählt. Die in den gezogenen Umschlägen enthaltenen Zahlen definieren dann die Gewinner. Wie groß ist die Wahrscheinlichkeit dafür, dass das Gewinnerpaar aus einer Frau und einem Mann besteht?

**Aufgabe 13.1** (Stetige Rechteckverteilung)

Ein Berufstätiger geht jeden Werktag zu einer Bushaltestelle, von der die Buslinie 112 zu seiner Firma fährt. Die Linie verkehrt alle 20 Minuten.

Der Fahrgäst schlendert in der Regel nach dem Frühstück ohne auf die Uhr zu schauen zur Bushaltestelle und nimmt den nächsten Bus der Linie 112. Die Wartezeit X lässt sich anhand der stetigen Gleichverteilung modellieren. Geben Sie die Dichtefunktion der Verteilung an. Berechnen Sie auch den Erwartungswert $E(X)$ und interpretieren Sie das Ergebnis.

Aufgabe 13.2 (Normalverteilung und Standardnormalverteilung)

- Eine Zufallsvariable X sei *normalverteilt* mit Erwartungswert $\mu = 3$ und Standardabweichung $\sigma = 4$. Berechnen Sie die Wahrscheinlichkeit $P(3 \leq X \leq 7)$ dafür, dass X im Intervall $[3; 7]$ liegt.
- Bestimmen Sie für eine *standardnormalverteilte* Zufallsvariable Z die fünf Wahrscheinlichkeiten $P(Z \leq 2,9)$, $P(0 \leq Z \leq 2,3)$, $P(-1,3 \leq Z \leq 0)$, $P(-0,8 \leq Z \leq 0,8)$ und $P(-1,3 \leq Z \leq 1,2)$.

Aufgabe 13.3 (Normalverteilung und Standardnormalverteilung)

In den Krankenhäusern einer Region wurde eine Erhebung zum Geburtsgewicht von Neugeborenen durchgeführt. Dabei blieben Frühgeborene unberücksichtigt. Die Untersuchung ergab, dass sich das in Gramm angegebene Geburtsgewicht X in guter Näherung durch eine Normalverteilung mit Erwartungswert $\mu = 2\,950$ und Standardabweichung $\sigma = 120$ modellieren lässt.

- Wie groß ist die Wahrscheinlichkeit, dass ein Neugeborenes nicht mehr als 2 800 Gramm wog?
- Wie groß ist die Wahrscheinlichkeit für ein Gewicht zwischen 2 800 und 3 250 Gramm?
- Was beinhaltet das 0,1-Quantil der Normalverteilung mit $\mu = 2\,950$ und Varianz $\sigma^2 = 120^2$ und welchen Wert hat es hier?

Anmerkung zu Teil a: Die gesuchte Wahrscheinlichkeit $P(X \leq 2\,800)$ stimmt mit $P(X < 2\,800)$ überein, wie man aus (13.8) mit $x_0 = 2\,800$ ersieht. Es ist also für das Ergebnis unerheblich, ob man bei der Aufgabenformulierung „nicht mehr als 2 800 Gramm“ oder „weniger als 2 800 Gramm“ verwendet.

Aufgabe 13.4 (Quantile von t - und Standardnormalverteilung)

Bei einem Test werde eine Teststatistik T eingesetzt, die bei Gültigkeit der Nullhypothese einer t -Verteilung mit $n = 10$ Freiheitsgraden folgt.

- Geben Sie einen Wert an, den eine Ausprägung der Testgröße T mit Wahrscheinlichkeit $\alpha = 0,05$ nicht überschreitet.
- Geben Sie ein bezüglich des Nullpunkts symmetrisches Intervall an, in dem eine Ausprägung von T mit Wahrscheinlichkeit $1 - \alpha = 0,95$ liegt. Wie groß ist die Wahrscheinlichkeit, mit der eine standardnormalverteilte Zufallsvariable in dieses Intervall fällt?

Aufgabe 14.1 (Kovarianz zweier Zufallsvariablen)

Es werden zwei „faire“ Münzen nacheinander geworfen, wobei das Ergebnis des ersten Wurfs durch eine Zufallsvariable X und das des zweiten Wurfs durch Y beschrieben sei. Die beiden möglichen Ausprägungen „Kopf“ und „Zahl“ von X und Y seien mit „1“ (Kopf) resp. mit „0“ (Zahl) codiert.



Kapitel 14

- a) Wie groß sind die Wahrscheinlichkeiten

$$p_{11} = P(X = 1; Y = 1), p_{12} = P(X = 1; Y = 0), \\ p_{21} = P(X = 0; Y = 1), p_{22} = P(X = 0; Y = 0),$$

durch die die gemeinsame Wahrscheinlichkeitsverteilung beider Zufallsvariablen bestimmt ist?

- b) Berechnen Sie die Kovarianz von X und Y .

Hinweis zu Aufgabenteil b: Wenn man (14.12) heranzieht, kann man den dort auftretenden Term $E(XY)$ analog zu (12.6) als Summe der vier möglichen Ausprägungen von XY ermitteln, wobei jeder Summand jeweils mit seiner Eintrittswahrscheinlichkeit p_{11} , p_{12} , p_{21} resp. p_{22} gewichtet wird.

**Aufgabe 15.1** (Punktschätzung von Kenngrößen)

Bei 24 Patienten wurde im Rahmen einer Studie u. a. das Gewicht X ermittelt. Es ergaben sich folgende Werte, jeweils auf volle kg gerundet (angelehnt an TOUTENBURG / SCHOMAKER / WISSMANN (2009, Abschnitt 10.4)):

Kapitel 15

45, 73, 70, 60, 62, 66, 85, 52, 49, 67, 70, 82, 91, 77, 76, 62, 55, 52, 59, 49, 62, 66, 94, 79.

- a) Berechnen Sie unter der Annahme, dass das Körpergewicht normalverteilt ist, eine unverzerrte Schätzung $\hat{\mu}$ für den Erwartungswert μ .
 b) Ermitteln Sie auch für die Varianz σ^2 und die Standardabweichung σ der Normalverteilung eine unverzerrte Schätzung. Hier genügt bei Fehlen von Software die Angabe der Bestimmungsformel, also des Lösungsansatzes.

Aufgabe 15.2 (Konfidenzintervalle für Erwartungswerte)

Bestimmen Sie mit den Daten aus Aufgabe 15.1 und der Normalverteilungsannahme für das Gewicht X auch ein Konfidenzintervall zum Niveau 0,95 für den unbekannten Parameter μ . Geben Sie die Grenzen des Intervalls auf eine Stelle nach dem Dezimalkomma genau an und interpretieren Sie Ihr Ergebnis.

Aufgabe 15.3 (Konfidenzintervalle für Anteilswerte)

Berechnen Sie auf der Basis der Daten des Politbarometers vom 8. Dezember 2017 (Tabelle 9.3) ein Konfidenzintervall zum Konfidenzniveau 0,95 für den Anteil p der Frauen in Deutschland mit SPD-Präferenz. Verwenden Sie dabei die Näherungsformel (15.20).

**Aufgabe 16.1** (einseitiger Gauß-Test)

Die nachstehende Aufgabe ist adaptiert aus CAPUTO / FAHRMEIR / KÜNSTLER / LANG / PIGEOT / TUTZ (2009, Kapitel 10):

Kapitel 16

Bei einer Studie zum Thema „Schwangerschaft“ mit 49 beteiligten Müttern wurde das Alter X der Frauen bei der Geburt des ersten Kindes festgestellt. Die Forschungshypothese beinhaltete, dass das Durchschnittsalter von Frauen bei der Erstgeburt oberhalb von 25 Jahren liegt. Bei den 49 befragten Frauen ergab sich der Mittelwert $\bar{x} = 26$ (Altersangaben in vollen Jahren).

- Testen Sie zum Signifikanzniveau $\alpha = 0,05$ die Hypothese $H_0 : \mu \leq 25$ gegen die Alternativhypothese $H_1 : \mu > 25$. Gehen Sie davon aus, dass X einer Normalverteilung mit Varianz $\sigma^2 = 9$ folgt.
- Was beinhalten bei diesem konkreten Test die Fehler 1. und 2. Art?

Aufgabe 16.2 (einseitiger Gauß-Test)

- Berechnen Sie für den linksseitigen Test (16.3) mit $\mu_0 = 2$ kg und $\alpha = 0,05$ aus Beispiel 16.3 die Wahrscheinlichkeit einer Verwerfung der Nullhypothese für den Fall, dass μ den Wert $\mu = 2,002$ kg hat.
- Wie groß ist diese Wahrscheinlichkeit für $\mu = 1,997$?
- Skizzieren Sie den Verlauf der Gütfunktion $G(\mu)$ des Tests.

Aufgabe 16.3 (zweiseitiger Gauß-Test)

Betrachten Sie wie in Beispiel 16.3 die industrielle Abfüllung von Zucker, der in 2-kg-Tüten in den Verkauf kommt (Sollwert $\mu_0 = 2$ kg). Das tatsächliche Füllgewicht X sei normalverteilt mit Standardabweichung $\sigma = 0,01$ kg. Verbraucher sind an einer Kontrolle von Sollwertunterschreitungen, Hersteller an einer Überwachung von Sollwertüberschreitungen interessiert.

- Anhand einer Stichprobe von 10 Tüten wurde für das Füllgewicht der Mittelwert $\bar{x} = 2,007$ kg ermittelt. Über einen zweiseitigen Test (16.1) mit $\mu_0 = 2$ kg soll geprüft werden, ob der Stichprobenbefund für oder gegen die Beibehaltung von H_0 spricht. Führen Sie den Test mit $\alpha = 0,05$ durch und interpretieren Sie das Ergebnis.
- Führen Sie den Test auch mit $\alpha = 0,01$ durch.



Aufgabe 17.1 (Kleinste-Quadrat-Schätzung)

Im Herzlabor eines Krankenhauses wird bei jedem Patienten eine Anamnese durchgeführt, bei der u. a. das Körpergewicht, die Körpergröße und der systolische Blutdruck festgestellt werden. Die Variablen „Körpergewicht“ und „Körpergröße“ können anhand des Body-Mass-Indexes zusammengeführt werden, dessen Wert eine erste Orientierung über das Vorliegen von Über- oder Untergewichtigkeit ermöglicht. Für 6 Männer wurden für den Body-Mass-Index X und den systolischen Blutdruck Y folgende Werte $(x_i; y_i)$ gemessen:

i	1	2	3	4	5	6
x_i	26	23	27	28	24	25
y_i	170	150	160	175	155	150

Gehen Sie davon aus, dass die Werte x_i und y_i über eine lineare Regression (17.1) verknüpft sind und schätzen Sie anhand des tabellierten Datensatzes des Umfangs $n = 6$ die Regressionskoeffizienten β und α unter Verwendung der KQ-Methode. Weisen Sie Ihre Schätzergebnisse $\hat{\beta}$ und $\hat{\alpha}$ auf zwei Stellen nach dem Dezimalkomma genau aus.

Aufgabe 17.2 (Kleinst-Quadrat-Schätzung und Bestimmtheitsmaß)

Das folgende Beispiel ist adaptiert aus CAPUTO / FAHRMEIR / KÜNSTLER / LANG / PIGEOT / TUTZ (2009, Kapitel 3):

In einer Region wurde anhand einer Studie untersucht, inwieweit das Geburtsgewicht Y Neugeborener (in Kilogramm) von verschiedenen sozioökonomischen Variablen abhängt, u. a. vom monatlichen Nettoeinkommen X der Eltern (in Tausend Euro). In der nachstehenden Tabelle sind für acht an der Studie beteiligte Kinder die Beobachtungsdaten ($x_i; y_i$) wiedergegeben ($i = 1, 2, \dots, 8$):

i	1	2	3	4	5	6	7	8
x_i	1,9	2,7	3,1	4,0	3,9	3,4	2,9	2,1
y_i	3,0	2,5	4,5	3,5	4,0	3,0	4,0	3,5

- Berechnen Sie unter Annahme des einfachen linearen Regressionsmodells (17.1) die KQ-Schätzungen für die Regressionskoeffizienten β und α .
- Quantifizieren Sie anhand des Bestimmtheitsmaßes R^2 aus (17.17) die Anpassungsgüte der Regressionsgeraden. Interpretieren Sie das Ergebnis.

Aufgabe 17.3 (Kleinst-Quadrat-Schätzung)

In Beispiel 17.1, das sich auf das einfache Regressionsmodell bezog ($k = 1$), wurden die KQ-Schätzformeln (17.6) und (17.7) auf einen sehr kleinen Datensatz angewendet. Leiten Sie die dabei errechneten Schätzwerte $\hat{\beta} = 0,125$ und $\hat{\alpha} = 0,25$ erneut her, nun aber unter Verwendung der KQ-Schätzformel (17.35) für das multiple Regressionsmodell. Notieren Sie die Formel (17.35) zunächst für den Spezialfall $k = 1$.

20.3 Lösungen zu Teil I

Lösung zu Aufgabe 1.1 (Statistical Literacy)

Richtig sind die Aussagen d und e.

Erläuterung zu d: Aus der Schlagzeile und Information 1 lässt sich berechnen, dass der aktuelle Frauenanteil $\frac{0,007}{0,0975} + 0,007$ beträgt, also aufgerundet 8%. Aus der Schlagzeile und Information 2 ergibt sich durch Addition von 7,3% und 0,7 Prozentpunkten ebenfalls ein aktueller Frauenanteil von 8%.

Kapitel 1

Erläuterung zu e: Da nur relative, aber keine absoluten Häufigkeiten gegeben sind, lässt sich die aktuelle Zahl von Frauen in Vorständen aus den gegebenen Informationen nicht berechnen.



Lösung zu Aufgabe 1.2 (Statistical Literacy)

- Die Aussage ist falsch. Tatsächlich stieg den publizierten Daten zufolge der Anteil von 50 Frauen unter 683 Vorständen auf 56 Frauen unter 697 Vorständen, also um 0,7 Prozentpunkte oder 9,75 Prozent.
- Bemerkenswert an der Schlagzeile ist, dass sie erst durch Kontextwissen über den derzeitigen Anteil von Frauen in Führungspositionen zur Nachricht wird: Bei fünf Prozent weiblichen Vorständen wäre eine Steigerung um 0,7 Prozent (!) in Zusammenhang mit der andauernden Diskussion über die Quote ein „Versagen“ wahlweise der Unternehmen oder der Frauen, bei 50 Prozent nicht der Rede wert, bei 95 Prozent zunehmend „bedrohlich“ für die Männer. Die Schlagzeile beinhaltet implizit die Botschaft, dass es sich um eine schwache Veränderung handelt. Dabei wird die vordergründige Aussage der Schlagzeile („Die Frauenquote steigt“) durch den Zusatz in Klammern massiv relativiert.
- In diesem Fall würde die Schlagzeile die implizite Botschaft kommunizieren, dass es sich um eine starke Veränderung handelt, obwohl faktisch eine Steigerung um knapp zehn Prozent bei einer kleinen Ausgangsbasis von 7,3 Prozent immer noch in einem einstelligen Frauenanteil resultiert.
- Eine Schlagzeile, die eine Befürwortung der Frauenquote signalisiert, könnte beispielsweise lauten: „Dank Frauenquote: Der Frauenanteil in den Vorständen steigt um fast zehn Prozent.“
- Eine Schlagzeile, die eine Ablehnung der Frauenquote signalisiert, könnte beispielsweise lauten: „Trotz Frauenquote: Der Frauenanteil in den Vorständen steigt um lediglich 0,7 Prozentpunkte.“



Lösung zu Aufgabe 2.1 (Grundbegriffe)

Kapitel 2 Die Grundgesamtheit ist durch alle in Deutschland lebenden Schulkinder definiert, die Schulkinder sind die statistischen Einheiten (Merkmalsträger). Interessierende Merkmale sind hier vor allem die Dauer des täglichen Fernsehkonsums (z. B. mit den Ausprägungen „Minuten“ oder „Viertelstunden“) und der Fernsehsender (evtl. nur mit der Differenzierung zwischen „privater Sender“ und „öffentlich-rechtlicher Sender“).

Als Teilgesamtheiten bieten sich Teilmengen an, zwischen denen man Unterschiede bezüglich des Fernsehverhaltens vermutet und entsprechende Hypothesen empirisch absichern will. Man könnte etwa zwischen Schulkindern in verschiedenen Schultypen oder Altersgruppen unterscheiden. Denkbar wäre auch eine Unterscheidung hinsichtlich der Zugehörigkeit der Kinder zu Sportvereinen oder des Bildungsstands der Eltern.

Lösung zu Aufgabe 2.2 (Skalenarten)

Höchster erreichter Schulabschluss: *ordinalskaliert*.

Gewählte Partei bei einer Kommunalwahl: *nominalskaliert*.

Bonität von Kunden einer Sparkasse: *ordinalskaliert*.

Verfallsdatum bei einer Konfitürensorte: *metrisch skaliert*.

Lösung zu Aufgabe 3.1 (Erhebungsarten)

Beispiele für vielbeachtete Zeitreihen: Zeitreihen aus dem Finanzmarktsektor (DAX und andere Aktienkursindizes, Entwicklung der Hypothekenzinssätze), Zeitreihen für den Arbeitsmarkt (z. B. monatliche Erwerbslosenquoten), Konjunkturindikatoren (Verbraucherpreisindex, Inflationsrate, Veränderungen beim Bruttoinlandsprodukt).



