LEHRBUCH

Hans-Joachim Mittag

Statistik

Eine Einführung mit interaktiven Elementen

4. Auflage

freier Zugang zu interaktiven Elementen inklusive

EBOOK INSIDE

Springer Spektrum

Springer-Lehrbuch

Hans-Joachim Mittag

Statistik

Eine Einführung mit interaktiven Elementen

4., wesentlich überarbeitete und erweiterte Auflage

Springer Spektrum

Hans-Joachim Mittag
FernUniversität in Hagen
Hagen, Deutschland

ISSN 0937-7433
ISBN 978-3-662-47131-9 ISBN 978-3-662-47132-6 (eBook)
DOI 10.1007/978-3-662-47132-6

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

Mathematics Subject Classification (2010): 62-01

Springer Spektrum

Planung: Iris Ruhmann

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Springer Berlin Heidelberg ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Inhaltsverzeichnis

Vorwort

Ursprung dieses Lehrtextes

Dieses Lehrbuch ist aus einem Kurs der FernUniversität Hagen hervorgegangen, der dort in mehreren Studiengängen zum Einsatz kam (Politikwissenschaft, Soziologie, Bildungswissenschaft, Psychologie, Umweltwissenschaft). Das Buch 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, statistische Informationen nutzen, sachadäquat interpretieren und verständlich kommunizieren zu können. Diese als *statistische Methodenkompetenz* bezeichnete Qualifikation ist heutzutage unverzichtbar.

Realisierung eines Medienmixes

Im März 2011 erschien die 1. Auflage, im Juli 2012 die 2. Auflage und im März 2014 die 3. Auflage dieses Lehrbuchs. Bereits die 1. Auflage wurde mit dem Comenius-EduMedia-Siegel 2011 der Gesellschaft für Pädagogik, Information und Medien ausgezeichnet. Als innovativ wurde der Medienmix gewürdigt, der das Lehrbuch klassischer Ausprägung (Printausgabe) mit interessanten Web-Adressen und interaktiven oder dynamischen Lernobjekten verknüpft. Letztere ermöglichen es, statistische Verfahren anhand von Experimenten „auszuprobieren“ oder statistische Konzepte über tongestützte Animationen nachzuvollziehen.

Die interaktiven Experimente waren anfangs alle Java-basiert (Java-Applets) und nur für Desktops konzipiert. Sie sind fast alle in einer unter http://www.fernuni-hagen.de/jmittag/bibliothek zugänglichen virtuellen Bibliothek zusammengefasst. Diese Sammlung von mehrsprachigen Java-Applets erhielt das Comenius-EduMedia-Siegel 2012. Einige Experimente und Animationen stammen aus einem älteren Multimedia-Projekt „Neue Statistik“. Alle Java-Applets wurden Anfang 2015 zertifiziert um verschärften Sicherheitsanforderungen für Java-Applikationen zu genügen.

Verknüpfung des Lehrbuchs mit mobilen Endgeräten über QR-Codes

Um die Lauffähigkeit interaktiver statistischer Experimente auch auf Smartphones und Tablets zu erreichen, wurden Lernobjekte auf der Basis von HTML5 und Javascript entwickelt. Dabei galt es beim Design die Touchfunktionalitäten mobiler Endgeräte und das kleinere Display von Smartphones zu berücksichtigen. Die Lernobjekte wurden zunächst *englischsprachig* programmiert und in einer Web-App zusammengefasst, deren Elemente in die 3. Auflage dieses Lehrbuchs eingingen. Die über http://www.fernuni-hagen.de/jmittag/app erreichbare App erhielt das Comenius-EduMedia-Siegel 2014.

Seit Januar 2015 gibt es unter https://www.hamburger-fh.de/statistik-app eine deutlich weiterentwickelte *deutschsprachige* Fassung der Web-App mit ganz neuen Lernobjekten. Deren Eingangsportal ist über den

QR-Code (deutschsprachige Statistik-App)

nebenstehenden QR-Code erreichbar (QR = Quick Response). Diese App enthält erstmals kurze Handhabungshinweise für die plattformunabhängig einsetzbaren Lernobjekte. Die Lernobjekte der deutschsprachigen Web-App sind via QR-Code mit der vorliegenden 4. Auflage dieses Manuskripts und mit Kursmaterialien der Hamburger Fern-Hochschule verknüpft.

Plattformunabhängige Lernobjekte im Einsatz auf mobilen Endgeräten

Bedeutung der Icons am Seitenrand

Wo *Animationen mit Ton*, besondere *Web-Links* sowie statistische *Experimente für Desktops* oder *Experimente für Desktops und mobile Endgeräte* zu finden sind, wird am Rand durch neu gestaltete runde Icons sichtbar gemacht. Bei den plattformunabhängigen interaktiven Elementen sind zusätzlich *QR-Codes* platziert, damit man sie direkt von der Printfassung des Lehrbuchs mit einem mobilen Endgerät aufrufen kann:

Icons mit Verlinkung: Animation mit Ton, externer Web-Link, interaktives Java-Applet, plattformunabhängiges interaktives Lernobjekt mit QR-Code

Anhand der runden Icons und QR-Codes sowie der zahlreichen in den Text eingestreuten Web-Links (grau bei der Printausgabe, magentafarben beim e-Buch) erkennt man, wo es Zusatzangebote gibt, die sich nur bei Nutzung des Computers erschließen lassen. Lediglich die über QR-Codes hervorgehobenen Lernobjekte kann man sofort von der Printfassung aufrufen, indem man den Code mit einem Smartphone oder – besser – mit einem Tablet einscannt. Um den Mehrwert der anderen Zusatzangebote zu nutzen, wechselt man von der Print- zur e-Buchfassung.

Neben den runden Icons und den QR-Codes 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 sowie auf *Aufgaben*, *Lösungen* oder ergänzende *Literatur*:

Weitere Icons: Vorschau, Aufgabe, Lösung, Literatur

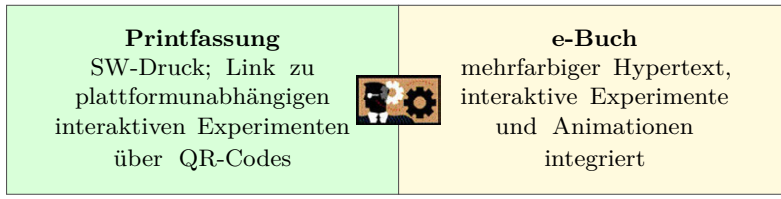
Was ist neu bei der 4. Auflage?

Das e-Buch und die Printfassung des Buches werden ab der 4. Auflage als Paket angeboten („**eBook Inside**“) – ohne Mehrpreis. In jeder Printausgabe ist ein individualisierter Code enthalten, der den Zugang zum e-Buch vermittelt. Die Printausgabe ist in Schwarzweiß gedruckt, das inhaltlich identische e-Buch ist mehrfarbig und interaktiv.

Die aktuelle Auflage unterscheidet sich von der 3. Auflage nicht nur durch die Einbindung deutschsprachiger interaktiver Lernobjekte via QR-Code und die eBook-Inside-Funktionalität, sondern auch durch zahlreiche Ergänzungen und Aktualisierungen. Völlig überarbeitet wurden u. a. die Ausführungen zum demografischen Wandel oder zu Indikatoren. Ganz neu sind z. B. ein Abschnitt zum Medienmix in der Statistikausbildung sowie Exkurse zu den Themen „Big Data“ und „Datenjournalismus“.

Vorzüge des e-Buchs

Das e-Buch ist im Gegensatz zur Printfassung nicht nur mehrfarbig, sondern auch interaktiv. Die interaktiven oder dynamischen Elemente sind 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. Das e-Buch ist ein Hypertext, bei dem man Querverweisen – im e-Buch blau markiert – zu Gleichungen, Abbildungen oder Aufgaben per Mausklick nachgehen oder vom Stichwortregister zu den dort aufgeführten Schlüsselbegriffen springen kann.



Unterschiede zwischen Buch und e-Buch

Aktualisierungen und Corrigenda

Bei Bedarf werden unter http://www.fernuni-hagen.de/jmittag/updates Aktualisierungen und Corrigenda zu diesem Lehrbuch eingestellt.

Betonung von Interdisziplinarität

Das vorliegende Werk illustriert anhand zahlreicher Beispiele, dass die Statistik alle Bereiche gesellschaftlichen Lebens durchdringt. Die verwendeten Beispiele und Exkurse – etwa zum demografischen Wandel, zur Messung von Armut und Einkommensungleichheit oder zu Risiken beim Screening zur Krebsfrüherkennung – sind aktuell und mit Hintergrundinformationen aus Online-Ausgaben überregionaler Zeitschriften verknüpft.

Einbindung von Daten der amtlichen Statistik

Exkurse können auch übersprungen werden. Einige Datensätze stammen von Eurostat, dem Europäischen Amt für Statistik in Luxemburg, an dem der Autor vier Jahre als nationaler Sachverständiger tätig war. Die Daten illustrieren die Bedeutung statistischer Informationen für die Planung und das Monitoring nationaler und supranationaler Politiken.

Struktur des Buchs

Das Buch ist in drei Teile gegliedert. Der erste Teil widmet sich der beschreibenden Statistik, der zweite Teil der Wahrscheinlichkeitsrechnung und der schließenden Statistik. Der dritte Teil (Anhänge) umfasst Aufgaben, Lösungen, statistische Tabellen und diverse Verzeichnisse.

Danksagungen

Dank für die Programmierung der interaktiven Experimente gebührt Herrn B. *Wallenborn* und Herrn A. *Michel*, beide Hagen. Herr Th. *Feuerstack*, ebenfalls Hagen, half bei der Gestaltung der LaTeX-Umgebung für dieses Manuskript. Frau A. *Dirks*, Hamburger Fern-Hochschule, gestaltete die Icons am Seitenrand. Herr Prof. Dr. R. *Münnich*, Trier, stellte Grafiken zur Armutsgefährdung zur Verfügung.

Dank gebührt ferner den nachstehenden Firmen und Institutionen, die kostenfrei Fotos oder andere Materialien zur Verfügung gestellt haben:

Fa. Böhme und Weihs GmbH, Sprockhövel (Herr Dr. N. *Böhme*)
Evonik Industries AG, Standort Essen (Herr Dr. W. *Wolfes*)
GfK, Nürnberg (Herr S. *Heller* und Herr R. *Nicklas*)
Hessischer Rundfunk, Frankfurt (Herr C. *Bender*)
JMP / SAS, Köln (Herr Dr. V. *Kraft*)
Kennesaw State University, USA (Dr. J. *McNeill*),
Q-DAS GmbH, Weinheim (Herr Dr. E. *Dietrich*)
Destatis, Wiesbaden (Frau B. *Sommer* und Herr Dr. F. *Rößger*)
TNS Infratest, München (Herr M. *Kögel*).

Zu danken ist schließlich noch Frau I. *Ruhmann*, Frau Dr. A. *Denkert*, Herrn Dr. A. *Rüdinger* und Frau A. *Herrmann* vom Springer Verlag für ihre Unterstützung bei der Vorbereitung der jetzigen 4. Auflage.

Wetter / Ruhr, im Mai 2015

Hans-Joachim Mittag
mail@mittag-statistik.de

Teil I

Beschreibende Statistik

Lernziele zu Teil I

Nach der Bearbeitung des ersten Teils dieses Manuskripts sollten Sie

- wissen, warum statistische 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;
- Datensätze für ein Merkmal 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

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 statistische Methodenkompetenz eine in immer mehr Arbeitsfeldern benötigte Schlüsselqualifikation darstellt, die auch im privaten Bereich nützlich ist.

Vorschau auf das Kapitel

1.1 Statistik im Alltag, in Politik und Gesellschaft

Die **Statistik** ist eine noch junge 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, die auf eine vergleichende Bewertung nationaler Bildungssysteme abzielen. 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, zur Entwicklung der Erwerbstätigkeit in Deutschland oder zur Nutzung sozialer Netzwerke. 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** oder den *Public Data Explorer* von Google.

Statistische Daten im Alltag

Die überragende gesellschaftliche Relevanz der Statistik spiegelt sich auch darin wider, dass das Jahr 2013 zum Internationalen Jahr der Statistik ausgerufen wurde. Im Laufe des Jahres gab es zu diesem Anlass zahlreiche Veranstaltungen von Universitäten, Statistikämtern, Unternehmen oder internationalen Institutionen.

2013 – Internationales Jahr der Statistik

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_1

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 Antworten zur „Sonntagsfrage“ vom 16. Oktober 2009 werden in diesem Manuskript mehrfach zur Illustration der Anwendung statistischer Konzepte herangezogen, etwa bei der Vorstellung grafischer Instrumente der beschreibenden Statistik oder 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. Rückrufaktionen und Gewährleistungsprozesse können 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, um Auffälligkeiten zu identifizieren, etwa die aktuellen „Renner“ und Ladenhüter. 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 psychologischer Interventionen zu evaluieren, z. B. die Effekte psychotherapeutischer Maßnahmen.

Statistik ist fachübergreifend

Die Statistik erfüllt für viele Wissenschaften eine wichtige Servicefunktion. In der *Soziologie*, der *Psychologie* oder auch der *Medizin* stützen sich Publikationen in den Fachzeitschriften maßgeblich auf Daten und deren statistischer Analyse. Die Versuchsplanung, bei der es 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, radioaktiver Zerfall), 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* oder die Gewinnung empirisch fundierter Aussagen zum Zustand von Wäldern in den *Umweltwissenschaften*.

Video von SAS zum Jahr der Statistik 2013

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

Statistik in der Politik

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). Auch der deutsche Beitrag zum Europäischen Stabilisierungsmechanismus (ESM), der die Stabilität des Euro sichern soll, hängt von Bevölkerungsdaten ab.

Beispiel 1.2: Monitoring strategischer Ziele der amtlichen Statistik

Interaktives Lernobjekt „Erwerbstätigkeit“

Interaktives Lernobjekt „Emission von Treibhausgasen“

Im Jahr 2000 verständigten sich die Staats- und Regierungschefs der Länder der EU auf die sogenannte *Lissabon-Strategie*, die Entwicklungsziele für Europa bis 2010 definierte. Die Ziele wurden aber nur teilweise erreicht. Inzwischen ist eine mit dem Etikett *Europa 2020* versehene neue Strategie vereinbart, die wirtschaftliche und soziale Kernziele für Europa bis 2020 festlegt und anhand von acht Leitindikatoren operationalisiert. Ein Ziel ist z. B. die Erhöhung der Beschäftigungsquote der als erwerbsfähig geltenden EU-Bevölkerung auf 75 %. Der Erreichungsgrad dieses Ziels wird von Eurostat über den Indikator „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 um 20 % gegenüber dem Stand von 1990.

Auch die *Vereinten Nationen* (UN) verfolgen globale Strategien und verknüpfen diese mit Indikatoren. Die UN wollen u. a. extreme Armut bekämpfen und gaben sich hier für den Zeitraum von 2000 bis 2015 Ziele vor. Inwieweit die Zielerreichung glückte, wurde anhand von acht Indikatoren, den *UN Millennium Development Goals*, in größeren Abständen quantifiziert.[1] Daten spielten hier ebenfalls eine Schlüsselrolle für das Politikmonitoring. Aggregate aus verschiedenen Indikatoren, sog. zusammengesetzte Indikatoren, werden vermehrt von internationalen Organisationen zur Beschreibung komplexer Entwicklungen eingesetzt, etwa zur Messung von Wohlfahrt oder von Innovation.

Exkurs 1.1: Datenjournalismus und Statistik-Blogs

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, die etwa 2009 begann, 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* oder auch bei der *New York Times* 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. Beispiele finden

[1] Die Bilanz Anfang 2015, kurz vor Ende der 5-Jahres-Periode, fiel gemischt aus. Bei einigen Millenniumszielen, z. B. der angestrebten Reduzierung der Mortalitätsrate bei Kindern unter 5 Jahren auf ein Drittel des Stands von 2000, wurden Erfolge erzielt, vor allem durch Verbesserung des Impfschutzes. Bei anderen Zielen, etwa der geplanten Halbierung des Anteils der unter extremer Armut lebenden Menschen, besteht weiterhin großer Handlungsbedarf.

sich in Exkurs 1.2 (Wohlstandsvergleich zwischen OECD-Ländern) und auch Exkurs 4.1 (Visualisierung von Altersstrukturen).

Es gibt inzwischen viele Blogs zur Statistik, die nicht nur für Experten interessant sind. Beispielhaft genannt seien hier der *Stats Blog* und der *Blog about Stats*. Der erstgenannte Blog ist eine Plattform für den allgemeinen Informationsaustausch zwischen Statistikern mit Querverbindungen zu zahlreichen spezialisierten Statistik-Foren. Der letztgenannte Blog widmet sich vor allem neueren Entwicklungen im Bereich der Kommunikation amtlicher Daten.

1.2 Aufgaben und Teilbereiche der Statistik

Aufgaben 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 von Daten und zum Lernen aus Daten bereit stellt.

Öffentliche Wahrnehmung des Fachs

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.

Mehrdeutigkeit des Begriffs „Statistik“

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), sowie für *Funktionen von Zufallsvariablen*, die zur Schätzung dieser Kenngrößen herangezogen werden. Im allgemeinen Sprachgebrauch wird auch häufig ein *Datensatz* als eine Statistik angesprochen, etwa ein Datensatz mit der Medaillenverteilung bei den Olympischen Sommerspielen oder Daten zu Bruttoverdiensten in der Europäischen Union. In diesem Manuskript wird „Statistik“ überwiegend im Sinne von „Wissenschaft“ verwendet.

Teilbereiche der Statistik:

Innerhalb der Statistik lassen sich idealtypisch 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 Aggregation auf wenige Kenngrößen, möglichst ohne größeren Informationsverlust. Das Europäische Amt für Statistik sammelt

Beschreibende Statistik

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.[2] Jede empirisch arbeitende Wissenschaft argumentiert mit Daten und bedient sich zwangsläufig der Instrumente der beschreibenden Statistik. Typisch für die beschreibende Statistik ist, dass sie keine Modelle benötigt. Letztere sind das Ergebnis von Bemühungen, reale Beobachtungen auf Gesetzmäßigkeiten zurückzuführen und diese zu formalisieren.

Explorative Datenanalyse und „Big Data“

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 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. 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, unstrukturierte oder sehr heterogene Datenbestände, die auch Unschärfen aufweisen können (etwa Daten von Biosensoren und Überwachungskameras, zum Kaufverhalten im Internet, zu Aktivitäten in sozialen Netzwerken oder zur Nutzung mobiler Endgeräte). Diese Daten werden computergestützt in Echtzeit mit mathematisch-statistischen Verfahren, u. a. mit Methoden des Data Mining, auf Zusammenhänge untersucht. Die Datenanalyse ist hier nicht hypothesengetrieben; vielmehr werden Hypothesen erst im Zuge der Analyse generiert.

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 und durch Verteilungsmodelle (engl: *probability distributions*) beschrieben. 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.

[2] In 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.

1.3 Methodenkompetenz als Lernziel

Seit den 90er Jahren wird über Schlüsselqualifikationen und Kompetenzen diskutiert, die Menschen dazu befähigen, den sich wandelnden Anforderungen des Berufs und, allgemeiner, 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 bloße Aneignung von Wissensinhalten hinaus. Es gibt unterschiedliche Arten von Schlüsselqualifikationen, etwa *soziale Kompetenz* (umfasst Kommunikationsfähigkeit im zwischenmenschlichen Bereich), *Medienkompetenz* (Fähigkeit zur effizienten Nutzung der kaum noch überschaubaren Informationsfülle) und *Methodenkompetenz* (Fähigkeit zur sachäquadaten Nutzung unterschiedlicher Werkzeuge, Arbeitstechniken und Theorien zur Lösung von Problemen). Mit dem **Bologna-Prozess**, der im Sommer 1999 mit einer gemeinsamen Erklärung der Europäischen Bildungsminister zur Schaffung eines europäischen Hochschulraums in Gang kam, wurde die *Beschäftigungsfähigkeit* (engl: *employability*) als neue Schlüsselqualifikation betont. Sie soll Hochschulabsolventen europaweit dazu befähigen, eine Beschäftigung auf dynamischen Arbeitsmärkten zu finden und das eigene Qualifikationsprofil fortlaufend den veränderten Gegebenheiten durch lebenslange Weiterbildung anzupassen.

Schlüsselqualifikationen und der Bologna-Prozess

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 *kommunalen Planungen* stützt man Entscheidungen über Investitionen auf Bevölkerungsdatendaten, 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*.

Methodenkompetenz als Basis für datengestützte Entscheidungsfindung

Allen genannten Beispielen ist gemeinsam, dass hier Wissen über statistische Methoden benötigt wird, um Daten zu gewinnen, auszuwerten und aus den Ergebnissen statistischer Analysen sachadäquate Schlüsse zu ziehen. Benötigt wird aber auch die Fähigkeit zur klaren und nachvollziehbaren Ergebniskommunikation. Das vorliegende Manuskript soll diese als **statistische Methodenkompetenz** (engl: *statistical literacy*) bezeichnete Qualifikation vermitteln. Da statistische Methodenkompetenz in immer mehr Berufsfeldern an Bedeutung gewinnt, hat sie einen positiven Einfluss auf die Schlüsselqualifikation „Beschäftigungsfähigkeit“.