Kapitel 3

Lösung zu Aufgabe 3.2 (geschichtete Zufallsauswahl)

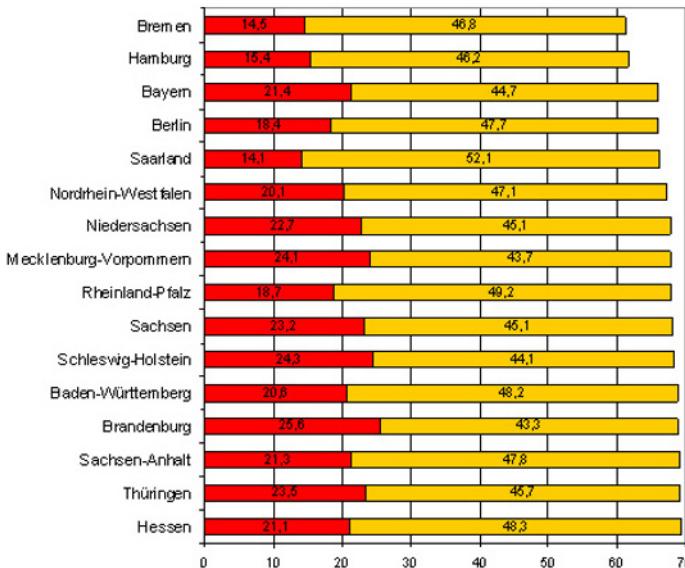
Bei proportionaler Schichtung entfallen $\frac{270}{600} \cdot 120 = 54$ Studierende auf Schicht 1, $\frac{180}{600} \cdot 120 = 36$ auf Schicht 2 und $\frac{150}{600} \cdot 120 = 30$ auf Schicht 3.



Kapitel 4

Lösung zu Aufgabe 4.1 (Nationale Verzehrstudie II)

Die folgende Abbildung zeigt die relativen Häufigkeiten für Männer (in %) in Form eines gestapelten Balkendiagramms. Die numerischen Werte der beiden dargestellten Teilhäufigkeiten sind jeweils eingeblendet. Die Bundesländer sind nach zunehmender Größe der Summe $f(a_2) + f(a_3)$ geordnet, also nach zunehmender Balkenlänge.



Man sieht, dass der Prozentsatz der Männer, die als übergewichtig oder gar als fettleibig zu klassifizieren waren, in allen Bundesländern oberhalb von 60% lag, meistens sogar in der Nähe von 70%. Die besten Werte mit unter 62% wurden in den beiden Stadtstaaten Bremen und Hamburg registriert. Für die geplante Verzehrstudie III, deren Ergebnisse bis 2023 vorliegen sollen, ist mit noch ungünstigeren Ergebnissen zu rechnen.

Lösung zu Aufgabe 4.2 (Gruppierung von Daten, Histogramme)

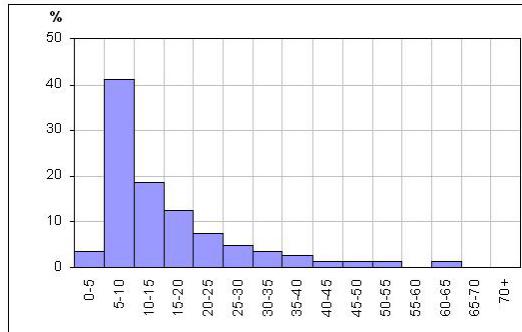
a) Merkmalsträger: Arbeitnehmer

Merkmal: Bruttoverdienst / Stunde (in EUR).

b)

Nr. der Klasse	Klassen-grenzen	Klassenbesetzungshäufigkeit absolut	Klassenbesetzungshäufigkeit relativ (in %)
1	0 bis unter 5,0	3	3,75
2	5,0 bis unter 10	33	41,25
3	10,0 bis unter 15,0	15	18,75
4	15,0 bis unter 20,0	10	12,5
5	20,0 bis unter 25,0	6	7,5
6	25,0 bis unter 30,0	4	5,0
7	30,0 bis unter 35,0	3	3,75
8	35,0 bis unter 40,0	2	2,5
9	40,0 bis unter 45,0	1	1,25
10	45,0 bis unter 50,0	1	1,25
11	50,0 bis unter 55,0	1	1,25
12	55,0 bis unter 60,0	0	0
13	60,0 bis unter 65,0	1	1,25
14	65,0 bis unter 70,0	0	0
15	70,0 und mehr	0	0

c)

**Lösung zu Aufgabe 4.3** (empirische Verteilungsfunktion)

a) Realisierbar sind 16 Ausprägungen, nämlich 3, 4, ..., 18.

b) Die empirische Verteilungsfunktion kann höchstens 16 Sprünge aufweisen.

Lösung zu Aufgabe 5.1

- a) Häufigkeitsverteilung für das Merkmal „Augenzahl“:

Augenzahlen	•	••	•••	••••	•••••	••••••
Absolute Häufigkeit	1	2	3	2	1	3
Relative Häufigkeit	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$



Kapitel 5

Interaktives Objekt
„Lageparameter“

- b) Wenn man die Augenzahlen nach Größe sortiert, erhält man eine Liste mit den Werten 1, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 6. Der Median ist nach (5.1) wegen $n = 12$ der Mittelwert aus dem 6. und 7. Element $x_{(6)}$ resp. $x_{(7)}$ der geordneten Liste, d. h. es ist $\tilde{x} = \frac{1}{2} \cdot (3 + 4) = 3,5$. Nach (5.2) erhält man dann $\bar{x} = \frac{1}{12} \cdot 45 = 3,75$. Wenn man alternativ von (5.4) ausgeht, ergibt sich dieser Wert wie folgt:

$$\bar{x} = \left(1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{4} \right) = \frac{45}{12} = 3,75.$$

- c) Für die Spannweite folgt nach (5.5) der Wert $R = 6 - 1 = 5$. Für die Berechnung der Varianz kann man jede der Formeln (5.6), (5.7) oder (5.10) heranziehen. Bei Verwendung von (5.10) ergibt sich

$$s^2 = \frac{(-2,75)^2}{12} + \frac{(-1,75)^2}{6} + \frac{(-0,75)^2}{4} + \frac{(0,25)^2}{6} + \frac{(1,25)^2}{12} + \frac{(2,25)^2}{4}.$$

Man errechnet hieraus $s^2 \approx 2,688$ und mit (5.8) dann $s \approx 1,640$.

Lösung zu Aufgabe 5.2 (Quantile und Boxplots)

- a) Die Quartile bestimmen sich nach (5.11). Da der geordnete Datensatz durch 1, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 6 gegeben ist, erhält man mit $p = 0,25$ aufgrund der Ganzzahligkeit von $n \cdot p$

$$x_{0,25} = \frac{1}{2} \cdot (x_{(3)} + x_{(4)}) = \frac{1}{2} \cdot (2 + 3) = 2,5.$$

Analog folgt für denselben Datensatz mit $p = 0,75$, wieder bei Beachtung der Ganzzahligkeit von $n \cdot p$

$$x_{0,75} = \frac{1}{2} \cdot (x_{(9)} + x_{(10)}) = \frac{1}{2} \cdot (5 + 6) = 5,5.$$

- b) Die 5 Charakteristika eines Boxplots sind in Abbildung 5.3 wiedergegeben. Es sind dies hier die beiden Extremwerte $x_{(1)} = 1$ und $x_{(12)} = 6$, die beiden Quartile $x_{0,25} = 2,5$ und $x_{0,75} = 5,5$ sowie der Median $\tilde{x} = 3,5$. Der Interquartilsabstand (5.12) beträgt $Q = x_{0,75} - x_{0,25} = 3$.
- c) Wenn man den um $x_{(13)} = 3$ erweiterten Datensatz nach aufsteigender Größe ordnet, hat man 1, 2, 2, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6. Die Quartile $x_{0,25}$ und $x_{0,75}$ bestimmen sich nach (5.11). Mit $n = 13$ und $p = 0,25$

oder $p = 0,75$ ist $n \cdot p$ nicht mehr ganzzahlig. Es ist daher die obere Hälfte von (5.11) anzuwenden. Man erhält

$$x_{0,25} = x_{([3,25]+1)} = x_{(4)} = 3; \quad x_{0,75} = x_{([9,75]+1)} = x_{(10)} = 5.$$

Für den Interquartilsabstand Q gilt $Q = x_{0,75} - x_{0,25} = 2$.

Lösung zu Aufgabe 5.3 (Quantile und Boxplots)

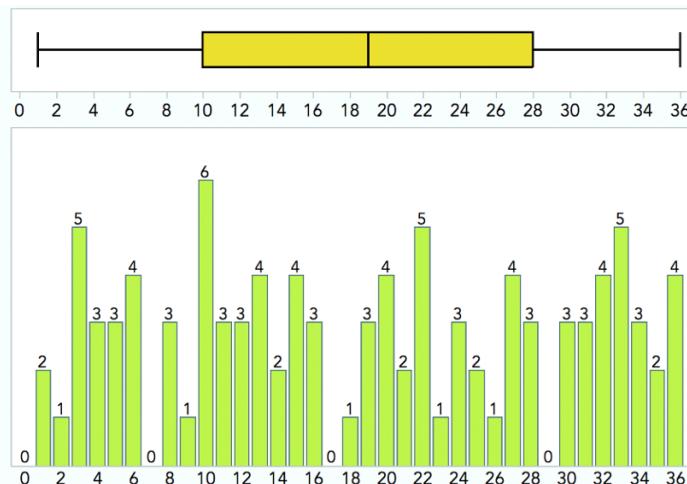
- a) Die 100 Werte der Rouletteserie lassen sich aus dem Balkendiagramm ablesen. Der Wert 0 tritt z. B. gar nicht auf, der Ausgang 1 insgesamt zweimal, der Ausgang 2 einmal etc. Man erhält den – hier nur unvollständig wiedergegebenen – Datensatz

$$1, 1, 2, 3, 3, 3, 3, 3, 4, 4, 4, \dots, 35, 35, 36, 36, 36, 36.$$

Für den Median folgt $\tilde{x} = x_{0,5} = 19$, für das untere und obere Quartil $x_{0,25} = 10$ bzw. $x_{0,75} = 28$ und für den Interquartilsabstand $Q = 18$.

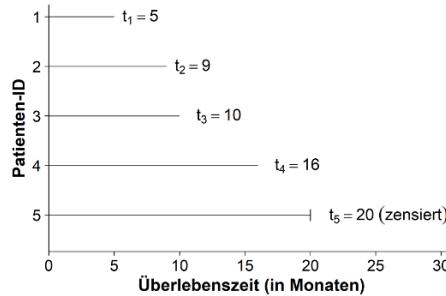
- b) Da die 0 gar nicht, der Ausgang 1 zweimal und der Ausgang 2 einmal auftrat, errechnet sich $F(2)$ als Summe der zugehörigen relativen Häufigkeiten, d. h. als $F(2) = 0 + 0,02 + 0,01 = 0,03$. Die relative Häufigkeit für den Ausgang 36 betrug 0,04. Hieraus leitet man für $F(35)$ den Wert $F(35) = 1 - 0,04 = 0,96$ ab.
- c) Der Boxplot weist die Extremwerte des Datensatzes, die Box und den innerhalb der Box liegenden Median aus. Der kleinste realisierte Ausgang ist hier 1, der größte 36. Die Box ist durch $x_{0,25} = 10$ und $x_{0,75} = 28$ begrenzt und der Median $\tilde{x} = 19$ liegt genau in der Mitte der Box.

Die folgende Grafik zeigt erneut das in der Aufgabe wiedergegebene Säulendiagramm, nun mit der dynamischen Statistiksoftware JMP erzeugt und mit zusätzlich oberhalb des Diagramms eingezeichnetem Boxplot.



Lösung zu Aufgabe 6.1 (Kaplan-Meier-Verfahren)

- a) Die folgende, zu Abbildung 6.1 analoge Grafik zeigt die Überlebenszeitverläufe für die 5 Patienten:



Kapitel 6

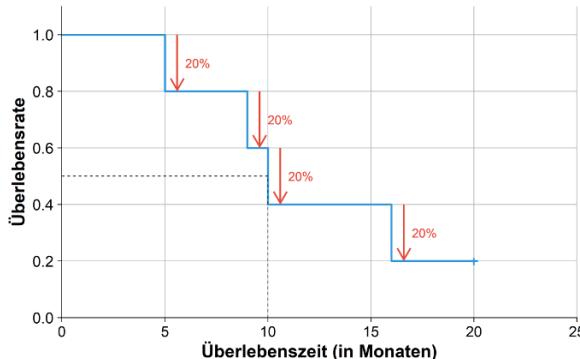
- b) Für jeden der 5 Ereigniszeitpunkte $t_1 = 5, t_2 = 9, t_3 = 10, t_4 = 16$ und $t_5 = 20$ kann man die Anzahl an Ereignissen und die Anzahl der bis zu dem jeweiligen Zeitpunkt unter Risiko stehenden Patienten abzählen:

$$\begin{aligned} d_1 &= 1, \quad d_2 = 1, \quad d_3 = 1, \quad d_4 = 1, \quad d_5 = 0, \\ n_1 &= 5, \quad n_2 = 4, \quad n_3 = 3, \quad n_4 = 2, \quad n_5 = 1. \end{aligned}$$

Damit lässt sich die Kaplan-Meier-Kurve $\hat{S}(t)$ nach (6.1) berechnen:

$$\hat{S}(t) = \begin{cases} 1 & \text{für } t \in [0; t_1], \\ 1 - \frac{d_1}{n_1} = \frac{4}{5} = 0,8 & \text{für } t \in [t_1; t_2], \\ (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2}) = \frac{3}{5} = 0,6 & \text{für } t \in [t_2; t_3], \\ (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})(1 - \frac{d_3}{n_3}) = \frac{2}{5} = 0,4 & \text{für } t \in [t_3; t_4], \\ (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})(1 - \frac{d_3}{n_3})(1 - \frac{d_4}{n_4}) = \frac{1}{5} = 0,2 & \text{für } t \geq t_4. \end{cases}$$

Die nachstehende Abbildung zeigt die Kurve:



- c) Der Median der Überlebenszeit beträgt 10 Monate. Man kann diesen Wert direkt aus obiger Kaplan-Meier-Kurve ablesen.

- d) Nach jedem Ereigniszeitpunkt stehen jeweils 20% weniger Patientinnen unter Risiko.

Lösung zu Aufgabe 6.2 (Kaplan-Meier-Verfahren)

- a) Da zu den ersten drei Zeitpunkten jeweils genau ein Ereignis eintritt

$$d_1 = d_2 = d_3 = 1 \Rightarrow n_2 = n_1 - 1, \quad n_3 = n_2 - 1 = n_1 - 2,$$

nimmt der Kaplan-Meier-Schätzer $\hat{S}(t)$ an den drei Ereigniszeitpunkten jeweils um $\frac{1}{n_1}$ ab. Es gilt

$$\begin{aligned}\hat{S}(t_1) &= \left(1 - \frac{1}{n_1}\right) = \frac{n_1 - 1}{n_1}; \\ \hat{S}(t_2) &= \left(1 - \frac{1}{n_1}\right) \cdot \left(1 - \frac{1}{n_2}\right) = \frac{n_1 - 1}{n_1} \cdot \frac{n_1 - 2}{n_1 - 1} = \frac{n_1 - 2}{n_1}; \\ \hat{S}(t_3) &= \left(1 - \frac{1}{n_1}\right) \cdot \left(1 - \frac{1}{n_2}\right) \cdot \left(1 - \frac{1}{n_3}\right) = \frac{n_1 - 3}{n_1}.\end{aligned}$$

- b) Die Berechnung von $\hat{S}(t)$ ist etwas komplizierter, wenn zensierte Daten vorkommen:

$$\begin{aligned}\hat{S}(t_5) &= \left(1 - \frac{1}{n_1}\right) \cdot \left(1 - \frac{1}{n_2}\right) \cdot \left(1 - \frac{1}{n_3}\right) \cdot \left(1 - \frac{0}{n_4}\right) \cdot \left(1 - \frac{1}{n_5}\right) \\ &= \frac{n_1 - 3}{n_1} \cdot \frac{n_5 - 1}{n_5} = \frac{n_1 - 3}{n_1} \cdot \frac{n_1 - 5}{n_1 - 4}.\end{aligned}$$

Die Brüche lassen sich im Allgemeinen nicht kürzen.



Lösung zu Aufgabe 7.1 (Gini-Koeffizient)

Kapitel 7

- a) Man erhält für die Ordinatenwerte v_1, v_2 und v_3 mit $p_4 = 200$:

$$v_1 = \frac{p_1}{p_4} = 0,1; \quad v_2 = \frac{p_2}{p_4} = 0,35; \quad v_3 = \frac{p_3}{p_4} = 0,65.$$

Da die Umsätze nach Größe geordnet vorliegen ($x_i = x_{(i)}$), folgt

$$q_4 = 1 \cdot 20 + 2 \cdot 50 + 3 \cdot 60 + 4 \cdot 70 = 580.$$



Mit der Merkmalssumme $p_4 = 200$ und (7.5) resultiert

$$G = \frac{1}{4} \cdot \left(\frac{2 \cdot 580}{200} - 1 \right) - 1 = 0,2.$$



Nach (7.8) folgt für den normierten Gini-Koeffizienten

$$G^* = \frac{4}{3} \cdot G = \frac{4}{15} \approx 0,267.$$

Interaktives Objekt
„Gini-Koeffizient“

- b) Der Inhalt A der markierten Fläche ist durch $A = \frac{G}{2} = 0,1$ gegeben.