Internationale Projekte zur Förderung von Methodenkompetenz

Der Stellenwert, den das Thema „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 auch Projekte auf nationaler und internationaler Ebene, die zur Verbesserung 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.[3]

Mit den vorstehend genannten Projekten und Aktivitäten hat dieses Manuskript eines gemeinsam – auch diese Einführung zielt auf die Entwicklung statistischer Methodenkompetenz ab, allerdings, anders als die schulbezogenen Projekte, auf der Ebene einer universitären Aus- und Weiterbildung. Es werden zwei übergeordnete Ziele angestrebt. Auf der ersten Stufe soll eine umfassende *Kenntnis* alternativer Möglichkeiten der Auswertung und Präsentation statistischer Information und die Fähigkeit zu einer sachadäquaten Ergebnisinterpretion erreicht werden. Man spricht in diesem Kontext von einer *passiven Methodenkompetenz*. Diese immunisiert z. B. vor manipulativem Umgang mit Daten in den Medien. Weitergehend ist das auf der nächsten Stufe angestrebte Ziel der Vermittlung *aktiver Methodenkompetenz*. Letztere ist eine Handlungskompetenz, die sich auf die Fähigkeit bezieht, im beruflichen Alltag Entscheidungen empirisch zu fundieren und nachvollziehbar zu kommunizieren.

Passive und aktive Methodenkompetenz

Beispiel 1.3: Fehlalarmhäufigkeiten bei der Krebsfrüherkennung

Die *Süddeutsche Zeitung* berichtete in der Ausgabe vom 9. März 2015 kritisch über den Nutzen von Untersuchungen zur Früherkennung von Brustkrebs. Ähnliche Artikel, auch zu Fehlalarmhäufigkeiten bei Prostatavorsorgeuntersuchungen, gab es schon am 20. April 2009 bei *Spiegel online* und in der

[3] Die CAUSE-Bibliotheken sind unter http://www.causeweb.org/ und die SOCR-Materialien unter http://www.socr.ucla.edu/ frei zugänglich.

Wochenzeitung *Der Freitag* vom 3. Dezember 2012. Alle Beiträge illustrieren die Bedeutung statistischer Methodenkompetenz auch im privaten Bereich.

In den Artikeln wird u. a. bemängelt, dass nicht nur Patienten, sondern auch Mediziner oft falsche Vorstellungen von der Treffsicherheit medizinischer Testverfahren haben und häufig mit falsch-positiven Befunden konfrontiert werden. Oft werde zudem mit relativen Risiken argumentiert, ohne dass die Bezugsbasis deutlich wird. Wenn sich z. B. mit Einsatz neuer Medikamente die Mortalitätsrate bei einer Erkrankung um 10 % vermindert, kann dies sowohl bedeuten, dass von 10 000 Personen im Mittel statt 10 Personen nur noch 9 der Erkrankung erliegen oder auch, dass sich die absolute Sterbeziffer von 3 000 auf 2 700 verringert. In einem Beitrag vom 25. Februar 2010 in der Zeitschrift *Die Zeit* wird auf Schwächen des Mammografie-Screenings hingewiesen, die auf mangelnder Koordination zwischen den Bundesländern beruhen oder auf Unterschieden bei den Qualitätstandards in spezialisierten Zentren und außerhalb solcher Referenzzentren zurückzuführen sind. Dass meist nur der potenzielle Nutzen und weniger die möglichen Risiken von Massenscreening-Aktionen kommuniziert werden, ist bemerkenswert, weil hier erhebliche finanzielle Ressourcen des Gesundheitsbereichs einfließen. Selbst Mediziner sind oft nicht in der Lage, betroffene Patienten sachadäquat über Screening-Risiken oder über Risiken von Medikamenten zu informieren. Letzteres erklärt sich z. T. daraus, dass Ärzte ihre Information oft aus nicht-neutralen Quellen beziehen.

Gigerenzer (2009) präsentiert ein Beispiel, das die Probleme beim Verständnis und bei der Kommunikation medizinischer Risiken illustriert. Bei dem Beispiel wird vorausgesetzt, dass in einer größeren Grundgesamtheit von N Frauen einer definierten Altersklasse 0,8 % der Frauen Brustkrebs haben und der Krebs in 90 % der Fälle bei einer Mammographie entdeckt wird. Allerdings weiß man auch, dass in der Teilpopulation ohne Erkrankung beim Screening im Mittel in 7 % aller Fälle ein Fehlalarm erfolgt.

Es wird nun eine Frau zufällig aus der Gesamtpopulation ausgewählt, die zu einem Screening geht und einen positiven Befund erhält. Wie groß ist die Wahrscheinlichkeit, dass sie trotzdem gesund ist, also ein falsch-positiver Befund vorliegt? [4] Hier ist nur Ihre intuitive Einschätzung gefragt. Würden Sie die Wahrscheinlichkeit eines falsch-positiven Befunds mit 1 %, 10 %, 30 %, 50 %, 70 %, 90 % oder gar mit 99 % beziffern? [5]

Exkurs 1.2: Durchführung und Rezeption der PISA-Studien

Die seit 2000 in 3-jährigem Turnus und inzwischen in fast 70 Ländern laufenden **PISA-Studien** (PISA = *P*rogramme for *I*nternational *S*tudent *A*ssessment) 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

[4] Unter „Wahrscheinlichkeit“ ist hier der Anteil der falsch-positiven Befunde (Fehlalarme) an der Gesamtzahl *aller* positiven Befunde zu verstehen.
[5] Die Frage wird in Kapitel 10 geklärt (s. dort die Aufgaben 10.6 und 10.7).

Wohlstandsvergleich

heutigen Wissensgesellschaft gerüstet sind. Operationalisiert wird diese Fähigkeit („literacy“) 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, 2003 und 2012 lag der Fokus 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. Bei der PISA-Studie 2009 zur Lesekompetenz bezogen sich die Zusatzfragen auf die Lebenssituation der Familien (Wohlstand, Zugang zu Wissen, Familienstruktur). Die Ergebnisse wurden ohne Datumsangabe bei *Zeit online* für die beteiligten Länder interaktiv präsentiert.

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 Züricher 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 7.1 und 7.2).

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. Die Berichterstattung über die im internationalen Vergleich gegenüber 2003 angeblich deutlich verbesserten Mathematikleistungen deutscher Schüler wurde in einer vom Rheinisch-Westfälischen Institut für Wirtschaftsforschung herausgegebenen *Pressemitteilung* vom 17. Dezember 2013 zur „Unstatistik des Monats“ erklärt. Der bessere Rangplatz Deutschlands sei, so der Tenor der Mitteilung, möglicherweise allein oder überwiegend auf die veränderte Zusammensetzung der Stichprobe zurückzuführen, nicht aber notwendigerweise auf eine Verbesserung des Mathematikunterrichts.

1.4 Medienmix 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 *Coursera*, *EdX*, *Udacity* und der europäischen Plattform *iversity*. 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.: *M*assive *O*pen *O*nline *C*ourses) 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 Artikel 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.

Granulare Online-Ressourcen

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* (*C*onsortium for the *A*dvancement of *U*ndergraduate *S*tatistics *E*ducation).

Online-Seminare

Neben den MOOCs haben auch sog. **Webinare** in der Methodenausbildung ihren Platz gefunden. Der Name leitet sich aus *Web* und Sem*inar* 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.

Einbezug mobiler Endgeräte

Online-Kurse und kleinteiligere 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.

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 einer Web-App zusammengefasst, die sich auch für den Mathematikunterricht der Sekundarstufe II eignet (s. Mittag 2015).

Statistik-Web-App

Abb. 1.2: *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.

2 Grundbegriffe der Statistik

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.

Vorschau auf das Kapitel

2.1 Statistische Einheit, Merkmal und Grundgesamtheit

Wozu braucht man eine statistische Terminologie?

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.

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.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_2

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 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, wie man die Werte für unterschiedliche Schadstoffe zu einem einzigen Wert aggregiert und 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 delinquentem Verhalten bei Jugendlichen oder nach Fahrzeugtyp bei Untersuchungen zu Schadstoffemissionen im Straßenverkehr.

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*.

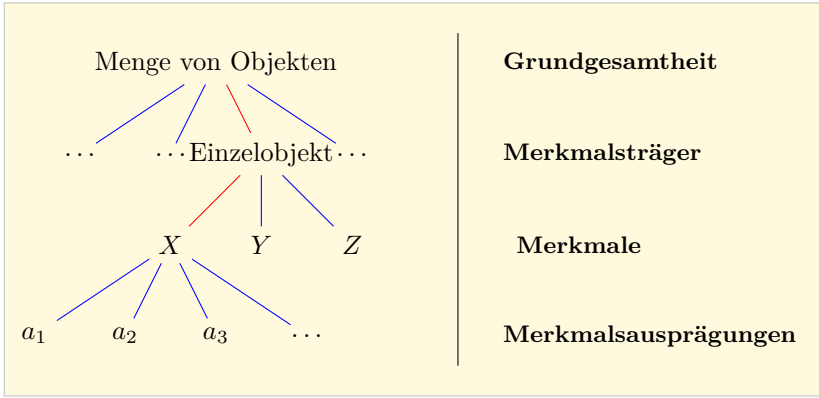


Abb. 2.1: *Begriffshierarchien für statistische Grundbegriffe*

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 **Urwerte**, **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 Januar 2015 in Sierra Leone neu gemeldeten Fälle der Viruserkrankung Ebola. 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 2013 in München ihren Erstwohnsitz hatten;
- Studierende einer Hochschule zu Beginn des Sommersemesters 2012, über die man via Telefonbefragung Informationen gewinnen will;
- die von einem stahlverarbeitenden Unternehmen im März 2012 produzierten Serienteile eines bestimmten Typs;
- die in Deutschland am 1. Januar 2015 gemeldeten PKWs.

Die statistischen Einheiten werden hier repräsentiert durch

- jede Person mit Erstwohnsitz in München am 1. Mai 2013;
- alle zum Sommersemester 2012 eingeschriebenen Studierenden;
- die im März 2012 gefertigten Serienteile;
- jeder am 1. Januar 2015 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.

Aufgabe 2.1

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.

2.2 Merkmalsklassifikationen und Skalen

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

Einteilung von Merkmalen nach der Anzahl der Ausprägungen

Ein **diskretes Merkmal** ist ein Merkmal, das nur endlich viele Ausprägungen oder aber 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 Merkmalen 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

[1] Der 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.

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 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 Merkmalen 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.

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

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.

Skala		sinnvolle Operationen			
		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*

Tabelle 2.1 macht deutlich, dass die genannten Skalenniveaus eine Hierarchie darstellen, bei der die Nominalskala das niedrigste Niveau und die Verhältnisskala – bzw. die Absolutskala als Sonderfall der Verhältnisskala – das höchste Niveau repräsentiert. Operationen, die für Daten eines bestimmten Skalenniveaus zulässig sind, sind stets auch auf Daten aller höheren Niveaus anwendbar. Man kann nämlich ein Merkmal, das ordinalskaliert ist, auf einer Nominalskala messen und ein metrisch skaliertes Merkmal stets auch auf einer Ordinalskala oder Nominalskala – allerdings auch hier wieder 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: Datenerhebungsinstrument 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 Zustimmung 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 ordinalskaliert, 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

Aufgabe 2.2

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 „CO_2-Emissionen von PKWs“ (Verhältnisskala).

Einteilung von Merkmalen 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 Falle 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 (2013, Abschnitt 3.5).