Lösung zu Aufgabe 7.2 (Herfindahl-Index)

- a) Für den Herfindahl-Index H erhält man mit $p_4 = 200$:

$$H = \frac{1}{p_4^2} \cdot \sum_{i=1}^4 x_i^2 = \frac{1}{200^2} \cdot (20^2 + 50^2 + 60^2 + 70^2) = 0,285.$$

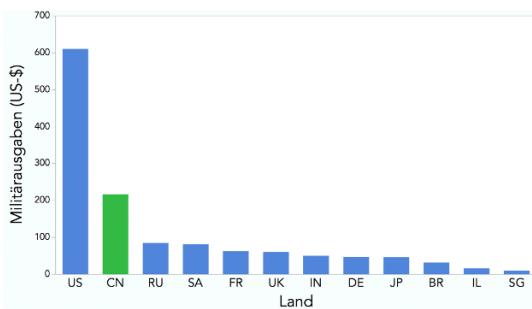
- b) Der Index H kann im Fall $n = 4$ nicht kleiner als 0,25 sein.

Lösung zu Aufgabe 8.1 (Militärausgaben 2014 im Ländervergleich)

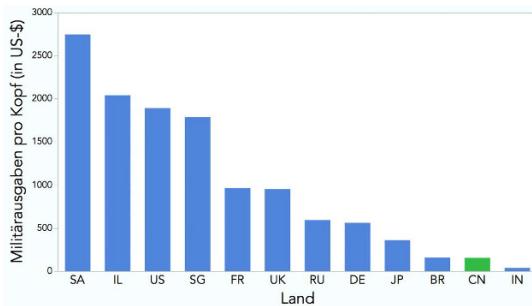
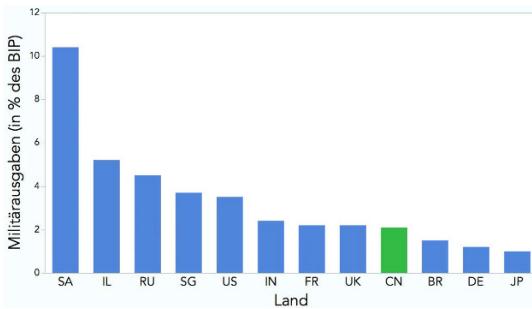
Die nachstehenden, mit JMP erzeugten Grafiken zeigen die nach absteigenden Werten geordneten Säulendiagramme. Die Betonung eines Landes (hier: China) macht gut sichtbar, dass sich die drei Ranglisten deutlich unterscheiden.



Kapitel 8



Interaktives Objekt
„Militärausgaben“



Lösung zu Aufgabe 8.2 (zusammengesetzte Indexzahlen)

- a) Gewichtet man die in Tabelle 8.2 aufgeführten Anzahlen für Gold, Silber und Bronze nach dem Schema 5 – 3 – 2 , resultiert folgende Rangfolge:

Rang	Nation	Gold	Silber	Bronze	Punkte
1.	🇨🇳 China	51	21	28	374
2.	🇺🇸 USA	36	38	36	366
3.	🇷🇺 Russland	23	21	28	234
4.	🇬🇧 Großbritannien	19	13	15	164
5.	🇦🇺 Australien	14	15	17	149
6.	🇩🇪 Deutschland	16	10	15	140
7.	🇫🇷 Frankreich	7	16	17	117
8.	🇰🇷 Südkorea	13	10	8	111
9.	🇮🇹 Italien	8	10	10	90
10.	🇯🇵 Japan	9	6	10	83



Interaktives Objekt
„Sommerolympiade
2016“

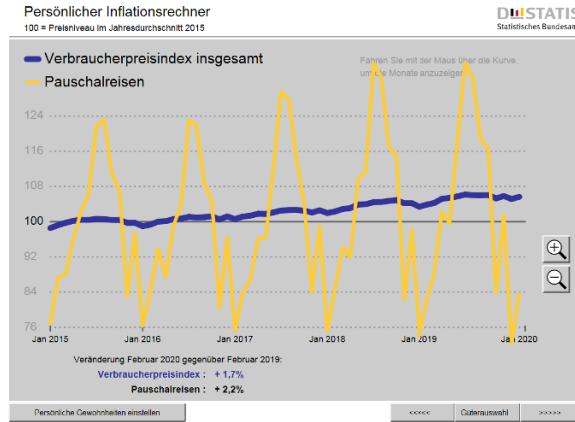
- b) Dividiert man die Punktzahlen der obigen Tabelle noch durch die in Klammern angegebene Einwohnerzahl (in Millionen) des jeweiligen Landes, resultiert eine ganz andere Rangfolge, bei der Länder mit hoher Einwohnerzahl erwartungsgemäß an Boden verlieren:

Rang	Nation	Gold	Silber	Bronze	Punkte
1.	🇦🇺 Australien (21,0)	14	15	17	7,10
2.	🇬🇧 Großbritannien (60,9)	19	13	15	2,69
3.	🇰🇷 Südkorea (48,4)	13	10	8	2,29
4.	🇫🇷 Frankreich (64,1)	7	16	17	1,83
5.	🇩🇪 Deutschland (82,4)	16	10	15	1,69
6.	🇷🇺 Russland (140,7)	23	21	28	1,66
7.	🇮🇹 Italien (58,1)	8	10	10	1,55
8.	🇺🇸 USA (303,8)	36	38	36	1,20
9.	🇯🇵 Japan (127,3)	9	6	10	0,65
10.	🇨🇳 China (1330,0)	51	21	28	0,28

Lösung zu Aufgabe 8.3 (Preisindex)

Man erkennt deutlich, dass sich die Preise für Pauschalreisen stets im Juli / August und, etwas weniger ausgeprägt, im Dezember nach oben verändern, offenbar aufgrund der höheren Nachfrage nach Reisen in den Sommerferien und um die Weihnachtszeit. Da die Sommerferien in Deutschland von Bundesland

unterschiedlich terminiert sind, verteilt sich die höhere Nachfrage hier auf einen längeren Zeitraum.



Lösung zu Aufgabe 9.1 (Randverteilungen)

Die Randverteilungen erhält man mit Aufsummieren der Zeilen resp. Spalten:

Kapitel 9

	b_1	b_2	Zeilensummen
a_1	62	96	158
a_2	14	188	202
Spaltensummen	76	284	n

Lösung zu Aufgabe 9.2 (Bedingte Häufigkeitsverteilungen)

Der Wert $f_X(a_5|b_1) = \frac{88}{824} \approx 0,107$ sagt aus, dass von den Personen in der Stichprobe, die männlichen Geschlechts ($Y = b_1$) waren, 10,7% die FDP favorisierten ($X = a_5$). Das Ergebnis $f_Y(b_1|a_2) = \frac{219}{353} \approx 0,620$ beinhaltet, dass von den Personen, die sich für die SPD ($X = a_2$) entschieden hatten, 62,0% Männer waren ($Y = b_1$).



Lösung zu Aufgabe 10.1 (Zusammenhangsmessung; Nominalskala)

- a) Man erhält mit den Werten der in Aufgabe 9.1 wiedergegebenen Vierfeldertafel bei Anwendung von (10.7) und Beachtung von $n = 360$

$$\chi^2 = \frac{360 \cdot (62 \cdot 188 - 96 \cdot 14)^2}{158 \cdot 202 \cdot 76 \cdot 284} = \frac{360 \cdot 10312^2}{158 \cdot 202 \cdot 76 \cdot 284} \approx 55,571.$$

- b) Für den Φ -Koeffizienten folgt nach (10.3)

$$\Phi = \sqrt{\frac{55,571}{360}} \approx 0,393.$$

Kapitel 10

Das Cramérsche Zusammenhangsmaß V aus (10.5) ist bei einer Vierfeldertafel wegen $M - 1 = 1$ mit dem Φ -Koeffizienten identisch.

Lösung zu Aufgabe 10.2 (Zusammenhangsmessung; metr. Skala)

Wenn man eine Arbeitstabelle anlegt, erhält man folgende Werte:

i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	-3,01	9,06	4,36	19,01	-13,12
2	-4,71	22,18	-1,74	3,03	8,20
3	1,29	1,66	-0,04	0,00	-0,05
4	1,09	1,19	4,46	19,89	4,86
5	4,79	22,94	1,76	3,10	8,43
6	-2,51	6,30	-3,64	13,25	9,14
7	2,29	5,24	2,26	5,11	5,18
8	-3,51	12,32	-4,44	19,71	15,58
9	3,49	12,18	-3,54	12,53	-12,35
10	0,79	0,62	0,56	0,31	0,44
Summe:		93,69	95,94	26,31	

Einsetzen der Summen am Tabellenende in (10.11) liefert

$$r = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{10} (y_i - \bar{y})^2}} = \frac{26,31}{\sqrt{93,69} \cdot \sqrt{95,94}} \approx 0,278.$$

Dieser Wert beinhaltet schwache Korrelation.

Lösung zu Aufgabe 10.3 (Zusammenhangsmessung; Ordinalskala)

Der Rangkorrelationskoeffizient kann nach (10.16) bestimmt werden, weil kein Rangplatz doppelt besetzt ist. Für die Anwendung von (10.16) sind die Rangplatzdifferenzen d_i und deren Quadrate zu ermitteln:

Mannschaft i	Rang (Halbturnier)	Rang (Freiluftturnier)	Rangdifferenz d_i	Quadrierte Rangdiff. d_i^2
A	1	2	-1	1
B	2	3	-1	1
C	3	1	2	4
D	4	5	-1	1
E	5	4	1	1

Hieraus folgt dann für das Zusammenhangsmaß r_{SP} :

$$r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^5 d_i^2}{5 \cdot (5^2 - 1)} = 1 - \frac{6 \cdot 8}{120} = 0,6.$$

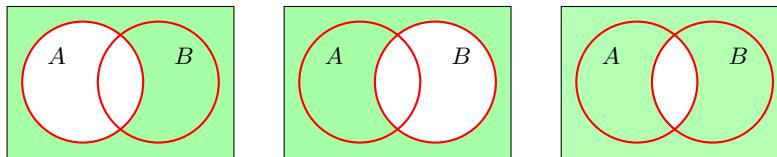
20.4 Lösungen zu Teil II



Lösung zu Aufgabe 11.1 (Venn-Diagramme)

Nur Aussage B ist unzutreffend. Dass C und D zutreffend sind, aber B nicht richtig ist, erkennt man leichter, wenn man zunächst \bar{A} und \bar{B} einzeln visualisiert. Die entsprechenden Venn-Diagramme sind nachstehend an erster und zweiter Stelle wiedergegeben. Das dritte Venn-Diagramm zeigt die Vereinigungsmenge von \bar{A} und \bar{B} . Die dort dunkel markierte Fläche stimmt nicht mit der Fläche überein, die im zweiten Venn-Diagramm der Aufgabe dunkel markiert war.

Kapitel 11



Lösung zu Aufgabe 11.2 (Ereignisse und Ereignisraum)

- a) Beim dreifachen Münzwurf ist die Ergebnismenge Ω durch die folgenden acht Tripel (Elementarereignisse) definiert:

$$\Omega = \{(Z, Z, Z), (Z, Z, K), (Z, K, Z), (K, Z, Z), \\ (Z, K, K), (K, Z, K), (K, K, Z), (K, K, K)\}$$

- b) Das Ereignis $A = \{\text{Bei mindestens zwei Würfen tritt „K“ auf}\}$ setzt sich zusammen aus den letzten vier der acht Tripel der Ergebnismenge Ω :

$$A = \{(Z, K, K), (K, Z, K), (K, K, Z), (K, K, K)\}$$

Lösung zu Aufgabe 11.3 (Laplace-Wahrscheinlichkeiten)

- a) Die Wahrscheinlichkeit dafür, bei *einmaligem* Würfeln mit einem fairen Würfel *keine* Sechs zu erhalten, beträgt $\frac{5}{6}$. Bei *dreimaligem* Würfeln errechnet sich die Wahrscheinlichkeit für das Auftreten von 0 Sechsen aufgrund der Unabhängigkeit der drei Ausgänge für jeden Wurf als

$$P = \left(\frac{5}{6}\right)^3 = \frac{125}{216} \approx 0,5787.$$



- b) Das Ereignis A umfasst 4 der 8 Elementarereignisse des dreifachen Münzwurfs. Jedes Tripel ist gleichwahrscheinlich. Nach (11.5) gilt also $P(A) = \frac{4}{8} = 0,5$ und damit auch $P(\bar{A}) = 0,5$.

Interaktives Objekt
„Binomialverteilung“

Hinweis zu Teil b: Das Ergebnis 0,5 lässt sich auch anhand der Binomialverteilung ableiten. Die Anzahl X der Ausgänge mit „Kopf“ beim dreifachen Wurf einer fairen Münze ist nämlich binomialverteilt mit $n = 3$ und $p = 0,5$. Gesucht ist die Wahrscheinlichkeit $P(X \geq 2) = 1 - P(X \leq 1)$. Die Wahrscheinlichkeit $P(X \leq 1)$ erhält man aus Tabelle 19.1 als Wert $F(1) = 0,5$ der Verteilungsfunktion $F(x)$ der $B(3;0,5)$ -Verteilung an der Stelle $x = 1$, d. h. es ist $P(X \geq 2) = 1 - 0,5 = 0,5$.

Lösung zu Aufgabe 11.4 (Kombinatorik)

Da Buchstaben mehrfach auftreten können und es hier auf die Reihenfolge der Buchstaben ankommt, liegt der Fall „Ziehen mit Zurücklegen und mit Berücksichtigung der Reihenfolge“ der Tabelle (11.1) vor. Da die Folge von A bis J insgesamt 10 Buchstaben umfasst, werden $n = 5$ Elemente aus einer Grundgesamtheit von $N = 10$ Elementen gezogen. Die Anzahl der Möglichkeiten beträgt insgesamt $10^5 = 100\,000$.

Lösung zu Aufgabe 11.5 (Bedingte Wahrscheinlichkeiten)

- a) Die Wahrscheinlichkeit $P(\bar{A})$ dafür, dass die zufällig ausgewählte Person keine Vollzeitbeschäftigung hat, ist nach (11.5) gegeben durch

$$P(\bar{A}) = \frac{160 - 64}{160} = \frac{96}{160} = 0,6.$$

- b) Die Wahrscheinlichkeit $P(A \cap B)$ dafür, dass sie sowohl vollzeitbeschäftigt als auch promoviert ist, errechnet sich zu

$$P(A \cap B) = \frac{40}{160} = 0,25.$$

- c) Für die Wahrscheinlichkeit $P(B|A)$, dass eine aus dem vollzeitbeschäftigen Lehrpersonal zufällig ausgewählte Person promoviert ist, ergibt sich nach (11.13)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0,25}{1 - 0,6} = \frac{0,25}{0,4} = 0,625.$$

Dasselbe Ergebnis ließ sich auch mit (11.11) ableiten. Man erhält

$$P(B|A) = \frac{40}{64} = 0,625.$$

- d) Wenn $P(A \cap B) = P(A) \cdot P(B)$ gilt, sind die Ereignisse A und B gemäß (11.16) unabhängig. Hier ist $P(A) = 0,4$, $P(B) = \frac{60}{160} = 0,375$ und folglich $P(A) \cdot P(B) = 0,4 \cdot 0,375 = 0,15$. Dieser Wert stimmt nicht mit $P(A \cap B) = 0,25$ überein, d. h. die Ereignisse A und B sind abhängig.
- e) Man erhält folgende Vierfeldertafel (Vorgaben dieser Aufgabe kursiv):

	m. Promotion (B)	o. Promotion (\bar{B})	Zeilensummen
Vollzeit (A)	40	24	64
Teilzeit (\bar{A})	20	76	96
Spaltensummen	60	100	160

Hieraus folgt

$$P(\bar{A}) = \frac{96}{160} = 0,6; \quad P(A \cap B) = \frac{40}{160} = 0,25; \quad P(B|A) = \frac{40}{64} = 0,625.$$

Lösung zu Aufgabe 11.6 (Bedingte Wahrscheinlichkeiten)

- a) Tabelle 11.3 ist eine Vierfeldertafel für absolute Häufigkeiten. In dieser bezeichnen N_1 und N_2 die Anzahl der Frauen in der Teilpopulation mit Krebskrankung resp. ohne Krebskrankung:

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächl. krank (A)	$r_p = 0,90 \cdot N_1$	$f_n = 0,10 \cdot N_1$	$N_1 = 0,008 \cdot N$
Tatsächl. gesund (\bar{A})	$f_p = 0,07 \cdot N_2$	$r_n = 0,93 \cdot N_2$	$N_2 = 0,992 \cdot N$
Spaltensummen	$r_p + f_p$	$f_n + r_n$	N

Nach Einsetzen von N_1 und N_2 im Innern der Tabelle folgt, wenn man zu relativen Häufigkeiten übergeht:

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächl. krank (A)	0,00720	0,0008	0,0080
Tatsächl. gesund (\bar{A})	0,06944	0,92256	0,992
Spaltensummen	0,07664	0,92336	1

- b) Sensitivität und Spezifität errechnen sich formal wie folgt:

$$\text{Sensitivität} = P(B|A) = \frac{r_p}{r_p + f_n} = \frac{0,0072}{0,008} = 0,90$$

$$\text{Spezifität} = P(\bar{B}|\bar{A}) = \frac{r_n}{f_p + r_n} = \frac{0,92256}{0,992} = 0,93.$$

Es wird somit bei 90 % der erkrankten Frauen die Krebskrankung entdeckt (richtig-positiver Befund), während bei 93% der gesunden Frauen ein richtig-negativer Befund resultiert.

- c) Für die Wahrscheinlichkeiten $P(B)$ und $P(\bar{B}) = 1 - P(B)$ eines positiven bzw. eines negativen Befunds bei einer zufällig aus der Grundgesamtheit ausgewählten Frau gilt, wie man aus der letzten Vierfeldertafel in Aufgabenteil a direkt ablesen kann,

$$P(B) = 0,07664; \quad P(\bar{B}) = 0,92336.$$

- d) Für die Wahrscheinlichkeit $P(\bar{A}|B)$ eines falsch-positiven Befunds ergibt sich aus der Vierfeldertafel

$$P(\bar{A}|B) = \frac{0,06944}{0,07664} \approx 0,906.$$

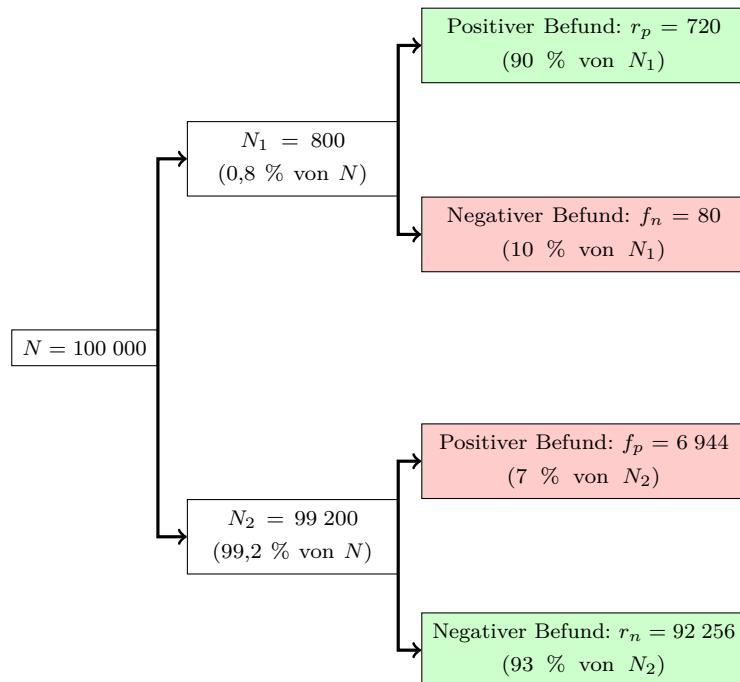
Dieses Ergebnis kann man auch aus (11.15) ableiten. Demnach ist bei ca. 90,6% (!) aller positiven Befunde der Befund falsch.

- e) Mit $P(A) = 0,006$ erhält man für $P(B)$ den Wert 0,07498 und für $P(\bar{A}|B)$

$$P(\bar{A}|B) = \frac{0,06958}{0,07498} \approx 0,928.$$

Lösung zu Aufgabe 11.7 (Baumdiagramm und Vierfeldertafel)

- a) Im nachstehenden Baumdiagramm stehen die Anzahlen f_n und f_p (Fehlerentscheidungen) in Rechtecken mit rötlichem Hintergrund.



- b) Vierfeldertafel für absolute Häufigkeiten bei Wahl von $N = 100\,000$:

	Test positiv (B)	Test negativ (\bar{B})	Zeilensummen
Tatsächl. krank (A)	$r_p = 720$	$f_n = 80$	$N_1 = 800$
Tatsächl. gesund (\bar{A})	$f_p = 6\,944$	$r_n = 92\,256$	$N_2 = 99\,200$
Spaltensummen	$r_p + f_p = 7\,664$	$f_n + r_n = 92\,336$	$N = 100\,000$

**Lösung zu Aufgabe 12.1** (Würfeln mit zwei Würfeln)

Kapitel 12

- a) Für die Eintrittswahrscheinlichkeiten $f(x)$ gilt nach (12.1)

$$f(x) = \begin{cases} \frac{1}{36} \approx 0,0277 & \text{für } x = 2 \text{ und für } x = 12; \\ \frac{1}{18} \approx 0,0556 & \text{für } x = 3 \text{ und für } x = 11; \\ \frac{1}{12} \approx 0,0833 & \text{für } x = 4 \text{ und für } x = 10; \\ \frac{1}{9} \approx 0,1111 & \text{für } x = 5 \text{ und für } x = 9; \\ \frac{5}{36} \approx 0,1388 & \text{für } x = 6 \text{ und für } x = 8; \\ \frac{1}{6} \approx 0,1667 & \text{für } x = 7; \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Die Funktion $f(x)$ ist symmetrisch bezüglich $x = 7$.

- b) Für die gemäß (12.3) definierte Verteilungsfunktion $F(x)$ gilt z. B. $F(0,5) = 0$, $F(3) = f(2) + f(3) = \frac{1}{12} \approx 0,0833$, $F(3,5) = F(3)$ und $F(6) = F(3) + f(4) + f(5) + f(6) = \frac{5}{12} \approx 0,41667$.
- c) Da die Augenzahlen bei den beiden Würfeln unabhängig voneinander sind und der Erwartungswert der Augenzahl eines Würfels jeweils den Wert 3,5 hat, besitzt der Erwartungswert der Augensumme X nach (12.13) den Wert 7.



Interaktives Objekt
„Augensummen“
(mit Modell)

Lösung zu Aufgabe 12.2 (Binomialverteilung)

Das Drehen des Glücksrades entspricht einem Bernoulli-Experiment (mögliche Ausgänge: eine bestimmte Farbe tritt auf / tritt nicht auf). Die Anzahl X des Auftretens einer bestimmten Farbe ist binomialverteilt mit $p = 0,25$ und $n = 4$, weil es vier Farben gibt (jede mit Eintrittswahrscheinlichkeit $p = 0,25$) und die Bernoulli-Kette vier Experimente umfasst. Daraus folgt:

- a) Die Wahrscheinlichkeit $P(X \leq 1)$, höchstens einmal die Farbe „Gelb“ zu erhalten, errechnet sich als Wert $F(1)$ der Verteilungsfunktion einer $B(4; 0,25)$ -verteilten Zufallsvariablen. Da Tabelle 19.1 keine Werte mit $p = 0,25$ ausweist, wird das nebenstehende Lernobjekt verwendet. Es resultiert $F(1) = 0,7383$. Die gesuchte Wahrscheinlichkeit $P(X \geq 2)$ dafür, dass mindestens zweimal die Farbe „Gelb“ erscheint, ist die Komplementärwahrscheinlichkeit von $P(X \leq 1)$, d. h. es gilt

$$P(X \geq 2) = 1 - P(X \leq 1) = 0,2617.$$



Interaktives Objekt
„Rechnen mit der
Binomialverteilung“

Lösung zu Aufgabe 12.3 (Hypergeometrische Verteilung)

- a) Die hypergeometrisch verteilte Zufallsvariable X kann alle ganzzahligen Werte im Intervall $T = \{x_{\min}, \dots, x_{\max}\}$ annehmen. Dabei ist $x_{\min} = \max[0; n - (N - M)]$ und $x_{\max} = \min(n; M)$. Mit $N = 10$, $M = 7$ und $n = 5$ folgt $x_{\min} = \max(0; 2) = 2$ und $x_{\max} = \min(5; 7) = 5$. Die Anzahl X der gezogenen roten Kugeln kann demnach nur 2, 3, 4 oder 5 sein.

Dass die Ausprägungen 0 und 1 hier ausgeschlossen sind, leuchtet sofort ein. Wenn man aus einer Urne, in der 7 rote und 3 schwarze Kugeln sind, nacheinander 5 Kugeln ohne Zurücklegen zieht, müssen mindestens zwei der gezogenen Kugeln rot sein. Es ist auch einsichtig, dass die Anzahl der gezogenen roten Kugeln den Wert $n = 5$ nicht übersteigen kann.

- b) Die Wahrscheinlichkeit $P(X = 4)$ dafür, dass genau 4 rote Kugeln gezogen werden, ist nach (12.26) gegeben durch

$$f(4) = \frac{\binom{7}{4} \binom{3}{1}}{\binom{10}{5}} = \frac{35 \cdot 3}{252} \approx 0,4167.$$

- c) Die Wahrscheinlichkeit $P(X \leq 4)$ dafür, dass *höchstens* 4 rote Kugeln gezogen werden, errechnet sich nach (12.27) als Summe aller von Null verschiedenen Werte der Wahrscheinlichkeitsfunktion $f(x)$ bis zur Stelle $x = 4$, d. h. als $F(4) = f(2) + f(3) + f(4)$. Die Werte $f(2)$ und $f(3)$ bestimmt man erneut mit (12.26):

$$f(2) = \frac{\binom{7}{2} \binom{3}{3}}{\binom{10}{5}} = \frac{2520 \cdot 1}{252} \approx 0,0833.$$

$$f(3) = \frac{\binom{4}{3} \binom{6}{2}}{\binom{10}{5}} = \frac{4 \cdot 15}{252} \approx 0,4167.$$



Interaktives Objekt
„Hypergeometrische
Verteilung“

Damit folgt schließlich $F(4) \approx 0,0833 + 0,4167 + 0,4167 = 0,9167$. Diesen Wert kann man auch unter Verwendung des nebenstehenden Lernobjekts erhalten, wenn man dort $N = 10$, $M = 7$, $n = 5$ und $x = 4$ einstellt.

Gleiches gilt für den Wert $f(4)$. Letzterer ergibt sich als Differenz $f(4) = F(4) - F(3)$ zweier benachbarter Werte der Verteilungsfunktion $F(x)$. Für $F(3)$ erhält man anhand des Lernobjekts bei Wahl von $N = 10$, $M = 7$, $n = 5$ und $x = 3$ den Wert $F(3) = 0,5000$. Für $f(4)$ resultiert somit der schon in Aufgabenteil a errechnete Wert $f(4) \approx 0,9167 - 0,5000 = 0,4167$.

Lösung zu Aufgabe 12.4 (Hypergeometrische Verteilung)

Die Anzahl X der Richtigen beim Spiel „6 aus 45“ ist $H(n; M; N)$ -verteilt mit $n = 6$, $M = 6$ und $N = 45$.

- Für $\mu = E(X)$ folgt nach (12.24), dass $\mu = 6 \cdot \frac{6}{45} = 0,8$.
- Die Anzahl der möglichen Ausgänge beim Spiel „6 aus 45“ ist nach Tabelle 11.1 – siehe dort den Fall „Ziehen ohne Zurücklegen und ohne Berücksichtigung der Anordnung“ – gegeben durch

$$\binom{45}{6} = \frac{45!}{39! \cdot 6!} = \frac{45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 8\,145\,060.$$

Da von den 8 145 060 möglichen Ausgängen, die alle gleichwahrscheinlich sind, nur ein einziger „6 Richtige“ beinhaltet, gilt nach (11.5) für die Wahrscheinlichkeit $f(6) = P(X = 6)$

$$f(6) = \frac{1}{8\,145\,060} \approx 0,12277 \cdot 10^{-6}.$$

Die Wahrscheinlichkeit beträgt also ca. $12,28 \cdot 10^{-6}$ %. Zum Vergleich: Beim deutschen Lottospiel „6 aus 49“ beträgt die Wahrscheinlichkeit für „6 Richtige“ nur $0,0715 \cdot 10^{-6}$, also $7,15 \cdot 10^{-6}$ % (s. Beispiel 12.5).

Lösung zu Aufgabe 12.5 (Hypergeometrische Verteilung)

Die gesuchte Wahrscheinlichkeit lässt sich unter Verwendung der hypergeometrischen Verteilung mit Parametern $N = 6$, $M = 2$ und $n = 2$ bestimmen. Der Parameter M entspricht hier der Anzahl der Männer in der Grundgesamtheit, aus der eine Stichprobe gezogen wird. Man erhält für das Ereignis „eine

Frau und ein Mann bilden das „Gewinnerpaar“ nach (12.26) bei Einsetzen der genannten Parameter und mit $\binom{2}{1} = 2$ sowie $\binom{4}{1} = 4$

$$f(1) = \frac{\binom{2}{1} \cdot \binom{6-2}{2-1}}{\binom{6}{2}} = \frac{2 \cdot 4}{\binom{6}{2}} = \frac{8}{15} \approx 0,533.$$

Diese Wahrscheinlichkeit lässt sich alternativ auch allein anhand kombinatorischer Überlegungen errechnen. Seien 1, 2, 3 und 4 die Zahlen, die den vier Frauen zugeordnet werden, und 5 resp. 6 die Codierungen für die beiden Männer. Es gibt 15 Möglichkeiten zwei verschiedene Zahlen auszuwählen, nämlich

$$(1; 2), (1; 3); (1; 4), (1; \mathbf{5}), (1; \mathbf{6}), (2; 3), (2; 4), (2; \mathbf{5}), \\ (2; \mathbf{6}), (3; 4), (3; \mathbf{5}), (3; \mathbf{6}), (4; 5), (4; \mathbf{6}), (5; 6).$$

Die Gesamtzahl der Möglichkeiten, aus der Gruppe von 6 Personen 2 Personen auszuwählen (Ziehen ohne Zurücklegen und *ohne* Berücksichtigung einer Reihenfolge), lässt sich auch nach Tabelle 11.1 ermitteln:

$$\binom{6}{2} = \frac{6!}{4! \cdot 2!} = 15.$$

Unter diesen 15 Wertepaaren sind 8 Paare, bei denen genau eine der beiden Zahlen 5 und 6 vorkommt (s. Markierungen durch fette Schrift). Man errechnet mit (11.5) für die gesuchte Wahrscheinlichkeit den Wert $\frac{8}{15} \approx 0,533$.



Lösung zu Aufgabe 13.1 (Rechteckverteilung)

Der Fahrgäst trifft mit Sicherheit innerhalb eines 20-Minuten-Intervalls ein, das durch die Abfahrtszeiten zweier aufeinanderfolgender Busse der Linie 112 begrenzt ist. Die Wartezeit X bis zum Eintreffen des nächsten Busses lässt sich anhand einer stetigen Gleichverteilung über $[0; 20]$ modellieren. Deren Dichtefunktion ist nach (13.6) gegeben durch

$$f(x) = \begin{cases} \frac{1}{20} & \text{für } 0 \leq x \leq 20 \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Mit (13.12) errechnet man dann $E(X) = 10$ (mittlere Wartezeit bei zufälligem Eintreffen an der Bushaltestelle).

Lösung zu Aufgabe 13.2 (Normalverteilung)

- a) Für die $N(3; 4^2)$ -verteilte Zufallsvariable X gilt mit (13.23)

$$P(3 \leq X \leq 7) = \Phi(1) - \Phi(0) = 0,8413 - 0,5 = 0,3413.$$

- b) Mit (13.20) – (13.23) und Tabelle 19.2 folgt:

$$\begin{aligned}P(Z \leq 2,9) &= \Phi(2,9) = 0,9981 \\P(0 \leq Z \leq 2,3) &= \Phi(2,3) - \Phi(0) = 0,9893 - 0,5 = 0,4893 \\P(-1,3 \leq Z \leq 0) &= \Phi(0) - [1 - \Phi(1,3)] = 0,4032 \\P(-0,8 \leq Z \leq 0,8) &= \Phi(0,8) - [1 - \Phi(0,8)] = 0,5762 \\P(-1,3 \leq Z \leq 1,2) &= \Phi(1,2) - [1 - \Phi(1,3)] = 0,7881.\end{aligned}$$

Anmerkung: Mit R lassen sich die obigen Wahrscheinlichkeiten besonders leicht bestimmen. Man erhält z. B. für die Wahrscheinlichkeit $P(3 \leq X \leq 7)$ aus Aufgabenteil a und die Wahrscheinlichkeit $P(-1,3 \leq Z \leq 1,2)$ aus Teil b

```
> pnorm(7, 3, 4) - pnorm(3, 3, 4)
[1] 0.3413447
> pnorm(1.2) - pnorm(-1.3)
[1] 0.7881298
```

Mit „pnorm(.)“ ist die Verteilungsfunktion der Normalverteilung bezeichnet.