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 Rechenfertigkeiten 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? Die Europäisierung der amtlichen Statistik wirkt der möglichen Manipulation durch Veränderung der Operationalisierung von Merkmalen entgegen. Eurostat besitzt Vollmachten für das Monitoring von Daten, die für die Stabilität der Eurozone besonders relevant sind – etwa Daten zur Entwicklung von Staatsschulden.

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 Sozialtransferzahlungen, 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 Lernobjekt „Erwerbstätigkeit“

Politisch brisanter ist die Operationalisierung von Erwerbs- oder Arbeitslosigkeit. Als *erwerbslos* gilt nach der z. Z. angewandten Definition der **International Labour Organization** (**ILO**, 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 **Bundesagentur für Arbeit** (**BA** in Nürnberg) 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

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.

Vorschau auf das Kapitel

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.

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.

Klassifikation von Erhebungen hinsichtlich der Datenquelle

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. Man unterscheidet entsprechend zwischen primär- und sekundärstatistischen Daten. Gelegentlich spricht man auch von **Tertiärerhebungen**, nämlich dann, wenn statistische Information aus vorhandenen Quellen geschöpft wird, aber nicht in Form der Originaldaten, sondern in aggregierter Form (z. B. gruppierte Daten).

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_3

Beispiel 3.1: Primär-, Sekundär- und Tertiärerhebungen

Die regelmäßig erscheinenden Berichte des Münchner 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, verwenden hingegen sekundärstatistische Daten und sind somit *Sekundärerhebungen*.

Die *Europäische Gehalts- und Lohnstrukturerhebung* erfasst Individualdaten für Millionen von Arbeitnehmern in fast allen europäischen Staaten. Die Verwendung der amtlichen Ergebnisse beinhaltet die Nutzung *tertiärstatistischer Daten*, weil die Ergebnisse der Erhebung aufgrund der Vertraulichkeit der originären Mikrodaten nur in aggregierter Form von Eurostat und den an der Erhebung beteiligten nationalen Statistikämtern kommuniziert werden. Die Vertraulichkeit der Mikrodaten ist durch Verordnungen mit Gesetzeskraft geregelt. Die Aussagekraft statistischer Auswertungen, die auf Tertiärdaten beruhen, ist natürlich reduziert, weil die ursprünglich vorhandene statistische Information verkürzt wird. Wissenschaftler sind an den Mikrodaten der Europäischen Wirtschafts- und Sozialstatistik interessiert, deren Vertraulichkeit – möglichst ohne nennenswerten Informationsverlust – durch geeignete Anonymisierungsverfahren zu sichern ist.

Klassifikation von Primärerhebungen nach der Art der Datengewinnung

Für die Forschung in den *Sozialwissenschaften*, der *Psychologie* und auch in der *Medizin* sind Primärerhebungen von besonderer Bedeutung. Man kann hier hinsichtlich der Art der Datengewinnung zwischen einer Befragung, einer Beobachtungsstudie und einem Experiment unterscheiden. Alle genannten Erhebungstypen können sich sowohl auf Einzelpersonen als auch auf Personengruppen beziehen.

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 geschlossenen Fragen erreichen.

Varianten der Befragung

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** (*c*omputer *a*ssisted

*p*ersonal *i*nterviewing) für das persönlich geführte Interview mit tragbarem Computer (meist Notebook) und **CATI** (*c*omputer *a*ssisted *t*elephone *i*nterviewing) 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 3.1 zeigt ein Telefonstudio eines Marktforschungsinstituts, in dem fernmündliche Befragungen durchgeführt werden.

Flash-Animation „Befragungen“

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.

Welche Befragungsart ist zu wählen?

Für welche Form einer Befragung man sich bei der Planung einer Erhebung entscheidet, hängt u. a. von der Größe des zu gewinnenden Datensatzes, von der Zielgruppe sowie vom verfügbaren Untersuchungsbudget und Zeitrahmen ab. Computerunterstützte Varianten werden jedenfalls immer wichtiger, z. B. etwa in der Markt- und Meinungsforschung.

Auch die **Beobachtung** ist ein verbreitetes Verfahren der Datenerhebung. Beobachtung kann sich auf ganz unterschiedliche Objekte beziehen,

[1] Der *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute* (ADM) gibt auf seiner Internetseite unter dem Link „Marktforschung in Zahlen“ an, dass 2014 in den Mitgliedsinstituten des ADM ca. 3900 CATI-Plätze eingerichtet und mehr als 8400 CAPI-Geräte im Einsatz waren (Stand: Mai 2015).

Wo werden Daten per Beobachtung gewonnen?

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, z. B. 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. Abbildung 3.2 illustriert dies anhand eines Fotos aus der Fertigungsüberwachung. Bei diesem Beispiel werden die Ausprägungen zweier qualitätsrelevanter geometrischer Merkmale (Durchmesser von Kurbelwellen) gleichzeitig erfasst und die Ergebnisse fortlaufend als Zeitreihe gespeichert. Auch hier können intervenierende Maßnahmen zum Zuge kommen, etwa das Nachjustieren einer Fertigungseinrichtung.

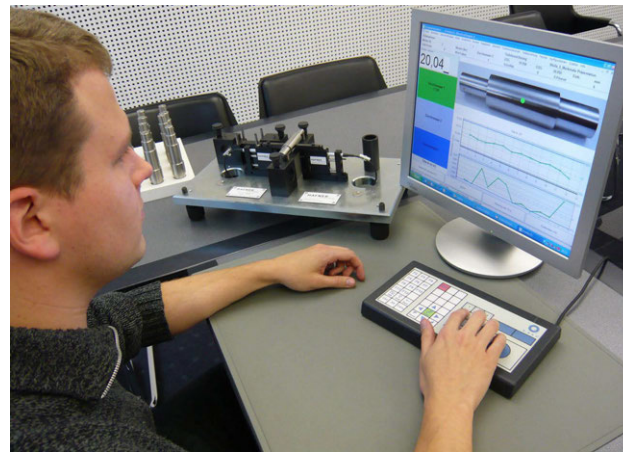


Abb. 3.2: *Simultane Erfassung zweier Merkmale eines Serienteils bei der industriellen Fertigungsüberwachung; Quelle: Fa. Q-DAS*

In den *Umweltwissenschaften* werden z. B. Schadstoffemissionen fortlaufend gemessen. Die Daten werden bei internationalen Klimakonferenzen als Diskussionsbasis herangezogen. 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.

Neuere Entwicklungen in der Markt- und Konsumforschung

In der Markt- und Konsumforschung gewinnen nicht-reaktive Methoden der Datengewinnung an Bedeutung, bei denen modernste Technik genutzt wird. *Google Analytics* ist z. B. ein Informationsdienst, der Verhaltensspuren im Internet auswertet. Als weiteres Beispiel genannt seien Frequenzzählungen oder Aufzeichnungen von Blickbewegungen und Blickwinkeln von Kunden in den Gängen von Supermärkten, mit denen Unternehmen Informationen zur Optimierung des Warensortiments gewinnen. Auch Geoinformationssysteme werden zunehmend zur Identifikation raumbezogener Zusammenhänge herangezogen, etwa bei der Messung von Pendlerströmen. Im Leipziger Zoo lief ein Projekt, bei dem GPS-Daten zur Gewinnung von Informationen zur Verweildauer von Zoobesuchern bei den einzelnen Tiergehegen genutzt wurden.

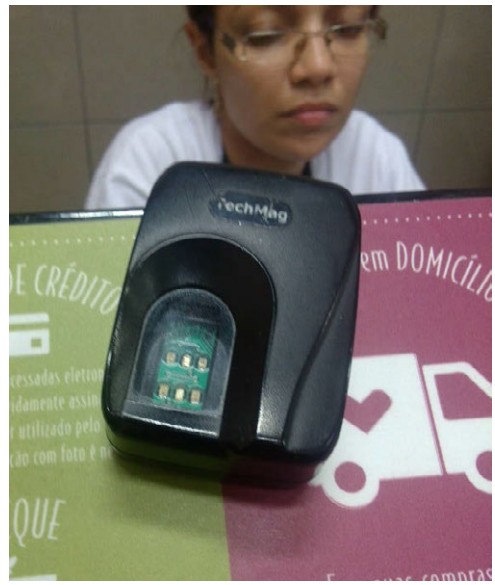


Abb. 3.3: *Überprüfung von Kundendaten per Fingerabdruck (Quelle: J. McNeill, Kennesaw State University, USA)*

Beim Bezahlen eines Einkaufs an der Kasse mit einer Bankkarte werden mancherorts zur Verifikation der auf dem Magnetstreifen gespeicherten Daten noch biometrische Daten herangezogen. Auf diese Weise wird Betrug mit gestohlenen Bankkarten erschwert. Abbildung 3.3 zeigt ein Lesegerät für Fingerabdrücke in einem brasilianischen Supermarkt

mit Lieferservice. Die anhand des Geräts gewonnenen Daten werden mit den Informationen auf dem Magnetstreifen abgeglichen. In etlichen US-amerikanischen Schulkantinen können Schüler an der Kasse allein mit ihrem Fingerabdruck bezahlen. In Deutschland ist eine Personenidentifikation anhand des elektronischen Reisepasses möglich, der ein digitalisiertes biometrisches Passbild und – optional – auch Fingerabdrücke speichert. Die für einen Abgleich benötigten Referenzdaten werden von den Meldebehörden erfasst.

Anwendungsfelder für Experimente

Ein ganz anderer Ansatz zur Datengewinnung ist der Einsatz von **Experimenten**. Diese wurden zuerst in den *Agrar-* und *Naturwissenschaften* und später in der *Technik* angewendet, sind 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 eine geeignete Untersuchungsanordnung den Einfluss weiterer Variablen möglichst ausschalten (Kontrolle von **Störvariablen**). Die Untersuchungsanordnung wird durch einen **Versuchsplan** festgelegt.

Flash-Animation „Versuchsplanung“

In den Wirtschafts- und Sozialwissenschaften sind Experimente kaum verbreitet, weil sich Forschung hier typischerweise auf Feldbeobachtung bezieht und selten auf Laborsituationen übertragen lässt. Bei den in der wirtschafts- und sozialwissenschaftlichen Forschung verbreiteten Beobachtungsstudien lässt ein beobachteter Zusammenhang zwischen zwei Merkmalen – anders als beim Experiment – nicht zwingend auf einen Kausalzusammenhang schließen, weil der Zusammenhang auch über eine dritte Variable vermittelt sein kann (vgl. hierzu Exkurs 9.2).

Fachspezifische Unterschiede beim Design von Experimenten

Zwischen Experimenten in den einzelnen Anwendungsbereichen, etwa in der Technik oder in der Psychologie und 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, z. B. bei Werkstoffen den Zusammenhang zwischen der Zusammensetzung von Legierungen und der Werkstoffeigenschaft „Reißfähigkeit“ oder „Härte“. Auch die planmäßige Veränderung von Formparametern eines Kraftfahrzeugs und die Untersuchung des Effekts auf den Luftwiderstand (c_w-Wert) im Windkanal ist ein Beispiel für eine industrielle Anwendung von Versuchsplänen. Die Messung der Merkmalsausprägungen anhand moderner Messtechniken ist hier i. d. R. kein Problem und auch Messwiederholungen lassen sich leicht realisieren.