Lösung zu Aufgabe 13.3 (Normalverteilung)

- a) Nach (13.21) gilt für die Verteilungsfunktion $F(x)$ der $N(2\,950; 120^2)$ -verteilten Zufallsvariablen X

$$F(x) = P(X \leq 2\,800) = \Phi\left(\frac{2\,800 - 2\,950}{120}\right) = \Phi(-1,25).$$

Mit (13.20) und Tabelle 19.2 folgt:



Interaktives Objekt
„Standardnormalverteilung“



$$\Phi(-1,25) = 1 - \Phi(1,25) = 1 - 0,8944 = 0,1056.$$

Die Wahrscheinlichkeit dafür, dass ein Neugeborenes ein Geburtsgewicht von höchstens 2 800 Gramm aufwies, betrug demnach 10,56%.

- b) Mit (13.23) verifiziert man, dass

$$\begin{aligned}P(2\,800 \leq X \leq 3\,250) &= \Phi\left(\frac{3\,250 - 2\,950}{120}\right) - \Phi\left(\frac{2\,800 - 2\,950}{120}\right) \\&= \Phi(2,5) - \Phi(-1,25).\end{aligned}$$

Erneuter Rückgriff auf (13.20) und Tabelle 19.2 ergibt

$$\Phi(2,5) - \Phi(-1,25) = \Phi(2,5) - 1 + \Phi(1,25) = 0,8882.$$

Die Wahrscheinlichkeit dafür, dass ein Neugeborenes zwischen 2 800 Gramm und 3 250 wog, betrug 88,82%.

- c) Das 0,1-Quantil $x_{0,1}$ der Normalverteilung ist mit dem 0,1-Quantil $z_{0,1}$ der Standardnormalverteilung über (13.26) verknüpft. Man errechnet mit $z_{0,1} = -z_{0,9} = -1,2816$ aus Tabelle 19.3 den Wert

$$x_{0,1} = 2\,950 + z_{0,1} \cdot 120 = 2\,950 - 1,2816 \cdot 120 \approx 2\,796,2.$$



Objekt „Quantile der Standardnormalverteilung“

Das 0,1-Quantil der Normalverteilung ist der Wert $x = x_{0,1}$, an dem die Verteilungsfunktion $F(x) = P(X \leq x)$ der Verteilung den Wert 0,1 annimmt. Wählt man also ein an der Untersuchung beteiligtes Neugeborenes zufällig aus, so hatte dieses mit einer Wahrscheinlichkeit von 10% ein Gewicht von höchstens 2 796,2 Gramm.

Lösung zu Aufgabe 13.4 (Quantile)

- Der Wert, den eine Ausprägung der als Testgröße fungierenden t_{10} -verteilten Zufallvariablen T mit Wahrscheinlichkeit $\alpha = 0,05$ nicht überschreitet, ist das 0,05-Quantil dieser Verteilung. Mit (13.29) und Tabelle 19.5 erhält man $t_{10;0,05} = -t_{10;0,95} = -1,812$.
- Das Intervall, in das eine Ausprägung von T mit Wahrscheinlichkeit $1 - \alpha = 0,95$ fällt, ist durch $[t_{10;0,025}; t_{10;0,975}]$, also durch $[-2,228; 2,228]$ gegeben.

Eine standardnormalverteilte Zufallvariable Z würde gemäß (13.23) und Tabelle 19.2 mit der Wahrscheinlichkeit

$$\Phi(2,228) - \Phi(-2,228) = \Phi(2,228) - [1 - \Phi(2,228)] \approx 0,974$$

in das durch die beiden Quantile der t -Verteilung definierte Intervall $[-2,228; 2,228]$ fallen.

Anmerkung zu Teil b: Während also die Realisation einer mit 10 Freiheitsgraden t -verteilten Zufallsvariablen mit einer Wahrscheinlichkeit von 0,05 (5%) außerhalb des Intervalls $[-2,228; 2,228]$ liegt, beträgt diese Wahrscheinlichkeit bei einer standardnormalverteilten Zufallsvariablen nur etwa $1 - 0,974 = 0,026$, d. h. 2,6%, weil die Dichte der Standardnormalverteilung im Vergleich zu der der t -Verteilung mit 10 Freiheitsgraden etwas steiler verläuft (vgl. auch Abbildung 13.8).

Lösung zu Aufgabe 14.1 (Kovarianz zweier Zufallsvariablen)

- Es gibt vier mögliche Ausgänge $(x; y)$, nämlich $(1; 1)$, $(1; 0)$, $(0; 1)$ und $(0; 0)$, die alle gleichwahrscheinlich sind. Die Wahrscheinlichkeiten p_{11} , p_{12} , p_{21} und p_{22} haben also alle den Wert 0,25.
- Die Kovarianz von X und Y kann bestimmt werden. Der Erwartungswert von X und Y ist jeweils 0,5 („faire“ Münzen). Der Erwartungswert $E(XY)$ errechnet sich analog zu (11.6) gemäß

$$E(XY) = p_{11} \cdot 1 \cdot 1 + p_{12} \cdot 1 \cdot 0 + p_{21} \cdot 0 \cdot 1 + p_{22} \cdot 0 \cdot 0 = 0,25.$$

Die Kovarianz hat somit den Wert $Cov(X, Y) = 0,25 - 0,5 \cdot 0,5 = 0$. Dieses Ergebnis hätte man aufgrund der Unabhängigkeit der Variablen X und Y auch direkt aus (14.13) erschließen können.



Interaktives Objekt
„Quantile der
 t -Verteilung“



Kapitel 14

Lösung zu Aufgabe 15.1 (Punktschätzung von Kenngrößen)



Kapitel 15

- a) Ein unverzerrter Punktschätzer $\hat{\mu}$ ist nach (15.6) durch die Ausprägung \bar{x} des in (14.3) eingeführten Stichprobenmittelwerts gegeben. Man errechnet bei Rundung auf drei Dezimalstellen $\bar{x} \approx 66,792$.
- b) Aus (15.9) ersieht man, dass die korrigierte Stichprobenvarianz s^{*2} aus (14.5) für die Varianz σ^2 der Normalverteilung eine unverzerrte Schätzung liefert. Die Summe (14.5) umfasst hier 24 Quadratterme und kann z. B. unter Verwendung von SPSS, Excel oder R ermittelt werden. Man erhält

$$s^{*2} := \frac{1}{23} \cdot \sum_{i=1}^{24} (x_i - 66,792)^2 \approx 180,346.$$

Für die Ausprägung der korrigierten Standardabweichung, die eine unverzerrte Schätzung für σ liefert, folgt dann $s^* \approx 13,429$.

Lösung zu Aufgabe 15.2 (Konfidenzintervalle für Erwartungswerte)

Das Intervall ergibt sich aus (15.16) mit $\alpha = 0,05$ und 23 Freiheitsgraden:



$$KI = \left[\bar{X} - t_{23;0,975} \cdot \frac{S^*}{\sqrt{24}}; \bar{X} + t_{23;0,975} \cdot \frac{S^*}{\sqrt{24}} \right].$$



Interaktives Objekt

„Quantile der
 t -Verteilung“

Setzt man für \bar{X} und S^* die aus den Daten errechneten Realisationen 66,792 resp. 13,429 und das Quantil $t_{23;0,975} = 2,069$ ein (s. Tabelle 19.5), so folgt bei Rundung auf eine Dezimalstelle:

$$KI = \left[66,792 - 2,069 \cdot \frac{13,429}{\sqrt{24}}; 66,792 + 2,069 \cdot \frac{13,429}{\sqrt{24}} \right] \approx [61,1; 72,5].$$

Ergebnisinterpretation: Die Grenzen des berechneten Konfidenzintervalls sind zufallsabhängig. Der unbekannte Parameter μ liegt nicht zwingend innerhalb des Intervalls. Das Verfahren der Intervallschätzung ist aber so angelegt, dass bei der Berechnung einer großen Anzahl von Konfidenzintervallen in $(1 - \alpha) \cdot 100\%$ der Fälle mit einer Überdeckung von μ zu rechnen ist.

Lösung zu Aufgabe 15.3 (Konfidenzintervalle für Anteilswerte)

Mit den Daten aus Tabelle 9.8 erhält man für den Anteil p der Frauen in Deutschland mit SPD-Präferenz zunächst die Punktschätzung $\hat{p} = \frac{134}{627} \approx 0,2137$ und damit gemäß (15.20) als approximatives 0,95-Konfidenzniveau

$$KI \approx \left[0,2137 - z_{0,975} \cdot \sqrt{\frac{0,2137 \cdot 0,7863}{627}}; 0,2137 + z_{0,975} \cdot \sqrt{\frac{0,2137 \cdot 0,7863}{627}} \right].$$

Mit $z_{0,975} = 1,96$ folgt schließlich

$$KI \approx [0,2137 - 0,0321; 0,2137 + 0,0321] \approx [0,182; 0,246].$$

Am genannten Stichtag hätte man für den fiktiven Fall einer unmittelbar bevorstehenden Bundestagswahl demnach den Anteil p der Frauen, die sich

für die SPD entscheiden, anhand dieses von etwa 18,2% bis 24,6% reichenden Intervalls geschätzt. Hätte man am gleichen Tag eine andere Stichprobe von 520 Frauen befragt, hätte sich eine andere Realisation für \hat{p} ergeben und damit auch andere Grenzen für die Intervallschätzung.

Lösung zu Aufgabe 16.1 (einseitiger Gauß-Test)

- a) Die Testvariable ist durch (16.2) gegeben, wobei dort $\mu_0 = 25$, $\sigma = 3$, $n = 49$ sowie $\bar{x} = 26$ einzusetzen ist. Man erhält

$$z = \frac{\bar{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \frac{26 - 25}{3} \cdot 7 \approx 2,333.$$

Die Ablehnung der Nullhypothese erfolgt, wenn $z > z_{0,95}$ gilt. Nach Tabelle 19.3 ist $z_{0,95} = 1,6449$ und H_0 folglich zu verwerfen. Dies impliziert, dass die Alternativhypothese H_1 als „statistisch gesichert“ gilt, d. h. als gesichert mit einer Irrtumswahrscheinlichkeit, deren Obergrenze bei dem hier durchgeführten einseitigen Test den Wert $\alpha = 0,05$ hat.

- b) Der Fehler 1. Art beinhaltet, dass man die Nullhypothese H_0 bei dem Test fälschlicherweise verwirft. Ein Fehler 1. Art kann im Falle $H_0 : \mu \leq 25$ offenbar nur für $\mu \leq 25$ auftreten.

Ein Fehler 2. Art liegt vor, wenn man die Nullhypothese H_0 bei dem Test fälschlicherweise *nicht* verwirft. Dies bedeutet bei dem in Rede stehenden einseitigen Test, dass man aufgrund der Realisation der Testgröße daran festhält, dass der Erwartungswert μ nicht über 25 Jahren liegt (Festhalten an H_0), obwohl er in Wirklichkeit oberhalb dieser Schranke liegt. Ein Fehler 2. Art kann hier nur im Falle $\mu > 25$ auftreten.

Die Wahrscheinlichkeit für das Eintreten eines Fehlers 2. Art hängt natürlich vom jeweiligen Wert μ ab; für $\mu = 28$ lässt sie sich z. B. gemäß (16.12) aus $\beta = P(\text{Nicht-Verwerfung von } H_0 | \mu = 28)$ errechnen.

Lösung zu Aufgabe 16.2 (einseitiger Gauß-Test)

- a) Im linksseitigen Gauß-Test aus Beispiel 16.3 war $\alpha = 0,05$, $n = 10$ und $\sigma = 0,01$. Setzt man neben den genannten Werten für α , n und σ noch $\mu = 2,002$ und $\mu_0 = 2,000$ in die Gütfunktion

$$G(\mu) = \Phi \left(-z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n} \right)$$

des Tests ein, so folgt für die Wahrscheinlichkeit $G(2,002)$ der Verwerfung der Nullhypothese für $\mu = 2,002$

$$G(2,002) = \Phi \left(-z_{0,95} - \frac{0,002}{0,01} \cdot \sqrt{10} \right) \approx \Phi(-2,277).$$

Wegen $\Phi(-2,277) = 1 - \Phi(2,277)$ folgt dann $G(2,002) \approx 0,0113$. Eine Ablehnung der Nullhypothese wäre im Falle $\mu = 2,002$ wegen $H_0 : \mu \geq 2$ eine Fehlentscheidung (Fehler 1. Art). Die Wahrscheinlichkeit hierfür beträgt also ca. 1,1%.

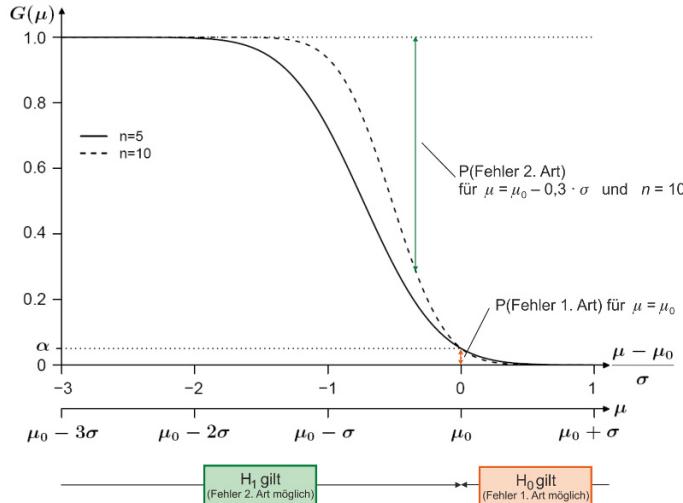


- b) Für $\mu = 1,997$ wäre eine Ablehnung der Nullhypothese hingegen eine korrekte Entscheidung. Sie tritt ein mit einer Wahrscheinlichkeit von

$$G(1,997) = \Phi\left(-z_{0,95} - \frac{-0,003}{0,01} \cdot \sqrt{10}\right) \approx \Phi(-0,696).$$

Mit $\Phi(-0,696) = 1 - \Phi(0,696)$ resultiert $G(1,997) \approx 0,242$. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art im Falle $\mu = 1,997$ und Wahl von $n = 10$ ist dann durch $1 - G(1,997) \approx 0,758$ gegeben. Dieser Wert ist in der folgenden Abbildung anhand eines vertikalen Pfeils veranschaulicht, der auf dem Niveau 1,0 endet.

- c) Der komplette Gütekontrollverlauf für den *rechtsseitigen* Gauß-Test war in Abbildung 16.6 für $n = 5$ und $n = 10$ und $\alpha = 0,05$ wiedergegeben. Für den *linksseitigen* Fall ergibt sie sich hieraus durch Spiegelung der Gütekurve des rechtsseitigen Tests an der vertikalen Geraden $\mu = \mu_0$. Die resultierende Grafik ist nachstehend wiedergegeben. Der Fall $n = 10$ ist durch die gestrichelte Kurve repräsentiert.



Setzt man in obiger Abbildung bei der unteren Abszissenachse $\mu_0 = 2$ und $\sigma = 0,01$ ein, so kann man die errechneten Wahrscheinlichkeiten $G(2,002) \approx 0,0113$ und $G(1,997) \approx 0,242$ auch als Werte der gestrichelten Kurve an den Stellen $\mu = 2,002$ resp. $\mu = 1,997$ zumindest grob ablesen.

Lösung zu Aufgabe 16.3 (zweiseitiger Gauß-Test)

- a) Die zu testenden Hypothesen sind durch (16.1) mit $\mu_0 = 2$ gegeben. Die Ablehnung der Nullhypothese $H_0 : \mu = \mu_0$ erfolgt nach (16.3) genau dann, wenn der Betrag $|z| = \left| \frac{\bar{x}-2}{0,01} \cdot \sqrt{10} \right|$ der Prüfstatistik aus (16.2) den aus Tabelle 19.3 ablesbaren Wert $z_{0,975} = 1,96$ überschreitet. Mit $\bar{x} = 2,007$ ergibt sich

$$|z| = \left| \frac{2,007 - 2}{0,01} \cdot \sqrt{10} \right| = 0,7 \cdot \sqrt{10} \approx 2,2136,$$

d.h., H_0 ist hier zu verwerfen. Die Alternativhypothese H_1 gilt dann als statistisch „bewiesen“ in dem Sinne, dass eine Irrtumswahrscheinlichkeit von $\alpha = 0,05$ vorbehalten bleibt.

- b) Bei Verwendung von $\alpha = 0,01$ ist $|z|$ mit dem Quantil $z_{0,995}$ der Standardnormalverteilung zu vergleichen. Dessen Wert 2,5758 wird von $|z| = 2,2136$ nun nicht mehr überschritten, d. h. man wird hier an der Nullhypothese H_0 festhalten und davon ausgehen, dass keine systematische Unter- oder Überschreitung des Soll-Füllgewichts vorliegt.