Experimente in der *Psychologie* beziehen sich hingegen auf individuelle Merkmale von Personen, etwa auf die Ausprägungen der latenten Varia-

blen „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**.

Eine ausführlichere Darstellung der vorgestellten Grundtypen „Befragung“, „Beobachtung“ und „Experiment“ findet man bei Sedlmeier / Renkewitz (2013, Kapitel 4 - 5). Die beiden erstgenannten Typen einschließlich einer Würdigung ihrer Vor- und Nachteile sind bei Schnell / Hill / Esser (2011, Abschnitte 7.1 - 7.2) detailliert behandelt. Der Gestaltung von Fragebögen, der Schulung von Beobachtern und Interviewern sowie dem Design von Experimenten widmet sich eine kaum zu überschauende Flut von Veröffentlichungen.

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 mit einem Notebook in die Haushalte und geben die Befragungsergebnisse sofort in mitgebrachte Notebooks ein (Datenerhebung via CAPI). Die Interviewsteuerung einschließlich der Prüfung der Antwortenkonsistenz 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

- nach dem zeitlichen Zusammenhang der Daten

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. 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 längeren 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.

- nach dem Umfang der erhobenen Daten

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 Verwaltungsregister zurückgriff (*registergestützter Zensus*), vor allem auf Melderegister und Register der Bundesagentur für Arbeit. In Deutschland beruhten die amtlichen Bevölkerungszahlen bis 2011 noch auf Fortschreibungen der Volkszählung von 1987 bzw. 1981 anhand der Mikrozensusdaten. Inzwischen waren die Hochrechnungen aber sehr unzuverlässig geworden und man vermutete, dass die für Ende 2011 angenommene amtliche Bevölkerungszahl für Deutschland zu hoch lag. Die seit Mai 2013 vorliegenden Ergebnisse zum Zensus von 2011 bestätigten dies. Danach lag die Einwohnerzahl Deutschlands Ende 2011 bei 80,3 Millionen und damit 1,5 Millionen niedriger als bisher angenommen. Zuverlässige Bevölkerungsdaten sind aber für viele

Flash-Animation „Mikrozensus und Zensus 2011“

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 und zum Euro-Rettungsschirm.

Vorteile und Grenzen von Stichprobenerhebungen

Stichprobenerhebungen sind vor allem bei sehr großen Grundgesamtheiten geboten oder oft auch der einzig 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 oder Einkommen 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. In solchen Fällen gibt es zur Stichprobenprüfung keine Alternative. Bei der Prüfung sicherheitsrelevanter Produkte, etwa bei Airbags oder Reißleinen von Fallschirmen, sind hingegen Vollerhebungen geboten, weil hier Restrisiken nicht vertretbar sind.

Aufgabe 3.1

Beispiel 3.3: SOEP und ALLBUS

Das *Sozioökonomische Panel* (*SOEP*) ist eine seit 1984 durchgeführte stichprobenbasierte Befragung von über 12 000 Haushalten (gleichbleibende Haushalte), die auf die Identifikation politischer und gesellschaftlicher Veränderungen in Deutschland abzielt. Die Befragung bezieht sich auf alle erwachsenen Haushaltsmitglieder 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 Zweijahresturnus 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 *Ge*sellschaft *S*ozialwissenschaftlicher *I*nfra*s*trukture*i*nrichtungen durchgeführt, die seit 2008 den Namen *GESIS – Leibniz-Institut für Sozialwissenschaften* trägt. Auch hier werden die Ergebnisse der Fachöffentlichkeit zugänglich gemacht.

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 z. B. 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.

Flash-Animation „Over- und Undercoverage“

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. Die dreiteilige Abbildung 3.4 veranschaulicht beide Phänomene. Die eigentlich angestrebte Grundgesamtheit ist jeweils anhand eines durchgezogenen, die tatsächlich bei der Stichprobenziehung verwendete Auswahlgesamtheit anhand eines gestrichelten Rahmens dargestellt.

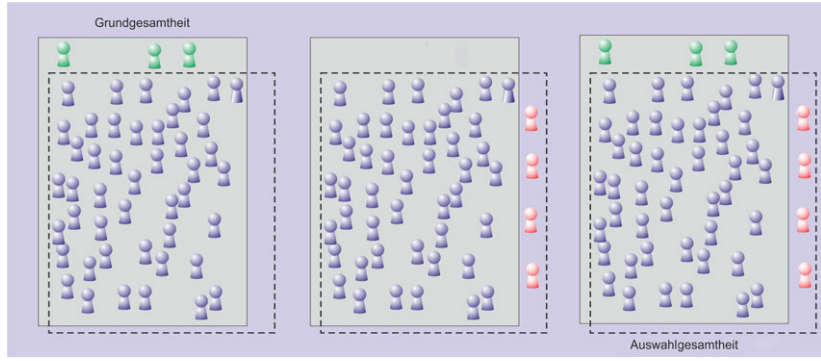


Abb. 3.4: *Undercoverage (linke Teilgrafik), Overcoverage (Mitte), gleichzeitiges Auftreten von Under- und Overcoverage (rechts)*

Warum zufällige Auswahl?

Um mit der Stichprobe ein repräsentives 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 Teilinformation der Merkmalsträger der Stichprobe und nicht die volle Information aller Merkmalsträger der Grundgesamtheit zur Verfügung steht. Man spricht in diesem Zusammenhang von einem **Stichprobenfehler**.

Flash-Animation „Inferenzschluss“

Wenn man einen Inferenzschluss auf eine Stichprobe stützt, die nicht repräsentativ ist für eine Grundgesamtheit (verzerrte Stichprobe), kommt zu dem unvermeidlichen Stichprobenfehler noch eine durch die **Verzerrung** (engl.: *bias*) der Stichprobe bedingte *systematische Verzerrung* hinzu, der sog. **Auswahlbias**. Der Inferenzschluss kann dann zu gravierenden Fehlschlüssen führen. Würde man z. B. in Finnland anhand eines Verzeichnisses stationärer Telefonanschlüsse eine Stichprobe auswählen, hieße dies, von vorneherein einen erheblichen Teil der Bevölkerung auszuschließen. Schon 2006 hatte nämlich bereits ca. 40 % der finnischen Bevölkerung nur noch ein Mobiltelefon. Vor allem der jüngere Teil der Bevölkerung wäre in der Stichprobe stark unterrepräsentiert.

Flash-Animation „Verzerrte Stichprobe“

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 neuerdings auch auf mobilen Endgeräten. Als Datenbasis dient eine Stichprobe von ca. 5000 Haushalten (AGF-Fernsehpanel), von der auf ca. 37 Millionen Fernsehhaushalte in Deutschland zurückgeschlossen wird.

Für die Gewinnung von Daten zur *stationären* TV-Nutzung wird von der GfK seit 2012 ein sog. Audio-Matching-Verfahren angewendet. Die Audiosignale der abzubildenden Sender werden in Form digitaler Signaturen aufgezeichnet und zentral gespeichert. Bei den Panelhaushalten werden solche „akustischen Fingerabdrücke“ anhand eines Messgeräts gewonnen, das die GfK den Haushalten zur Verfügung stellt. Abbildung 3.5 zeigt ein derartiges Messgerät im Einsatz.

Durch einen Abgleich („Matching“) 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, können die Panelhaushalte mit Netzwerk-Routern und On-Device-Software ausgestattet werden. Diese filtern bestimmte Markierungen („Tags“) aus dem Datenstrom und ermöglichen die Identifikation von Nutzungsart und Nutzungsdauer.

Abb. 3.5: *Gewinnung von Daten zur Nutzungsdauer von TV-Sendern (Quelle: GfK Nürnberg)*

Einschaltquoten vom Vortag

Für jede Sendung lässt sich so am Ende eines Fernsehtages eine Einschaltquote ermitteln, die aufgrund der unterstellten Repräsentativität der Panelhaushalte auf die Grundgesamtheit aller Fernsehhaushalte bezogen wird. Die per Hochrechnung ermittelten Einschaltquoten werden jeden Morgen veröffentlicht. Sie beeinflussen die Preise für Fernsehwerbung.

Zur Sicherung der Repräsentativität des Fernsehpanels werden Gewichtungen vorgenommen, z. B. zum Ausgleich unterschiedlicher regionaler Verteilungen der Haushalte. Die Zusammensetzung des Panels und die verwendete Gewichtung werden regelmäßig überprüft. So soll gewährleistet werden, dass das Panel bei einer sich wandelnden Struktur der Gesamtbevölkerung die Veränderungen möglichst gut nachvollzieht.

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 in unzureichendem Maße 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 Zuge oder nacheinander n Elemente gezogen werden. Die Ziehung der Lottozahlen ist z. B. so organisiert.

Zweistufige Verfahren

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.

Aufgabe 3.2

Formal ist eine geschichtete Stichprobenauswahl 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, ...,N_L zerlegt wird ($N_1+N_2+..+N_L = N$), aus denen im zweiten Schritt Zufallsstichproben des Umfangs n_1, n_2, ...,n_L gezogen werden ($n_1 + n_2 + .. + n_L = n$). Je nachdem, ob der Anteil $\frac{n_i}{N_i}$ ($i = 1, 2, .., L$) der einer Schicht entnommenen Stichprobenelemente fest ist oder nicht, liegt eine *proportional geschichtete Stichprobe* resp. eine *disproportional geschichtete Stichprobe* vor. Abbildung 3.6 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 Auswahlsatz 20 % der Elemente einer Schicht.

Java-Applet „Geschichtete Stichproben“

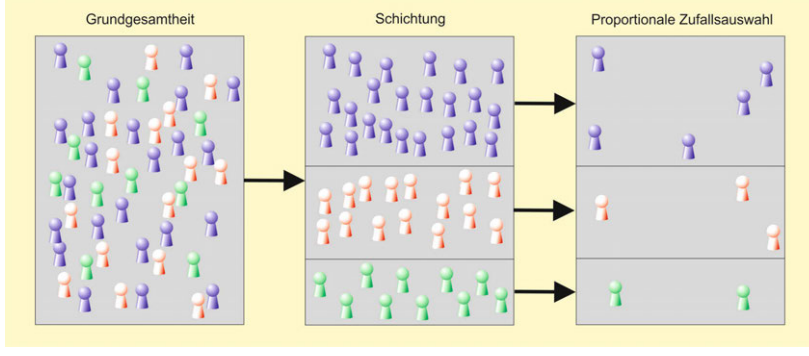


Abb. 3.6: *Schichtung mit proportionaler Stichprobenauswahl*

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 aber 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.

Flash-Animation „Klumpenstichprobe“

Systematische Auswahlprozeduren

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 (2012, 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).

3.3 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.

Organisation der amtlichen Statistik in Deutschland

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

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.

Träger nicht-amtlicher Statistik

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* in Essen (RWI), das *Institut für Weltwirtschaft* in Kiel (IfW) und das *Institut für Wirtschaftsforschung Halle* (IWH).