Lösung zu Aufgabe 17.1 (KQ-Schätzung)

Man kann eine Arbeitstabelle anlegen, wenn man die KQ-Schätzungen manuell berechnen will. Mit $\bar{x} = 25,5$ und $\bar{y} = 160$ erhält man:

Kapitel 17



i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0,5	0,25	10	5,0
2	-2,5	6,25	-10	25,0
3	1,5	2,25	0	0
4	2,5	6,25	15	37,5
5	-1,5	2,25	-5	7,5
6	-0,5	0,25	-10	5,0
Summe:		17,5		80

Für die KQ-Schätzung $\hat{\beta}$ von β folgt dann gemäß (17.6) zunächst

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{40}{3} \cdot \frac{12}{35} = \frac{32}{7} \approx 4,57.$$

Hieraus erhält man mit $\bar{x} = 25,5$ und $\bar{y} = 160$ für $\hat{\alpha}$ nach (17.7)

$$\hat{\alpha} = 160 - \hat{\beta} \cdot 25,5 = 160 - \frac{32}{7} \cdot 25,5 \approx 43,43.$$

Lösung zu Aufgabe 17.2 (KQ-Schätzung; Bestimmtheitsmaß)

- a) Mit $\bar{x} = 3,0$ und $\bar{y} = 3,5$ resultiert folgende Arbeitstabelle für die manuelle Berechnung der KQ-Schätzungen:

i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	-1,1	1,21	-0,5	0,55
2	-0,3	0,09	-1,0	0,30
3	0,1	0,01	1,0	0,10
4	1,0	1,00	0	0
5	0,9	0,81	0,5	0,45
6	0,4	0,16	-0,5	-0,20
7	-0,1	0,01	0,5	-0,05
8	-0,9	0,81	0	0
Summe:		4,1		1,15

Für die KQ-Schätzung von β folgt dann nach (17.6)

$$\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{0,14375}{0,5125} \approx 0,28$$

und für die KQ-Schätzung von α mit (17.7)

$$\hat{\alpha} = 3,5 - \hat{\beta} \cdot 3 \approx 3,5 - 0,84 = 2,66.$$

- b) Um das Bestimmtheitsmaß zu ermitteln, könnte man die Arbeitstabelle noch um eine Spalte $(y_i - \bar{y})^2$ erweitern. Die Spaltensumme wäre 3, d. h. es ist $s_y^2 = 0,375$. Nach (17.18) folgt

$$R^2 = \frac{(s_{xy})^2}{s_x^2 \cdot s_y^2} = \frac{0,14375^2}{0,5125 \cdot 0,375} \approx 0,108.$$

Der Wert bedeutet, dass der einfache lineare Regressionsansatz nur etwa 10,8% der Gesamtvariation der Daten erklärt (schwacher Erklärungsbeitrag). Es ist daher anzunehmen, dass noch andere Einflussgrößen bei der Modellspezifikation zu berücksichtigen sind.

Lösung zu Aufgabe 17.3 (KQ-Schätzung)

Die Matrizen \mathbf{X} und $\mathbf{X}'\mathbf{X}$ sowie der Vektor \mathbf{y} haben hier die Gestalt

$$\mathbf{X} = \begin{pmatrix} 1 & 10 \\ 1 & 30 \\ 1 & 50 \end{pmatrix} \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 3 & 90 \\ 90 & 3500 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix}$$

– vgl. auch (17.29) mit $n = 3$ und den Werten aus Tabelle 17.2. Mit $\alpha := \beta_0$ und $\beta := \beta_1$ folgt für die KQ-Schätzung des Vektors β der Regressionskoeffizienten

$$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} 3 & 90 \\ 90 & 3500 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 12 \\ 460 \end{pmatrix}.$$

Die Inversion der obigen (2×2) -Matrix kann man unter Heranziehung einer Software durchführen, etwa der freien Statistiksoftware *R*. Für die Regressionskoeffizienten α und β resultieren erneut die in Beispiel 17.1 schon ohne Verwendung von Matrizen errechneten Schätzwerte $\hat{\alpha} = 0,25$ und $\hat{\beta} = 0,125$:

$$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \frac{35}{24} & -\frac{3}{80} \\ -\frac{3}{80} & \frac{1}{800} \end{pmatrix} \cdot \begin{pmatrix} 12 \\ 460 \end{pmatrix} = \begin{pmatrix} \frac{35}{2} - \frac{69}{4} \\ -\frac{9}{20} + \frac{23}{40} \end{pmatrix} = \begin{pmatrix} 0,25 \\ 0,125 \end{pmatrix}.$$



21 Verzeichnisse und Internet-Ressourcen



Vorschau auf
das Kapitel

Im Schlusskapitel findet man neben einem Literatur-, Autoren- und Sachregister auch eine Zusammenstellung gängiger Statistiksoftware und interessanter Online-Ressourcen. Da eventuell nicht allen Lesern die in der Statistik häufig verwendeten griechischen Buchstaben und mathematischen Symbole geläufig sind, wurde auch ein Symbolverzeichnis aufgenommen.



Gesamtverzeichnis

21.1 Literaturverzeichnis

ASENDORPF, J. B. / F. J. NEYER (2018): *Psychologie der Persönlichkeit*, Springer Verlag, 6. Auflage, Berlin - Heidelberg.

BAMBERG, G. / F. BAUR / M. KRAPP (2017): *Statistik*, 18. Auflage, Oldenbourg Verlag, München.

BÜNING, H. / G. TRENKLER (1994): *Nichtparametrische statistische Methoden*, 2. Auflage, de Gruyter Verlag, Berlin.

CAPUTO, A. / L. FAHRMEIR / R. KÜNSTLER / S. LANG / I. PIGEOT / G. TUTZ (2009): *Arbeitsbuch Statistik*, 5. Auflage, Springer Verlag, Berlin - Heidelberg.

EID, M. / M. GOLLWITZER / M. SCHMITT (2015): *Statistik und Forschungsmethoden*, 4. Auflage, Beltz Verlag, Weinheim - Basel.

FAHRMEIR, L. / A. HAMERLE / G. TUTZ (Hrsg.; 1996): *Multivariate statistische Verfahren*, 2. Auflage, Springer Verlag, Berlin - Heidelberg - New York.

FAHRMEIR, L. / T. KNEIB / S. LANG (2009): *Regression – Modelle, Methoden und Anwendungen*, 2. Auflage, Springer Verlag, Berlin - Heidelberg - New York.

FAHRMEIR, L. / C. HEUMANN / R. KÜNSTLER / I. PIGEOT / G. TUTZ (2016): *Statistik*, 8. Auflage, Springer Verlag, Berlin - Heidelberg.

HANDL, G. / H. KUHLENKASPER, (2018): *Einführung in die Statistik – Theorie und Praxis mit R*, Springer Verlag, Berlin - Heidelberg.

HANDL, G. / H. KUHLENKASPER, (2017): *Multivariate Analysemethoden – Theorie und Praxis mit R*, 3. Auflage, Springer Verlag, Berlin - Heidelberg.

KAUERMANN, G. / H. KÜCHENHOFF (2011): *Stichproben – Methoden und praktische Umsetzung mit R*, Springer Verlag, Berlin - Heidelberg.

LUHMANN, M. (2020): *R für Einsteiger – Einführung in die Statistiksoftware für die Sozialwissenschaften*, Beltz Verlag, Weinheim - Basel.

- MAWSON, A. / B. D. RAY / A. R. BUIYAN / B. JACOB (2017): Pilot comparative study on the health of vaccinated and unvaccinated 6- to 12-year-old US-children. *Journal of Transnational Science*, Vol. 3, S. 2 -12, Open Access.
- MAYER-SCHÖNBERGER, V. / K. CUPIER (2017): *Big Data – Die Revolution, die unser Leben verändern wird*, 3. Auflage, Redline Verlag, München.
- MITTAG, H.-J. (2017): Interaktive Visualisierung statistischer Konzepte und gesellschaftsrelevanter Daten im Statistikunterricht. *Mathematik im Unterricht*, Heft 8, S. 173 – 180, Universität Salzburg, Österreich.
- MOSLER, K. / F. SCHMID (2009): *Beschreibende Statistik und Wirtschaftsstatistik*, 4. Auflage, Springer Verlag, Berlin - Heidelberg.
- MOSLER, K. / F. SCHMID (2011): *Wahrscheinlichkeitsrechnung und schließende Statistik*, 4. Auflage, Springer Verlag, Heidelberg.
- RAHLF, TH. (2018): *Datenvisualisierung mit R*, 2. Auflage, Springer Verlag, Berlin - Heidelberg.
- RASCH, D. / K. D. KUBINGER / YANAGIDA, T. (2011): *Statistik in der Psychologie – vom Einführungskurs zur Dissertation*, Hogrefe Verlag, Göttingen.
- RINNE, H. / H.-J. MITTAG (1995): *Statistische Methoden der Qualitäts sicherung*, 3. Auflage, Hanser Verlag, München.
- RINNE, H. / H.-J. MITTAG (1999): *Prozessfähigkeitsmessung für die industrielle Praxis*, Hanser Verlag, München.
- SCHLITTGEN, R. (2012): *Einführung in die Statistik – Analyse und Modellierung von Daten*, 12. Auflage, Oldenbourg Verlag, München.
- SCHLITTGEN, R. (2013): *Regressionsanalysen mit R*, Oldenbourg Verlag, München.
- SCHNELL, R. / P. B. HILL / E. ESSER (2018): *Methoden der empirischen Sozialforschung*, 11. Auflage, De Gruyter / Oldenbourg Verlag, München.
- SEDLMEIER, P. / F. RENKEWITZ (2018): *Forschungsmethoden und Statistik in der Psychologie*, 3. Auflage, Pearson Verlag, München.
- STELAND, A. (2016): *Basiswissen Statistik - Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 4. Auflage, Springer Verlag, Berlin - Heidelberg.
- STORM, R. (2007): *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle*, 7. Auflage, Fachbuchverlag Leipzig / Carl Hanser Verlag, München - Wien.
- TOUTENBURG, H. / C. HEUMANN (2008): *Induktive Statistik - Eine Einführung mit R und SPSS für Windows*, 4. Auflage, Springer Verlag, Berlin - Heidelberg.
- TOUTENBURG, H. / C. HEUMANN (2009): *Deskriptive Statistik - Eine Einführung in Methoden und Anwendungen mit SPSS*, 7. Auflage, Springer Verlag, Berlin.

TOUTENBURG, H. / M. SCHOMAKER / M. WISSMANN (2009): *Arbeitsbuch zur deskriptiven und induktiven Statistik*, 2. Auflage, Springer Verlag, Berlin.

TSCHIRK, W. (2018): *Bayes-Statistik für Human- und Sozialwissenschaften*, Springer Verlag, Berlin.

WILKINSON, R. / K. PICKETT (2010): *The Spirit Level – Why Greater Equality Makes Societies Stronger*, Penguin Books, London.

WOLLSCHLÄGER, D. (2020): *Grundlagen der Datenanalyse mit R*, 5. Auflage, Springer Verlag, Berlin - Heidelberg.

ZUCCHINI, W. / A. SCHLEGEL / O. NENADIC / S. SPERLICH (2009): *Statistik für Bachelor- und Masterstudenten*, Springer Verlag, Berlin - Heidelberg.

21.2 Statistik-Software und Online-Ressourcen

Statistik-Software



Das R-Projekt



Aufgaben mit R-Code
zu den Lösungen



RStudio mit
nützlichen Paketen

In den letzten Jahren hat sich zunehmend **R** etabliert – eine sehr flexible Programmiersprache für die Auswertung und grafische Darstellung von Daten. Sie ist als Open-Source-Software kostenfrei installierbar und heute in der statistischen Praxis und in der Statistikausbildung nicht mehr wegzudenken. Neue Verfahren können über einen umfassenden Katalog an Paketen sofort genutzt werden. Mitte März 2020 waren über 14 000 Pakete verfügbar. Zu den in **R** enthaltenen Funktionen und Paketen ist eine umfangreiche Dokumentation vorhanden – eine Art **Nachschlagewerk und Kochbuch**. Viele der Datenanalysen und Grafiken im vorliegenden Buch wurden mit **R** erzeugt. Der R-Code zu den **Lösungen** der Übungsaufgaben ist ebenfalls Teil dieses Werks.

Während **R** keine grafische Benutzeroberfläche hat und über eine Kommandozeile bedient wird, ist **RStudio** eine integrierte Entwicklungsumgebung für **R** mit grafischer Benutzeroberfläche. **RStudio** vereinfacht das Arbeiten mit **R** und kann ebenfalls kostenfrei heruntergeladen werden. Für interaktive Webanwendungen in **R** gibt es unter anderem das Paket **Shiny**. Nutzer können im Browser beispielsweise Parameter einer Verteilungsfunktion oder Ausprägungen einer Variablen ändern und die damit verbundenen Effekte sofort sehen. **Shiny** bietet aufgrund einer eigenen, mit **R** kompatiblen Syntax die Möglichkeit, dass **R**-Nutzer mit geringer Erfahrung in Webentwicklung relativ schnell ansprechende interaktive Visualisierungen erstellen können. Es gibt eine offen zugängliche Webseite mit **Shiny-Anwendungen**, auf der man auch den Quellcode für die aufgeführten Beispiele findet.

Excel ist ein weit verbreitetes kommerzielles Programm von Microsoft zur Tabellenkalkulation. Nahezu alle Statistikprogramme können Excel-Dateien einlesen. Die Software zeichnet sich durch eine einfache Datenvizualisierung aus. Über das Funktionsmenü lassen sich auch elementare Berechnungen statistischer Kennwerte ausführen. Die Anwendung einzelner induktiver Verfahren – z. B. t-Test, lineare Regression oder Korrelationsanalyse – ist ebenfalls möglich. Allerdings ist Excel für darüber hinausgehende Berechnungen eher ungeeignet. Für diese werden spezielle Zusatzmodule (Add-Ons, Plug-Ins) benötigt, die teilweise von Drittanbietern entwickelt wurden und in der Regel kostenpflichtig sind (z.B. **XLfit** oder **XLSTAT**).

Es ist eine große Stärke und zugleich eine Schwäche von Excel, dass sich Berechnungen schnell und pragmatisch durchführen lassen, ohne dass besondere formale Regeln, etwa zum Aufbau einer Tabelle, beachtet werden müssen. Als analytischer „Skizzentisch“ leistet Excel durchaus gute Dienste, verleitet aber dazu, die Dokumentation und das saubere wissenschaftliche Arbeiten zu vernachlässigen. Spätestens wenn man nach einigen Monaten nicht mehr überblickt, was man eigentlich auswerten wollte, zeigt sich der Nachteil solcher „quick-and-dirty“-Analysen.

Python ist wie **R** eine Open-Source-Programmiersprache. Sie hat viele verschiedene Anwendungsgebiete und vermag auch im Bereich „Datenwissenschaft“ einiges zu bieten. So verfügt Python über eine große Zahl an Bibliotheken

zum Maschinellen Lernen, teilweise in Verbindung mit neuronalen Netzen (sog. „Deep Learning“). In Python sind auch statistische Verfahren enthalten oder über zusätzliche Pakete zugänglich. Anders als in R sind in Python die in den Datenwissenschaften häufig verwendeten Methoden und Modelle in nur wenigen Paketen zusammengefasst. Die Programmiersprache zeichnet sich durch eine übersichtliche Syntax aus und kann über eine Entwicklungsumgebung genutzt werden, zum Beispiel mit Jupyter oder PyCharm. Visualisierungen sind anhand von Paketen wie Matplotlib oder Plotly einfach realisierbar.

SPSS („Statistical Packages for Social Sciences“) ist eine kostenpflichtige Software-Plattform, die seit einigen Jahren zu IBM gehört und vor allem in den Sozialwissenschaften verwendet wird. Sie verfügt über eine nutzerfreundliche Oberfläche und bietet Schritt-für-Schritt-Anleitungen für eine Reihe statistischer Verfahren. So ist es neben der Programmierung mit einer eigenen Syntax-Sprache auch möglich, statistische Auswertungen über die Fensterführung durchzuführen, ohne Programmierkenntnisse zu haben. Die wichtigsten Verfahren aus der Statistik und Datenmanipulation sind im Programm vorhanden, allerdings teilweise nur über ebenfalls kostenpflichtige Erweiterungspakete.

Stata wird vor allem in den Wirtschaftswissenschaften verwendet. Das kostenpflichtige Programm bietet die Möglichkeit zur Verwendung sowohl über die Syntax als auch über ein Menü. Zusätzlich zu den Standardverfahren eignet sich Stata auch für die Analyse von Ereignis- und Paneldaten. Da alle Stata-Pakete erst nach weitläufigen Tests erscheinen, sind neuere statistische Verfahren oft erst verspätet nutzbar.

SAS („Statistical Analysis System“) ist eine kommerzielle Programmiersprache zur Datenanalyse und -visualisierung. Das Programm kann sowohl über eine grafische Benutzeroberfläche wie auch durch das Schreiben von Code in einer eigenen Programmiersprache mit Makrofunktionalität verwendet werden. SAS verfügt über eine große Bibliothek vorgefertigter Verfahren, ist allerdings wie SPSS und Stata kostenpflichtig.

Aus SAS ging die Statistiksoftware **JMP** hervor, die vor allem für die explorative Datenanalyse und interaktive Visualisierung verwendet wird und mit einer grafischen Benutzeroberfläche arbeitet.

Da die neuesten statistischen Verfahren in Statistikpaketen wie SPSS, Stata, SAS oder JMP nicht immer sofort verfügbar sind, ermöglichen viele kommerzielle Anbieter von Statistikprogrammen mittlerweile die Integration von R-Code. Mit dem Add-On RExcel kann R auch innerhalb von Excel-Anwendungen genutzt werden.



Statistik-Web-App

Online-Ressourcen in diesem Lehrtext

Im Vorwort wurde darauf hingewiesen, dass dieses Lehrbuch in einer Printfassung erscheint, die über einen individualisierten Zugangscode auch den Zugriff auf eine interaktive pdf-Version erlaubt („eBook Inside“). Die Online-Variante weist gegenüber der Printfassung einen bedeutenden Mehrwert auf. Dieser beruht darauf, dass hier direkte Verknüpfungen zu interessanten Web-Adressen sowie zu interaktiven statistischen Experimenten, Lehrvideos und tagestützten Animationen realisiert wurden. Die statistischen Experimente stammen teilweise aus einer unter

- <https://www.hfh-fernstudium.de/statistik-app>



Weiterentwickelte Statistik-Web-App

eingestellten virtuellen Bibliothek („Statistik-Web-App“) mit interaktiven Lernobjekten, die für die Hamburger Fern-Hochschule bis Ende 2018 realisiert wurde. Die Elemente der App sind plattformunabhängig einsetzbar. Die in diesem Buch verwendeten Lernobjekte stammen zu einem größeren Teil aus einer weiterentwickelten, unter

- <https://stat.iks-hagen.de/app>

frei zugänglichen virtuellen Sammlung („neue Statistik-Web-App“). Diese entstand ab 2019 in Zusammenarbeit mit dem Institut für Kooperative Systeme, einem Aninstitut der FernUniversität Hagen.

Die erwähnten Lehrvideos findet man unter

- <http://www.mittag-statistik.de/videos>.

Sonstige Online-Ressourcen

Eine Vielzahl von Statistikvorlesungen und Statistikkursen steht heute online in der Gestalt von *MOOCs* zur Verfügung. Beispiele zur Statistik sind auf den E-Learning-Plattformen *Coursera* und *EdX* zu finden, kostenpflichtigen Kurse z. B. unter

- <https://www.coursera.org/specializations/statistics>;
- <https://www.edx.org/course/introduction-statistics-descriptive-uc-berkeleyx-stat2-1x#.VSUBUBzwCpo>.

Neben Videos mit mehrteiligen Vorlesungen gibt es auch zahlreiche Einzelvideos und Animationen, die sich nur auf einzelne statistische Verfahren beziehen, sowie Sammlungen solcher Ressourcen. Beispiele für letztere sind die virtuellen Bibliotheken *SOCR* (*Statistics Online Computational Resources*) und *CAUSE* (*Consortium for the Advancement of Undergraduate Statistics Education*).

Eine dynamische Visualisierung ausgewählter Daten der internationalen amtlichen Statistik wird von der schwedischen *Gapminder-Stiftung* angeboten. Man findet hier z. B. ein dynamisches Blasendiagramm, das für die Staaten dieser Welt für die letzten zwei Jahrhunderte zeigt, wie sich die Lebenserwartung Neugeborener in Abhängigkeit vom Wohlstand der Bevölkerung verändert hat. Der Flächeninhalt der Kreise (Blasen) spiegelt die Bevölkerungsgröße wider.

Wie sich die Lebenserwartung Neugeborener in der Welt entwickelt hat, kann man auch unter

- <http://www.worldlifeexpectancy.com/country-history>

- <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html>

studieren. Vor allem in afrikanischen Ländern ist die Lebenserwartung noch immer sehr niedrig.

Die freie Enzyklopädie „Wikipedia“ umfasst ein informatives Statistikportal, zugänglich via

- <http://de.wikipedia.org/wiki/Portal:Statistik>.

Erwähnt sei auch *Tableau Public*, ein Portal für die interaktive Online-Visualisierung von Daten:

- <http://www.tableausoftware.com/public/>.

Interessant ist auch das für Laien konzipierte Wissensportal „Statistics Explained“ von Eurostat. Man findet es unter

- http://ec.europa.eu/eurostat/statistics-explained/index.php/Main_Page.

Eine Veranschaulichung von Regionaldaten für die EU unter Einbezug von Landkarten gibt es bei Eurostat unter

- <http://ec.europa.eu/eurostat/cache/RCI/#?vis=nuts1.labourmarket>.

Genannt sei auch ein mit EU-Mitteln gefördertes Projekt *ProCivicStat*, das sich der Förderung von Daten- und Methodenkompetenz widmete.

Gelingene interaktive Präsentationen von Daten der deutschen amtlichen Statistik sind auf den Internetseiten des Statistischen Bundesamts eingestellt. Es sei hier nochmals auf die dynamische Visualisierung der **Bevölkerungsstruktur für Deutschland** hingewiesen sowie auf das **Preiskaleidoskop**, das eine benutzerfreundliche Visualisierung der Ausgabenanteile des Warenkorbs des „Durchschnittsverbrauchers“ bietet.

Erwähnt sei schließlich auch die von der Europäischen Zentralbank unter

- http://www.ecb.europa.eu/stats/macroeconomic_and_sectoral/hicp/html/inflation.en.html

angebotene Visualisierung der Inflationsentwicklung in der Eurozone oder eine ähnliche Präsentation von Daten zur Entwicklung der Staatsfinanzen unter

- http://www.ecb.europa.eu/stats/macroeconomic_and_sectoral/government_finance/html/dashboard.en.html.

Es gibt auch Ansätze, interaktive Datenvisualisierung mit Kunst zu verbinden. Ein Beispiel ist die Visualisierung des Well-Being-Indexes der OECD unter

- <http://truth-and-beauty.net/projects/oecd-better-life-index>.

21.3 Symbolverzeichnis

Griechische Buchstaben

In der Statistik werden Merkmale oder Zufallsvariablen, deren Ausprägungen und auch Kenngrößen häufig mit griechischen Buchstaben belegt, z. B. wird Φ für die Verteilungsfunktion der Standardnormalverteilung und μ sowie σ für den Erwartungswert resp. die Standardabweichung einer Zufallsvariablen verwendet. Da das griechische Alphabet eventuell nicht allen Lesern vollständig geläufig ist, ist es nachstehend mit Aussprachehinweisen wiedergegeben.

Klein- buchstabe	Groß- buchstabe	Aussprache	Klein- buchstabe	Groß- buchstabe	Aussprache
α	A	Alpha	ν	N	Nü
β	B	Beta	ξ	Ξ	Xi
γ	Γ	Gamma	o	O	Omikron
δ	Δ	Delta	π	Π	Pi
ϵ	E	Epsilon	ρ	P	Rho
ζ	Z	Zeta	σ	Σ	Sigma
η	H	Eta	τ	T	Tau
θ	Θ	Theta	y	Y	Ypsilon
ι	I	Iota	ϕ	Φ	Phi
κ	K	Kappa	χ	X	Chi
λ	Λ	Lambda	ψ	Ψ	Psi
μ	M	Mü	ω	Ω	Omega

Mathematische Symbole und Schreibweisen

Auch die in der Mathematik gängigen Abkürzungen und Schreibweisen sind möglicherweise nicht jedem Leser sehr vertraut. Daher sind in den beiden folgenden Tabellen einige der in diesem Manuscript häufiger auftretenden Notationen zusammengestellt. Die erste Tabelle fasst Schreibweisen für Mengen und für Operationen mit Mengen zusammen.

Symbol	Beschreibung
$\{\dots\}$	Menge von Objekten
$a \in A; a \notin A$	a ist ein bzw. ist kein Element der Menge A
$A \subset B;$	A ist Teilmenge von B
$A \cap B; A \cup B$	Schnittmenge bzw. Vereinigungsmenge der Mengen A und B
$A \setminus B$	Differenzmenge von A und B

In der folgenden Tabelle sind weitere Schreibweisen aus Mathematik und Statistik wiedergegeben, z. B. aus der mathematischen Logik, zum Größenvergleich von Termen sowie Notationen für Vektoren und Matrizen oder für Verteilungsaussagen.

Symbol	Beschreibung
(a, b)	geordnetes Paar
$f : A \rightarrow B$	Funktion f , bildet A nach B ab
$\Rightarrow; \Leftrightarrow$	Implikation (daraus folgt); Äquivalenz (genau dann)
$a = b; a \neq b$	a und b sind gleich; a und b sind ungleich
$a := b; a \approx b$	a ist durch b definiert; a und b sind näherungsweise gleich
$a < b; a > b$	a ist kleiner bzw. größer als b
$a \leq b; a \geq b$	a ist kleiner oder gleich b ; a ist größer oder gleich b
∞	unendlich
$\sum_{i=1}^n a_i$	Summe der Terme a_1, a_2, \dots, a_n
$\sum_{i=1}^{\infty} a_i$	Summe der Terme a_1, a_2, \dots
$n!$	Produkt $n \cdot (n - 1) \cdot \dots \cdot 1$ der n ersten natürlichen Zahlen
$\binom{n}{k}$	Binomialkoeffizient; Quotient aus $n!$ und $(n - k)! \cdot k!$
$\sqrt{a}; a $	Wurzel aus a ; Betrag von a
$\exp x, e^x$	Exponentialfunktion
$\mathbf{a}, \mathbf{A}, \mathbf{I}$	Vektor, Matrix, Einheitsmatrix
$rg \mathbf{A}$	Rang der Matrix \mathbf{A}
$X \sim N(\mu; \sigma^2)$	X ist normalverteilt mit Erwartungswert μ und Varianz σ^2
$X \sim B(n; p)$	X ist binomialverteilt mit Parametern n und p
$X \sim H(n; M; N)$	X ist hypergeometrisch verteilt mit Parametern n, M, N
$Y \sim \chi_n^2; Y \sim t_n$	Y ist χ_n^2 - bzw. t -verteilt mit n Freiheitsgraden
$Y \sim F_{m;n}$	Y ist F -verteilt mit m und n Freiheitsgraden
$x_p; z_p$	p -Quantil der Normal- und der Standardnormalverteilung
$\chi_{n;p}^2; t_{n;p}; F_{m;n;p}$	p -Quantil der χ^2 -, t - und F -Verteilung

21.4 Autorenregister

A	
Asendorpf, J. B.	260
B	
Baggini, J.	29
Bamberg, G.	45, 104, 227, 347
Baur, F.	45, 104, 227, 347
Boecking, B.	16
Brown, E. N.	16
Buiyan, A. R.	48
Busch, P.	13f
Büning, H.	281
C	
Caputo, A.	172, 353, 355
Cukier, K.	10
D	
Dräger, J.	19
E	
Esser, E.	37
F	
Fahrmeir, L.	53, 69, 102, 172, 182, 215, 231, 288, 291, 303, 306, 318, 353, 355
G	
Gowda, T.	16
H	
Hamerle, A.	318
Handl, A.	152, 318
Heumann, C.	53, 69, 94, 102, 142, 166, 182, 215, 228, 231, 288, 291, 301, 313, 318, 347
Hill, P. B.	37
Hindinger, C.	13f
Hundman, K.	16
J	
Jacob, B.	48
Jahnke, T.	18
K	
Kauermann, G.	45
L	
Kejriwal, M.	16
Kneib, T.	303, 306
Krapp, M.	45, 104, 227, 347
Kubinger, K. D.	310
Kuhlenkasper, T.	152, 318
Küchenhoff, H.	45
Künstler, R.	53, 69, 102, 172, 182, 215, 231, 288, 291, 318, 353, 355
M	
Lang, S.	172, 303, 306, 353, 355
Lankau, R.	19
Layard, R.	29
Luhmann, M.	196
N	
Nenadic, O.	221, 223, 226, 228
Neyer, F. J.	260
P	
Pickett, K.	108
Pigeot, I.	53, 69, 102, 172, 182, 215, 231, 288, 291, 318, 353, 355
R	
Rahlf, Th.	58
Rasch, D.	310
Ray, B. D.	48
Renkewitz, F.	28, 37
Rinne, H.	200, 245
S	
Sarrazin, T.	68
Schlegel, A.	221, 223, 226, 228
Schlittgen, R.	78, 182, 252, 259, 318
Schmid, F.	169, 199, 233, 243, 259, 279
Schnell, R.	37
Schomaker, M.	346, 353
Schüller, K.	13f
Sedlmeier, P.	28, 37

- Sperlich, S. 221, 223, 226, 228
Steland, A. 53, 85
Storm, R. 200

T

- Toutenburg, H. 94, 102, 142, 166, 228,
231, 291, 301, 313, 346f, 353
Trenkler, G. 281
Tschirk, W. 172, 256
Tutz, G. 53, 69, 102, 172, 182, 215,
231, 288, 291, 318, 353, 355

W

- Wilkinson, R. 108
Wissmann, M. 346, 353
Wollschläger, D. 196

Y

- Yanagida, T. 310

Z

- Zucchini, W. 221, 223, 226, 228

21.5 Sachregister

A	
ABC-Analyse	102
Ablehnbereich	263
Absolutskala	26
Abweichung	
mittlere quadratische	81, 232, 243
AIC	306
Aktienindex	116
ALLBUS	32, 40
Altenquotient	64
Alternativhypothese	260
Alternativtest	258
Annahmebereich	263
ANOVA	308
Anpassungstest	182, 258
arithmetisches Mittel	75
getrimmtes	78
gewichtetes	78
Armut	79
Auswahl	
geordnete	167
ungeordnete	167
Auswahlbias	46
Auswahlgesamtheit	41
Autokorrelation	296
Axiome von Kolmogoroff	163
B	
Balkendiagramm	53
gestapeltes	57
Baumdiagramm	128
Bayes-Statistik	172, 256
bedingte Wahrscheinlichkeit	171
Befragung	32
Beobachtung	33
Bernoulli-Experiment	189, 193, 198, 246
Bernoulli-Kette	189, 198, 246
Bernoulli-Verteilung	188, 203, 246, 303
Bestimmtheitsmaß	292, 302
Better-Life-Index	121
Bevölkerungspyramide	63f, 66f
Beziehungszahl	110
Bias	241
Big Data	8, 10
Binomialkoeffizient	168
Binomialtest	275
Binomialverteilung	194, 202, 321
Approximation	234
Binärvariable	302
Biometrie	9, 283
Blasendiagramm	53, 139, 386
Boxplot	87, 317
C	
CAPI	32
CATI	32
Chi-Quadrat	
-Koeffizient	142
-Test	258, 276
-Verteilung	221, 331
Koeffizient	281
Cramér's V	143
D	
Data Literacy	13
Data Mining	8
Daten	
gruppierte	52
klassierte	52
Primär-	23
Roh-	23
zensierte	92
Datenanalyse	
bivariate	123
explorative	7
multivariate	51
univariante	51
Datenerhebung	
Primärerhebung	31
Sekundärerhebung	31
Tertiärerhebung	31
Datengewinnung	
anhand von Stichproben	39
durch Befragung	32
durch Beobachtung	33
durch Teilerhebung	39
durch Vollerhebung	39
mit nicht-reaktiven Verfahren	34
per Experiment	35
Datenjournalismus	10

Datenkompetenz	13	einer diskreten Zufallsvariablen	
Datenwissenschaft	9	190	
DAX	38, 116	einer stetigen Zufallsvariablen	211
Determinationskoeffizient	292	globaler	311
Dezil	86	zweier Zufallsvariablen	191
Dichtefunktion	180, 208	ethische Kompetenz	15
bedingte	231	Eurostat	29, 32, 50, 111, 387
der Chi-Quadrat-Verteilung	221f	Experiment	31, 35
der F-Verteilung	226	Bernoulli-	189, 246
der Normalverteilung	213	interaktives	19
der Rechteckverteilung	209	nach Laplace	163
der Standardnormalverteilung	215,	Quasi-	37
224, 328		Zufalls-	162
der t-Verteilung	223f	explorative Datenanalyse	7
gemeinsame	231	Exponentialverteilung	182
Donut-Diagramm	53	Exzess	228
Dow Jones Index	116		
Durchschnitt	75	F	
gleitender	79	F-Test	258, 314
		F-Verteilung	226, 314, 334
E		Faktor	308
einfache Zufallsstichprobe	44	Faktorenanalyse	318
Einflussfaktoren	36	Faktorstufe	308
Einkommensverteilung	103, 105, 108	Fehlalarmrate	176
Einstichproben-Test	258, 277	Fehler	
empirische Verteilungsfunktion	70	α	263
Ereignis	160	β	263
disjunktes	160	1. Art	263
Elementar-	159	2. Art	263
Komplementär-	160	mittlerer quadratischer	242
sicheres	160	Freiheitsgrade	
unabhängiges	172, 229	der Chi-Quadrat-Verteilung	222
unmögliches	160	der F-Verteilung	226
Ergebnismenge	160	der t-Verteilung	223
Erwartungstreue	241	Full Fact	78
asymptotische	241		
Erwartungswert	181	G	
der Bernoulli-Verteilung	246	Gauß-Test	258, 263, 277
der Binomialverteilung	194	geometrische Verteilung	182
der Chi-Quadrat-Verteilung	222	geometrisches Mittel	79
der hypergeometrischen Verteilung	199	geschichtete Stichprobe	44
der Normalverteilung	213	GESIS	111
der Null-Eins-Verteilung	192	Gini-Koeffizient	103
der Rechteckverteilung	212	normierter	105
der t-Verteilung	223	Gleichverteilung	
des Stichprobenmittelwerts	232,	diskrete	180, 185
243		stetige	208
		gleitender Durchschnitt	79