In die Markt- und Meinungsforschung, die im Auftrag von Unternehmen oder öffentlichen Einrichtungen erfolgt, werden erhebliche Summen investiert. Der *Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute* in Bonn (ADM) bezifferte das Geschäftsvolumen für Marktforschung für das Jahr 2012 in Europa auf ca. 11,85 Milliarden Euro, wobei 24% dieser Summe auf Deutschland entfielen. Relativ bekannte Institute sind z. B. die *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

Träger amtlicher und nicht-amtlicher Statistik gibt es weltweit. 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 **Organisation für wirtschaftliche Zusammenarbeit und Entwicklung** (OECD, engl.: *O*rganisation for *E*conomic *C*o-operation and *D*evelopment) bereit gestellt. Die OECD ist aufgrund stärkerer Marketingaktivitäten häufiger als Eurostat als Datenquelle in den Medien genannt, z. B. im Zusammenhang mit den PISA-Studien, wirkt aber nicht aktiv an der Harmonisierung von Datenerhebungen auf nationaler Ebene mit. Zu erwähnen ist auch die **UN Statistics Division**, das Statistikreferat der Vereinten Nationen.

4 Univariate Häufigkeitsverteilungen

Vorschau auf das Kapitel

In diesem Kapitel geht es um die Beschreibung und grafische Darstellung eines Datensatzes für ein einziges Merkmal. 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 Säulen- oder Balkendiagrammen visualisieren. Illustriert wird dies an Datensätzen zum Politbarometer und zur Nationalen Verzehrstudie II.

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 Würfelexperimenten 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 Einzelmerkmale erhoben werden, fällt i. a. 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.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_4

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 $x_1, \ldots, x_n$ 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“).

Klassenbildung bei stetigen Merkmalen

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 $a_1, \ldots, a_k$, so ist die im Folgenden mit

Verteilung von absoluten und relativen Häufigkeiten

$$h_i := h(a_i) \qquad i = 1, 2, \ldots, k \tag{4.1}$$

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. Um Häufigkeiten auch für Datensätze unterschiedlichen Umfangs direkt vergleichbar zu machen, teilt man die absoluten Häufigkeiten durch den Umfang n der Beobachtungsreihe. Die resultierenden

relativen Häufigkeiten

$$f_i := f(a_i) = \frac{h(a_i)}{n} \qquad i = 1, 2, \ldots, k \tag{4.2}$$

repräsentieren Anteile, die man manchmal in Form von Prozentwerten ausweist (Multiplikation mit 100).[1] Häufigkeiten lassen sich in Tabellenform ausweisen. Dabei resultieren **Häufigkeitsverteilungen** für absolute oder relative Häufigkeiten. Eine Häufigkeitsverteilung für ein Merkmal X wird auch als **empirische Verteilung** für dieses Merkmal bezeichnet. Es ist sofort einsichtig, dass sich die absoluten Häufigkeiten zu n und die relativen Häufigkeiten zu 1 addieren.

Visualisierung von Häufigkeitsverteilungen

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 Mittelpunktswinkel α_i, der die Größe des Kreissektors definiert, ist sowohl bei absoluten Häufigkeiten h_i als auch bei relativen Häufigkeiten f_i durch $f_i \cdot 360°$ gegeben. 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 auch mit Landkarten verknüpfen. Am 29. April 2015 veröffentlichte *Die Zeit* z. B. ein Blasendiagramm, das den Spargelkonsum in den deutschen Bundesländern darstellte. Die Mittelpunkte der Kreise waren hier in der Karte so platziert, dass sie in dem betreffenden Bundesland lagen.

Flash-Animation „Landtagswahl 2010 in NRW – Säulendiagramm“

Als Alternative zum Kreisdiagramm werden aber meist eher Stab- oder Säulendiagramme verwendet. Beim **Stabdiagramm** werden die Häufigkeiten durch vertikale dünne Stäbe (Striche), beim **Säulendiagramm** durch vertikale dicke Stäbe (Rechtecke) dargestellt. Ein Säulendiagramm wird auch **Balkendiagramm** genannt. Wenn die Merkmalsausprägungen Kategorien mit längeren Namen sind (etwa Namen von Staaten, Bundesländern oder Parteien), empfiehlt es sich Codes zu verwenden. Alternativ kann man ein Säulen- bzw. Balkendiagramm um 90° drehen und die Namen der Kategorien dann waagerecht präsentieren.[2]

[1] Die Bezeichnungen für Häufigkeiten sind in der Literatur nicht immer einheitlich. 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 / Künstler / Pigeot / Tutz (2010, Abschnitt 2.1.1) oder Steland (2013, Abschnitt 1.5.2).

[2] In einigen Lehrbüchern und auch in der Tabellenkalkulationssoftware Excel wird die Bezeichnung „Balkendiagramm“ nur auf diese gedrehten Säulendiagramme bezogen, der Begriff „Säulendiagramm“ also nur für Darstellungen mit vertikal angeordneten Säulen verwendet.

Java-Applet „Bruttoverdienste in Europa 2002“

Abbildung 4.1 zeigt mittlere Bruttostundenverdienste in Euro von Arbeitnehmern im Bereich „Industrie und Dienstleistungen“ in 27 europäischen Ländern für 2002 anhand eines Säulendiagramms. Die Ländernamen sind codiert; z. B. steht „AT“ für Österreich (Austria). Neben den Ländern der EU-25 des Jahres 2006 (ohne Malta), in dem die Ergebnisse veröffentlicht wurden, waren auch die damaligen Kandidatenländer Bulgarien und Rumänien sowie das EFTA-Land Norwegen an der Erhebung beteiligt. Staaten, die 2006 keine EU-Mitglieder waren, sind am Ende eingereiht.[3]

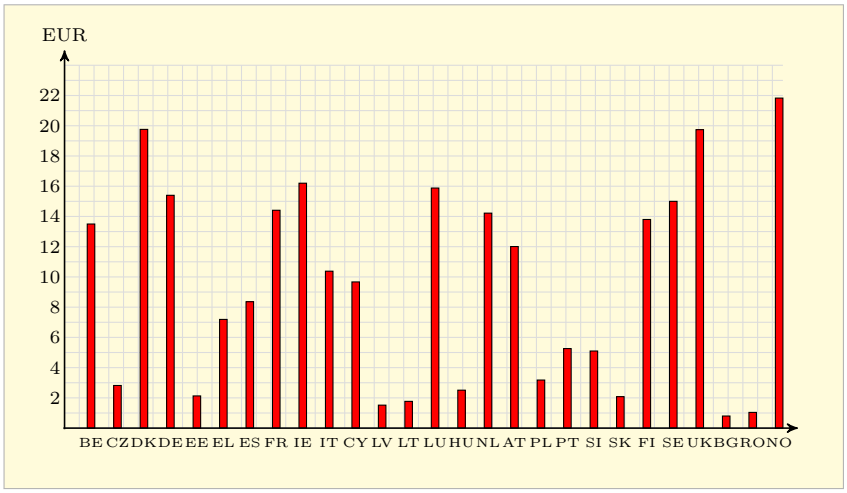


Abb. 4.1: *Bruttostundenverdienste in Europa* 2002 *(Säulendiagramm)*

In Zeitungen sowie im Fernsehen sieht man manchmal dreidimensional ausgestaltete Kreis- oder Säulendiagramme, bei denen die dritte Dimension ohne inhaltliche Bedeutung ist. Bei Kreisdiagrammen sollte diese Darstellungsform generell vermieden werden, weil sie reine Effekthascherei ist. Hingegen kann man 3D-Säulendiagramme für Häufigkeitsverteilungen durchaus in Betracht ziehen, wenn die dritte Dimension inhaltlich interpretierbar ist. Dies ist z. B. der Fall, wenn man zwei Häufigkeitsverteilungen in einer einzigen Grafik präsentiert und dies dadurch realisiert, dass man zwei Säulendiagramme hintereinander anordnet (s. Abbildung 4.6) oder jeweils zwei Säulen nebeneinander stellt.

Flash-Animation „Landtagswahl 2010 in NRW; verändertes Säulendiagramm“

Gewollt oder ungewollt manipulativ können Säulendiagramme oder Zeitreihengraphen sein, bei denen die vertikale Achse nicht bei 0 beginnt. Vergleicht man z. B. bei der NRW-Landtagswahl vom Mai 2010 die zwei Monate vor dem Wahltag erhobenen Präferenzen einer Stichprobe von Wählern für die beiden damaligen Bewerber um das Ministerpräsidentenamt anhand eines Säulendiagramms, so wird die Wirkung der

[3] Die 2002 erhobenen Daten sind Ergebnisse einer im 4-Jahres-Turnus durchgeführten Europäischen Verdienststrukturerhebung, die von EUROSTAT ausgewertet wird.

Grafik deutlich verändert, wenn man die Achse mit den Prozentwerten kappt und z. B. erst bei 25% beginnen lässt.

Mehr Nutzerfreundlichkeit in der amtlichen Statistik

Die amtliche Statistik bedient sich heute einer nutzerfreundlichen und interaktiven Datenkommunikation. So werden z. B. grafische Darstellungen von Häufigkeitsverteilungen mit Landkarten verknüpft. Mit der Maus lassen sich einzelne Regionen ansteuern und Daten für die ausgewählte Region grafisch präsentieren. Abbildung 4.2 zeigt eine frühere Online-Präsentation des amtlichen Endergebnisses der Bundestagswahl 2013. Im oberen Teil des umrahmten rechten Bereichs ließ sich hier eine Variable auswählen, etwa die Wahlbeteiligung oder der Erst- oder Zweitstimmenanteil einer Partei. Hier wurde die Variable „Zweitstimmenanteil der SPD“ eingestellt. Im unteren Teil des umrahmten rechten Felds ist ein Balkendiagramm zu sehen, welches den Zweitstimmenanteil aller Parteien visualisiert, die die Fünfprozenthürde überspringen konnten. Mit der Maus konnte ein Wahlkreis auf der Karte ausgewählt und die Wahlergebnisse für diesen Wahlkreis angezeigt werden.

Bundestagswahl 2013

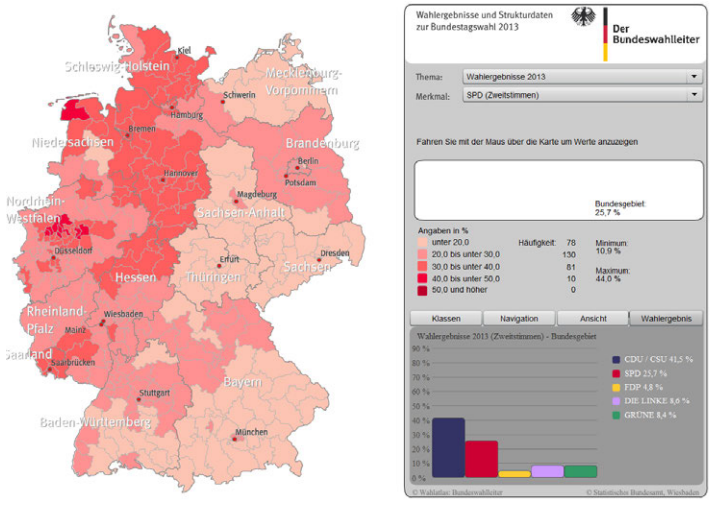


Abb. 4.2: *Ergebnisse der Bundestagswahl 2013 (links: eingefärbte Landkarte mit Ausweis des Zweitstimmenanteils der SPD in allen Wahlkreisen; rechts unten: Säulendiagramm mit den Zweitstimmenanteilen der Parteien (Quelle: Statistisches Bundesamt)*

Die Karte zeigt das Niveau der ausgewählten Variablen für alle Wahlkreise auf der Basis einer vom Nutzer veränderbaren Klassenbildung. Die einzelnen Klassen sind dabei farblich gestuft. Für die in Abbildung 4.2 gewählte Variable „Zweitstimmenanteil der SPD“ sind fünf Klassen eingestellt und in der Karte durch unterschiedliche Grautöne repräsentiert (Rottöne in der e-Buchfassung), wobei eine helle Farbausprägung einen niedrigen und ein dunklere einen hohen Zweitstimmenanteil widerspiegelt.

Beispiel 4.1: Ergebnisse des ZDF-Politbarometers vom 16. 10. 2009

Flash-Animation „Absolute und relative Häufigkeiten“

Bei der bekannten „Sonntagsfrage“ – einer im Auftrag des ZDF im Zwei-Wochen-Turnus durchgeführten Telefonbefragung – wird die Wahlentscheidung für den fiktiven Fall erfragt, dass am nächsten Sonntag Bundestagswahlen stattfinden. Abbildung 4.3 zeigt die am 16. Oktober 2009 veröffentlichten Ergebnisse einer solchen Befragung, die in der Zeit vom 13. - 15. Oktober lief. Die Ausprägungen $a_1, \ldots, a_6$ des Merkmals „präferierte Partei“ stehen für die CDU/CSU, SPD, FDP, die Linken, die Grünen resp. für „Sonstige“. Angegeben sind die absoluten Häufigkeiten $h(a_i)$ und die aus diesen abgeleiteten relativen Häufigkeiten $f(a_i)$ $(i = 1, 2, \ldots, 6)$, letztere auf drei Dezimalstellen gerundet.

Befragt wurden 1298 Personen, die per Zufallsauswahl aus der Grundgesamtheit aller Wahlberechtigten ausgewählt wurden. Hiervon sahen sich 277 Personen außerstande eine Präferenz zu nennen oder gaben an, überhaupt nicht zur Wahl gehen zu wollen. Diese Personen blieben unberücksichtigt. Die veröffentlichten Ergebnisse basierten also auf einer Stichprobe von $n = 1021$ Personen, die sich für eine Partei entscheiden konnten.

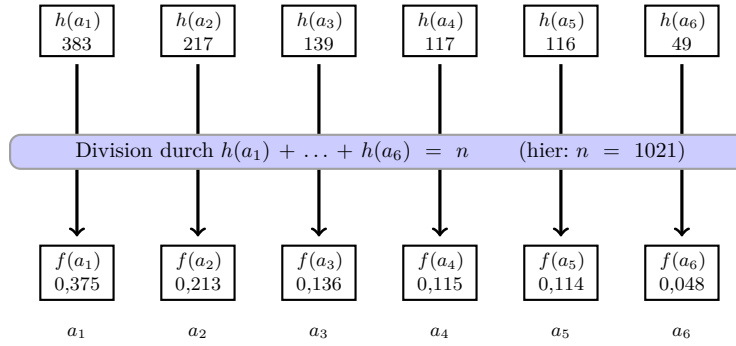


Abb. 4.3: *Häufigkeiten beim ZDF-Politbarometer vom 16. Oktober 2009; Quelle: Forschungsgruppe Wahlen*

Die Häufigkeitstabelle für das nominalskalierte Merkmal „Parteipräferenz“ ist in Tabelle 4.1 wiedergegeben. In der letzten Spalte der Tabelle sind – zum Vergleich mit den relativen Häufigkeiten in der dritten Tabellenspalte – auch die Zweitstimmenanteile bei der Bundestagswahl vom 27. September 2009 ausgewiesen (in Klammern die Zweitstimmenanteile der jüngsten Bundestagswahl vom 22. September 2013). Da zwischen der Befragung im Oktober 2009 und der vorausgegangenen Bundestagswahl nur knapp drei Wochen lagen, ist es nicht verwunderlich, dass die relativen Häufigkeiten in den letzten beiden Tabellenspalten nicht stark differieren. Unterschiede resultieren natürlich auch daraus, dass die „Sonntagsfrage“ nur eine relativ kleine Stichprobe der Wahlberechtigten umfasst.

		„Sonntagsfrage“ vom 16. 10. 2009		Bundestagswahl
	Merkmalsausprägung	Absolute Häufigkeit	Relative Häufigkeit	Zweitstimmenanteil
	a_i	$h(a_i)$	$f(a_i)$	2009 (2013)
CDU/CSU	a_1	383	0,375	0,338 (0,415)
SPD	a_2	217	0,213	0,230 (0,257)
FDP	a_3	139	0,136	0,146 (0,048)
DIE LINKE.	a_4	117	0,115	0,119 (0,086)
BÜNDNIS 90/DIE GRÜNEN	a_5	116	0,114	0,107 (0,084)
Sonstige	a_6	49	0,048	0,060 (0,110)
	Summe	$n = 1021$	1	1 (1)

Tab. 4.1: *Politbarometer vom 16. Oktober* 2009 *und Bundestagswahlergebnisse vom September* 2009 *(in Klammern: September* 2013*)*

Abbildung 4.4 zeigt für die obige Häufigkeitsverteilung je ein Kreis-, Stab- und Säulendiagramm. Beim Kreisdiagramm lassen sich Anteile ähnlicher Größe, etwa $f(a_3)$ und $f(a_4)$, nicht so gut unterscheiden wie beim Stab- oder Säulendiagramm.

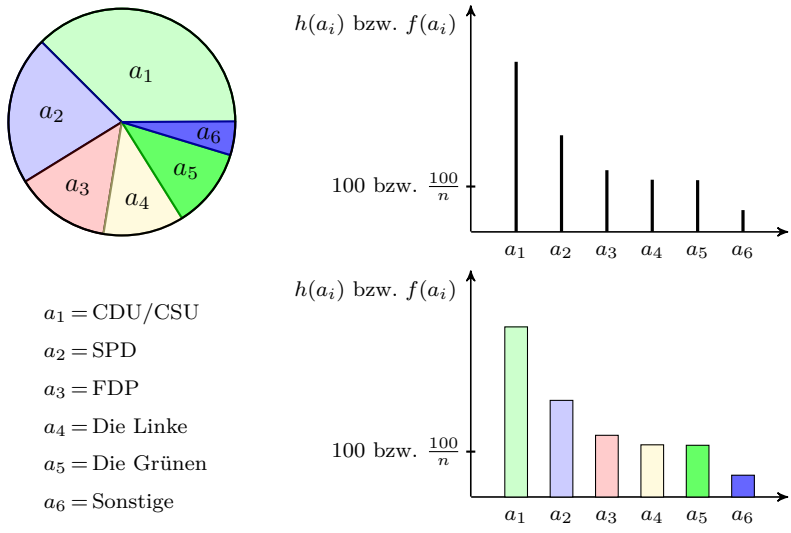


Abb. 4.4: *Kreis-, Stab- und Säulendiagramm (Politbarometerdaten)*

Stapelung von Häufigkeiten

Säulen- oder Balkendiagramme sind grafische Darstellungen von Häufigkeitsverteilungen, die in der Praxis auch in Modifikationen anzutreffen sind, bei denen die Säulen in zwei oder mehr Teile zerlegt und übereinander gestapelt sind. Man hat dann ein **gestapeltes Säulendiagramm** oder auch **gestapeltes Balkendiagramm**. Die Komponenten können durch unterschiedliche Schraffierung oder Färbung unterschieden werden. Hat man z. B. Stichproben von Personen für mehrere Regionen, so lassen sich die Personen der einzelnen Stichproben unter Verwendung einer geeigneten Operationalisierungsvorschrift drei Gewichtsklassen zuordnen, etwa a_1 (unter- oder normalgewichtig), a_2 (übergewichtig) und a_3 (fettleibig / stark übergewichtig / adipös). Anhand eines einfachen Balkendiagramms könnte man dann z. B. die Anzahl der nicht zu a_1 gehörenden Personen darstellen. Differenziertere Information erhält man, wenn man für jede Region die relativen Besetzungshäufigkeiten für alle drei Klassen anhand einer dreiteiligen Säule veranschaulicht oder für jede Stichprobe nur die Besetzungshäufigkeiten für a_2 und a_3 ausweist. Bei einer Zerlegung in mehr als zwei Komponenten wird ein gestapeltes Säulendiagramm schnell unübersichtlich. Durch Einblendung der numerischen Angaben kann dieser Nachteil gemildert werden (vgl. Abbildung 4.5).

Beispiel 4.2: Ergebnisse der Nationalen Verzehrstudie II

Im Auftrag des Bundesministeriums für Landwirtschaft und Ernährung wurde von Ende 2005 bis Anfang 2007 eine ca. 20 000 Personen umfassende Stichprobe der Bevölkerung Deutschlands nach ihrem Ernährungsverhalten befragt. Dabei wurde auch der Anteil der Übergewichtigen und Fettleibigen anhand des sog. *Body-Mass-Index* ermittelt.[4] Die 2008 veröffentlichten Ergebnisse dieser Nationalen Verzehrstudie II flossen in ein europäisches Gesundheits- und Ernährungsmonitoring ein.

Tabelle 4.2 zeigt die Ausprägungen des BMI-Wertes für die an der Studie beteiligten Männer in den deutschen Bundesländern. Bei den BMI-Werten wurde hier nur zwischen drei Ausprägungen a_1, a_2 und a_3 unterschieden (Gruppierung der Daten): a_1 entspricht Unter- oder Normalgewicht, a_2 bedeutet Übergewicht und a_3 Fettleibigkeit. Die Tabelle fasst für das klassierte Merkmal „BMI-Wert“ insgesamt 16 absolute und relative Häufigkeitsverteilungen zusammen – je eine pro Bundesland. Die Verteilungen umfassen bei der vorgenommenen Bildung von drei Klassen Häufigkeiten für jeweils drei Ausprägungen. Die Summe der absoluten Häufigkeiten in jeder Zeile von Tabelle 4.2 ergibt die Anzahl der in

[4] Der Body-Mass-Index BMI ist definiert als $BMI = m/l^2$, wobei m das Körpergewicht in kg und l die Körpergröße in Metern bezeichnet. Es wurde eine alters- und geschlechtsunabhängige Klassifikation herangezogen, nach der Personen mit einem BMI-Wert unter 18,5 als untergewichtig, 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 gelten. In Beispiel 4.2 werden die beiden erstgenannten Klassen zu einer Klasse zusammengefasst (Ausblendung des Problems von Untergewichtigkeit).

einem Bundesland befragten Männer. Für Bayern verifiziert man z. B., dass sich die Häufigkeiten $h(a_1) = 345$, $h(a_2) = 455$ und $h(a_3) = 218$ zu 1018 addieren.[5]

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	

Tab. 4.2: *Häufigkeiten für auffällige BMI-Werte bei Männern (ungewichtete Daten; Quelle: Mitteilung des Max-Rubner-Instituts)*

Abbildung 4.5 zeigt die relativen Häufigkeiten in Form gestapelter Säulendiagramme. Die Häufigkeiten $f(a_1)$ wurden hier unterdrückt und die Balken nach zunehmendem Wert von $f(a_2) + f(a_3) = 1 - f(a_1)$ geordnet. Die Grafik weist also im Gegensatz zu Tabelle 4.2 keine vollständigen Verteilungen aus.[6] Die Säulen beginnen jeweils mit den relativen Häufigkeiten $f(a_3)$ für Fettleibigkeit. Die Anteile $f(a_2)$ der Übergewichtigen lassen sich nicht leicht unterscheiden, weil die sie repräsentierenden Säulen unterschiedliche Anfangspunkte aufweisen. Deshalb sind auch die numerischen Werte eingeblendet. 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.