Gliederungszahl	109	K	
Global Competitiveness Index	121	Kacheldiagramm	58
Grundgesamtheit	22	Kaplan-Meier-Kurve	94
dichotome	189, 193	Kardinalskala	25
Gütfunktion	267	Kerndichteschätzer	68
		Kerzendiagramme	89
H		KI	8
Häufigkeit		Clumpenstichprobe	45
absolute	52, 124	Kombinatorik	166
bedingte	130, 132	Konfidenzintervall	240
relative	53	für den Erwartungswert	249
Häufigkeitsverteilung	53	für einen Anteilswert	252
absolute	69, 124	Konfidenzniveau	249, 252
bedingte	131f	Kontingenztabelle	124, 175, 281
relative	69, 124	Kontingenztafel	124, 231, 281
Haupteffekte	318	Kontingenztest	281
Herfindahl-Index	106	Kontrollgruppe	37
Histogramm	61	Korrelation	48, 148
Homoskedastizität	296	partielle	153
Human Development Index	50, 118f	Schein-	48
Human Poverty Index	118	Korrelationskoeffizient	148, 235, 293
hypergeometrische Verteilung	199	partieller	152
Hypothese		Kovarianz	
Alternativ-	260	empirische	146, 235
Null-	260	theoretische	234
Häufigkeit		KQ-Schätzung	287, 300
relative	124, 190	der Regressionskoeffizienten	288, 291
Häufigkeitsverteilung		der Varianz der Störvariablen	289
relative	184	Eigenschaften	291
I		Kreisdiagramm	53
ILO	30	3D-Darstellung	57
Imputation	298	Kreuztabelle	124
Indikatoren	111	Kryptowährungen	89
Leit-	6	Kurtosis	228
zusammengesetzte	6, 114		
Inferenz	42	L	
Inflationsrate	116	Laplace-Experiment	163
Inflationsrechner	116	Letalität	111
Information Literacy	14	Likert-Skala	27, 121
Informationskompetenz	14	logistisches Regressionsmodell	304
Intelligenzmessung	219	Logit-Modell	304
Interquartilsabstand	86	Lognormalverteilung	182
Intervallschätzung	240, 248	Lorenzkurve	100
Intervallskala	25	Längsschnittstudie	38
J		M	
Jugendquotient	64	MANOVA	308

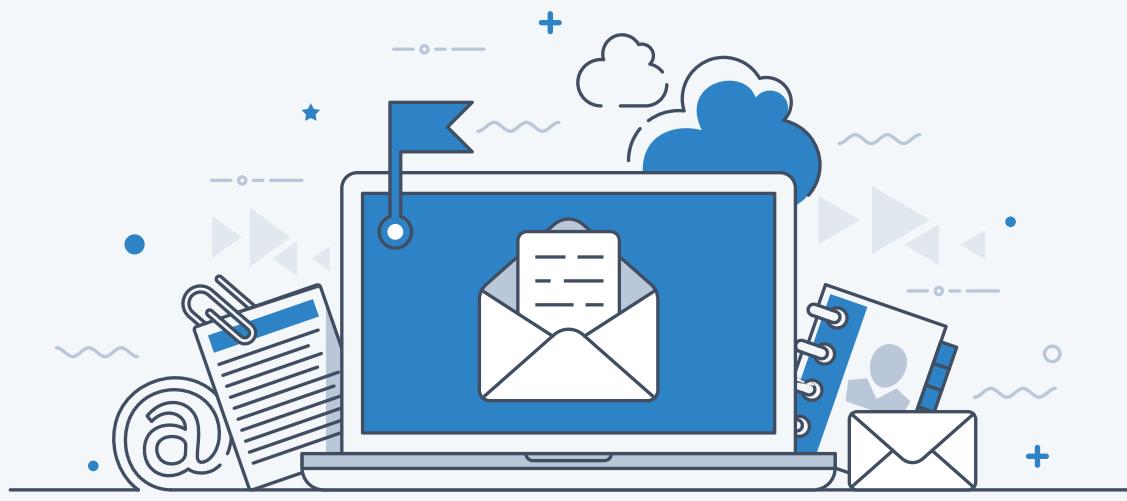
Marimekko-Diagramm	58
maschinelles Lernen	8
Matrix	
der Regressoren	295
Maximum-Likelihood-Methode	288
Maßzahl	109
Median	74, 190, 192, 212
mittlere absolute Abweichung	84
Medianalter	64
Merkmal	22
Ausprägung	22
binäres	127, 175
dichotomes	127, 175
diskretes	24
qualitatives	28
quantitatives	28
stetiges	24
Merkmalsträger	22
Methode der kleinsten Quadrate	287, 300
Methodenkompetenz	
statistische	13
metrische Skala	25
Mikrozensus	37, 39
Minimum-Varianz-Portfolio	236
Mittelwert	75, 190
bei gruppierten Daten	76
getrimmter	78
gewichteter	78
Stichproben-	232, 243
mittlere quadratische Abweichung	81, 232, 243
mittlerer quadratischer Fehler	242
ML-Schätzung	288
Modalwert	74
Modell	
deterministisches	179
parametrisches	8
stochastisches	179
Moderatorvariable	153
Modus	74
MOOC	18, 386
Mortalität	111
Mosaik-Plot	58, 129
MSE	242
Multikollinearität	296
N	
nicht-reaktives Erhebungsverfahren	34
Nominalskala	25
Normalverteilung	180, 213, 328
Null-Eins-Verteilung	188, 194, 303
Nullhypothese	260
O	
Objektivität	28
OECD	50, 111, 118, 121
Oekonometrie	9
Operationalisierung	28
Ordinalskala	25
Overcoverage	42
P	
p-Quantile	212, 330ff, 334
p-Wert	270
Panel	38
partielle Korrelation	153
Permutation	168
Phi-Koeffizient	142
PISA-Studien	17
Poisson-Verteilung	182, 303
Politbarometer	4, 55, 74, 126, 133, 136, 143, 247, 282, 353
Population	22
Portfoliotheorie	236
Preiskaleidoskop	117
Primärdaten	23
Primärerhebung	31
Proxyvariable	153
Prozessfähigkeitsindizes	244
Prädiktor	285
Prüfplan	200
Prüfstatistik	231, 261
Prüfvariable	261
Psychometrie	9
Punktschätzung	240, 248
Q	
Quantile	
der Chi-Quadrat	
-Verteilung	331
der Chi-Quadrat-Verteilung	223, 277, 331
der F-Verteilung	226, 334
der Normalverteilung	218

der Standardnormalverteilung	217, 225, 330	gestapeltes	57
der t-Verteilung	224, 332	mit Doppelsäulen	137
einer empirischen Verteilung	85	Satz von Bayes	172
einer theoretischen Verteilung	181, 192	Scheinkorrelation	48, 151
p-	85, 192, 212	Schichtung	44
Quartil		Schiefe	
oberes	86, 192	empirische	228
unteres	86, 192	theoretische	227
Quartilsabstand	86	Schwankungsintervall	181
Quasi-Experiment	37	der Standardnormalverteilung	218
Querschnittsstudie	38	für den Anteilswert	253
Quotenauswahl	45	Schätzfunktion	231, 241, 290
R		Schätzung	
R ...	20, 201, 203, 275, 280, 305f, 384	der Varianz	243, 313
Randhäufigkeiten		des Erwartungswerts	243
absolute	125	für Anteilswerte	246
relative	125	Intervall-	240, 248
Randverteilung	125, 231	KQ-	287, 300
Rangkorrelationskoeffizient	155	ML-	288
Rangskala	25	Punkt-	240, 248
Ratingskala	27	von Effekten	313
Ratioskala	25	Sekundärerhebung	31
Rechteckverteilung	208	Sensitivität	176
Regressionsanalyse	284	Signifikanzniveau	265
Regressionsfunktion	284	empirisches	270
Regressionsgerade	285	Signifikanztest	258
Regressionshyperebene	301	SIPRI	88, 112
Regressionskoeffizient	285	Skala	
Regressionsmodell		Absolut-	26
einfaches	284	Intervall-	25
kategoriales	304	Kardinal-	25
lineares	285, 287	Likert-	27, 121
logistisches	304	metrische	25
Logit-	304	Nominal-	25
multiples	284, 294	Ordinal-	25
nicht-lineares	285	Rang-	25
verallgemeinertes lineares	303	Rating-	27
Reliabilität	28	Ratio-	25
Residuen	287, 292, 300f	Verhältnis-	25
Ringdiagramm	53	SOEP	29, 40
Rohdaten	23	Sozialkreditsystem	113
S		Spannweite	81, 190
Saeulendiagramm	53	Spezifität	176
3D-Darstellung	57, 137	Stabdiagramm	53
Standardabweichung	82, 190f, 211	Standardabweichung	82, 190f, 211
eines Schätzers	241	eines Schätzers	241
empirische	82	empirische	82
korrigierte	82, 273	korrigierte	82, 273

Stichproben-.....	273	T	
theoretische.....	181	t-Test	258, 273, 277
Standardfehler	241, 247	für verbundene Stichproben ..	279
des Stichprobenmittelwerts ..	243	t-Verteilung	223
Standardisierung	191, 211	Teilerhebung	39
Standardnormalverteilung ..	215, 328,	Tertiärerhebung	31
330		Test	
Statistical Literacy	13	Alternativ-	258
Statistik		Anpassungs-	182, 258
Anwendungsfelder	3	Binomial-	275
beschreibende	7	Chi-Quadrat-	258, 276, 281
deskriptive	7	einseitiger	258
induktive	8	Einstichproben-	258, 277
schließende.....	8, 23, 42, 179	F-	258, 314
statistische Einheit	22	für Anteilstwerte	258
Statistisches Bundesamt	3, 30, 49, 62,	für Erwartungswerte	258
111, 116f, 387		für Varianzen	258, 276
Stichprobe		Gauß-	258, 263, 277
einfache Zufalls-.....	44, 166	Kontingenz-	281
geschichtete	44	mit Messwiederholung	279
Klumpen-.....	45	nicht-parametrischer	258, 281,
mit Berücksichtigung der Anord-		316	
nung	167	parametrischer	258
mit Zurücklegen	166	Signifikanz-	258
ohne Berücksichtigung der Anord-		t-	258, 273, 277, 307
nung	167	Trennschärfe	267
ohne Zurücklegen	167	Unabhängigkeits-	258
Quotenbildung	45	zweiseitiger	258
systematische	45	Zweistichproben-	258, 279, 307
Zufalls-	42, 166	Teststatistik	231, 261
Stichproben	23	theoretische Verteilungsfunktion ..	180
abhängige	279	Trägermenge	
unabhängige	278	bei Binomialverteilung	195
unverbundene	278	bei hypergeometrischer Verteilung	
verbundene	279	200, 202	
Stichprobenerhebung	39	diskrete Zufallsvariable	183
Stichprobenfehler	42, 121	stetige Zufallvariable	207
Stichprobenfunktion	231, 241		
Stichprobenmittelwert	232, 243	U	
Stichprobenstandardabweichung ..	82,	Ueberlebensrate	94
233, 273		Ueberlebenszeitanalyse	92
Stichprobenvarianz	232, 244	Ueberlebenszeitfunktion	94
Stochastik	179	UN	50, 108, 118f
Streudiagramm	138	Agenda 2030	6, 50
Student-Verteilung	223	Unabhängigkeit	
Störvariable	36	empirische	135, 141
		von Ereignissen	172, 229
		von Zufallsvariablen	230

Unabhängigkeitstest	258, 281	eines Schätzers	241
Undercoverage	42	empirische	81, 191
Unverzerrtheit	241	korrigierte	82
asymptotische	241	Stichproben-	82, 244
Urliste	23	theoretische	181, 191
bivariate	124, 138	Varianzanalyse	307
univariate	52	einfaktorielle	308, 310
Urnenmodell	44, 166, 198	Haupteffekte	318
mit Berücksichtigung der Anordnung	167	Interaktionseffekte	318
mit Zurücklegen	166, 198	mehr faktorielle	308
ohne Berücksichtigung der Anordnung	167	mit balanciertem Design	310
ohne Zurücklegen	167, 199	mit festen Effekten	308
Urwerte	23	mit Messwiederholungen	310
mit zufälligen Effekten	308	Modell in Effektdarstellung	311, 317
V		Wechselwirkungseffekte	318
Validität	28	Variationskoeffizient	84
Value at Risk	221	Vektor	
Variable	22	der Regressionskoeffizienten	295
abhängige	36	der Störvariablen	294
Binär	302	Residuen	301
dichotome	302	Spalten	294
latente	26, 318	Venn-Diagramm	160
manifeste	29, 318	Verbraucherpreisindex	116
Moderator	153	Verhältnisskala	25
Proxy	153	Verhältniszahl	109
Prüf	261	Versuchsgruppe	37
Response	302	Versuchsplan	36, 48, 254
Stichproben	23	Versuchsplanung	35
Stör	36, 285	Verteilung	180
unabhängige	36	asymmetrische	87, 223
Zufalls	23, 179	Bernoulli	188, 203, 246, 303
Varianz	190	Binomial	194, 202, 321
bei gruppierten Daten	84	Chi-Quadrat	221, 331
der Binomialverteilung	194	diskrete	180
der Chi-Quadrat-Verteilung	222	diskrete Gleich	185
der hypergeometrischen Verteilung	199	empirische	53, 124, 180
der Normalverteilung	213	Exponential	182
der Null-Eins-Verteilung	192	F	226, 314, 334
der Rechteckverteilung	212	geometrische	182
der t-Verteilung	223	hypergeometrische	199
des Stichprobenmittelwertes	243	linksschiefe	86, 223, 227
des Stichprobenmittelwerts	232	linkssteile	86, 223, 227
einer diskreten Zufallsvariablen	190	Lognormal	182
einer stetigen Zufallsvariablen	211	Normal	180, 213, 328
		Null-Eins	188, 194, 303
		Poisson	182, 303

Rechteck-.....	208	Wilson-Intervall.....	252, 256
rechtsschiefe.....	86, 223, 227	World Press Freedom Index.....	122
rechtssteile.....	86, 223, 227	World Values Survey.....	122
Standardnormal-	215, 328, 330	Wölbung	
stetige.....	180	empirische.....	228
stetige Gleich-	208	theoretische.....	228
Student-.....	223		
t-.....	223		
theoretische.....	180	Z	
Zweipunkt-.....	188	z-Transformation.....	84, 212
Verteilungsfunktion		Zeitreihe.....	38
der Binomialverteilung .	195, 321	Zensierung.....	92
der Chi-Quadrat-Verteilung .	222	Zentraler Grenzwertsatz.....	233
der F-Verteilung	226	Zufallsexperiment.....	162
der hypergeometrischen Verteilung		Zufallsstichprobe.....	42, 166
201		Zufallsvariable	23, 179
der Normalverteilung	214	Ausprägung.....	179
der Rechteckverteilung.....	209	binäre	188
der Standardnormalverteilung	215,	diskrete	24, 179, 183, 207
328		Realisierung.....	23, 179
der t-Verteilung.....	224	stetige.....	24, 179, 207
diskreter Zufallsvariablen	180,	Zusammenhangsmaß	
184		Chi-Quadrat-Koeffizient.....	142
empirische.....	70, 180, 184	Cramér's V	143
gemeinsame	229, 231	für Zufallsvariablen	235
stetiger Zufallsvariablen.....	207	Kontingenzkoeffizient K.....	145
theoretische	180, 184	metrisch skalierte Merkmale	148
Verweildaueranalyse	92	nominalskalierte Merkmale..	142
Verzerrung.....	241	von Spearman	155
Vierfeldertafel.....	127, 145, 175	Zuverlässigkeitssanalyse.....	92
Volatilität	89	Zweipunkt-Verteilung.....	188
Vollerhebung.....	39	Zweistichproben-Test	258
		Gauß-	279
		t-	279
W			
Wahrscheinlichkeit	162		
bedingte	171, 231		
Wahrscheinlichkeitsfunktion.	180, 183,		
207			
der Binomialverteilung .	195, 321		
der hypergeometrischen Verteilung			
200, 202			
gemeinsame	231		
Wahrscheinlichkeitsrechnung...	8, 179		
Wahrscheinlichkeitsverteilung	180		
Warenkorb	116f		
Webinar	19		
Well-Being-Index.....	121, 387		
WHO.....	110		



Willkommen zu den Springer Alerts

Unser Neuerscheinungs-Service für Sie:
aktuell | kostenlos | passgenau | flexibel

Mit dem Springer Alert-Service informieren wir Sie individuell und kostenlos über aktuelle Entwicklungen in Ihren Fachgebieten.

Abonnieren Sie unseren Service und erhalten Sie per E-Mail frühzeitig Meldungen zu neuen Zeitschrifteninhalten, bevorstehenden Buchveröffentlichungen und speziellen Angeboten.

Sie können Ihr Springer Alerts-Profil individuell an Ihre Bedürfnisse anpassen. Wählen Sie aus über 500 Fachgebieten Ihre Interessensgebiete aus.

Bleiben Sie informiert mit den Springer Alerts.

Jetzt anmelden!

Mehr Infos unter: springer.com/alert

Part of SPRINGER NATURE