[5] Die Häufigkeiten in Tabelle 4.2 sind reale Beobachtungen. Diese wurden vor der Weitergabe an die Presse über Gewichtungsfaktoren korrigiert, um unterschiedliche Auswahlwahrscheinlichkeiten für die Zielpersonen der Stichprobe auszugleichen.

[6] Komplette relative Häufigkeitsverteilungen erhält man, wenn man die Abszissenachse bis 100 verlängert und jeweils einen dritten, bei 100 endenden Teilbalken anhängt.

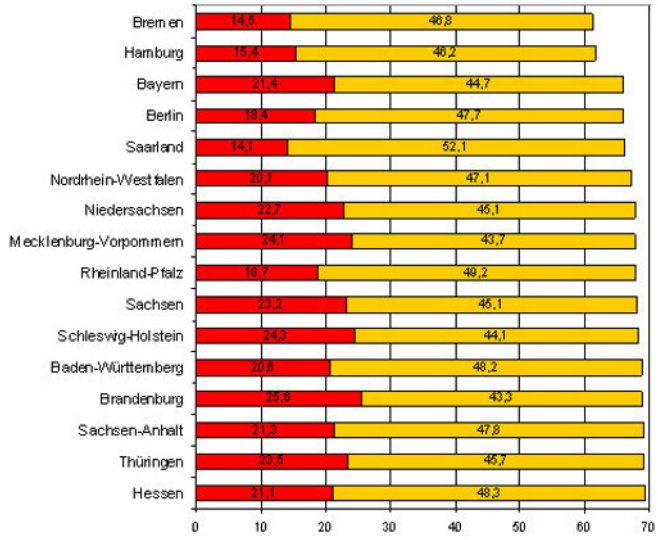


Abb. 4.5: *Ergebnisse der Nationalen Verzehrstudie II für Männer (gestapeltes Säulendiagramm; Teilbalken rechts: Anteil der Übergewichtigen; links: Anteil der Fettleibigen; in %)*

Aufgabe 4.1

Erwähnt sei, dass die Werte für Frauen besser ausfielen. Während der Prozentsatz der Männer mit Übergewicht oder Fettleibigkeit durchweg über 60% lag, wurde diese Quote bei den Frauen nur in einem Bundesland erreicht.

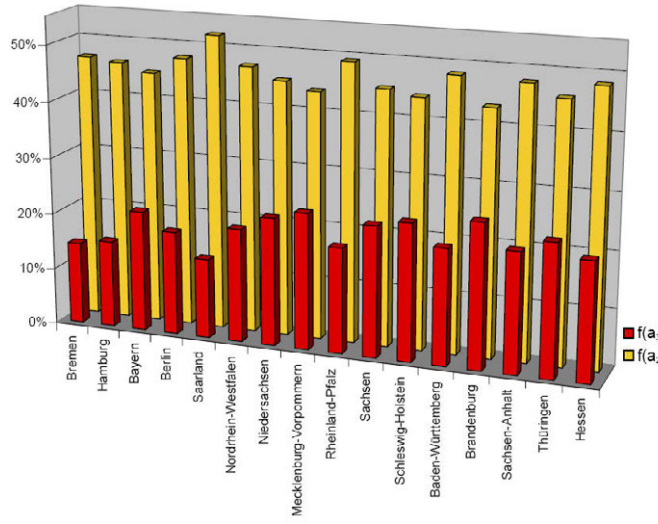


Abb. 4.6: *Ergebnisse der Nationalen Verzehrstudie II für Männer (3D-Säulendiagramm; Teilbalken hinten: Anteil der Übergewichtigen; vorne: Anteil der Fettleibigen; in %)*

Als Alternative zu Abbildung 4.5 kann man die Häufigkeiten für a_2 und a_3 hintereinander stellen statt sie zu stapeln, also ein **3D-Säulendiagramm** heranziehen. Dies ist in Abbildung 4.6 realisiert. Bei dieser Abbildung stören

die z. T. sehr langen Ländernamen – hier bietet sich eine Codierung an (z. B. „HH“ für Hamburg). Generell sollte man 3D-Säulendiagramme zurückhaltend verwenden, weil sie stets mit perspektivischen Verzerrungen einhergehen, die das Ablesen numerischer Werte erschweren. Dieser Nachteil lässt sich abmildern, wenn man die numerischen Werte zusätzlich präsentiert.

Übergewichtigkeit ist heute ein Problem vieler Länder, das sich – wenn nicht gegengesteuert wird – in der Zukunft noch verschärfen wird. Die Medien malen schon jetzt Schreckensszenarien an die Wand – vgl. etwa den als „UK fat alert“ apostrophierten Warnruf vom 26. August 2011 in der britischen Tageszeitung *The Independent*. Der Beitrag „Ein Drittel der Menschheit ist zu dick“ in der *Zeit* vom 29. Mai 2014 weist für 2013 in der Altersklasse ab 20 Jahren für die USA einen Anteil von 31,7 % fettleibiger Männer und 33,9 % fettleibiger Frauen aus (BMI-Index ≥ 30). Für Deutschland betrugen die entsprechenden Prozentwerte 21,9 % resp. 22,5 %, für Großbritannien 24,5 % und 25,4 %. Eine Warnung vor einer drastischen Zunahme von Übergewichtigkeit und Fettleibigkeit in Deutschland fand man auch am 19. Juli 2014 bei *Spiegel online*.

4.2 Häufigkeitsverteilungen für klassierte Daten

Bei klassierten Daten bezieht sich eine Häufigkeitsverteilung auf Klassenbesetzungshäufigkeiten. Auch hier kann man die absoluten oder relativen 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, anders als im Säulendiagramm aus Abbildung 4.4. 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.

Visualisierung von Einkommensverteilungen

Abbildung 4.7 zeigt erneut Bruttoverdienste von Arbeitnehmern in Europa im Bereich „Industrie und Dienstleistungen“ für 2002, nun nur für Spanien und Portugal. Die Jahreseinkommen umfassen auch Sonderzahlungen, etwa Boni, Weihnachtsgeld und Urlaubsgeld. Die Daten werden anhand von Histogrammen visualisiert (Klassierung der Individualdaten).[7] Die letzte der 15 Klassen, zu der im Vergleich zur vorletzten Klasse ein etwas höheres Rechteck gehört, ist hier nach oben offen. Würde man die Anzahl der Klassen deutlich erhöhen, entfiele die leichte Auffälligkeit der Höhe des letzten Rechtecks.[8]

Aufgabe 4.2

[7] Die Zusatzinformationen oberhalb der beiden in Abbildung 4.7 wiedergegebenen Histogramme werden in Kapitel 5 erläutert.

[8] Die Grafiken stammen aus einer Schrift von Eurostat zur Europäischen Verdienststrukturerhebung; s. Mittag (2006). Die 5000-Euro-Intervalle schließen die rechte Intervallgrenze nicht ein.

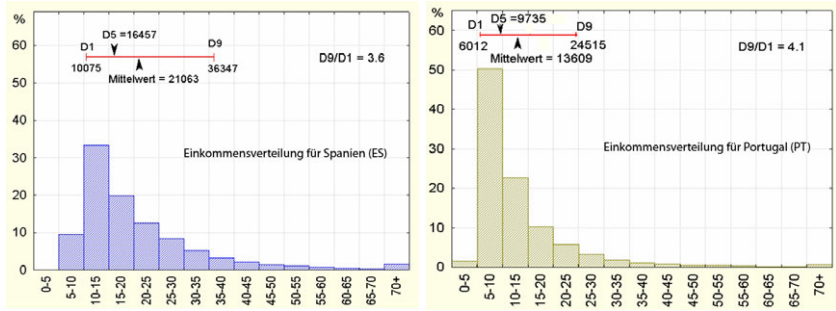


Abb. 4.7: *Bruttojahresverdienste* 2002 *in Spanien und Portugal in Tausend Euro (Histogramme der Einkommensverteilungen)*

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. im Internet eine ansprechende interaktive Visualisierung der bereits beobachteten bzw. vorausberechneten 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. Die aktuelle, mit den Statistischen Landesämtern koordinierte 13. Bevölkerungsvorausberechnung für Deutschland wurde Ende April 2015 in einer Pressemitteilung der Öffentlichkeit vorgestellt. Die Ergebnisse des europaweit durchgeführten Zensus 2011, die für Deutschland eine Korrektur der Bevölkerungszahlen nach unten mit sich brachten, sind hier erstmals vollständig eingearbeitet. Die Berechnungen basieren auf einem Set von je einer von zwei alternativ herangezogenen Annahmen zur Geburtenhäufigkeit pro Frau, zur Lebenserwartung Neugeborener und zur Stärke von Zuwanderungsbewegungen:

- *Geburtenhäufigkeit je Frau*:

  Annahme G1: konstant bei ca. 1,4 bei Anstieg des durchschnittlichen Geburtsalters um 13 Monate, G2: bis 2028 Anstieg auf etwa 1,6 bei Anstieg des durchschnittlichen Geburtsalters um ca. 8 Monate.

- *Lebenserwartung Neugeborener im Jahr* 2060:

  Annahme L1: 84,8 für Jungen / 88,8 für Mädchen; L2: 86,7 für Jungen / 90,4 für Mädchen).

- *Wanderungssaldo pro Jahr*:

  Annahme W1: Der positive Saldo von 500 000 Personen in 2015 schwächt sich bis 2021 langsam auf 100 000 ab und bleibt dann auf diesem Niveau. W2: Der positive Saldo von 500 000 Personen schwächt sich allmählich bis 2021 auf 200 000 Personen ab mit anschließender Konstanz dieses Wertes.

Abbildung 4.8 zeigt die Bevölkerungsstruktur von Deutschland für den Annahmenset G1-L1-W1. Eine Darstellung des in der Abbildung wiedergegebenen Typs wird meist **Bevölkerungspyramide** genannt, obwohl die Form der Grafik für die deutsche Bevölkerung zunehmend eher einem Pilz ähnelt. [9]

Animation „Altersstruktur für Deutschland“

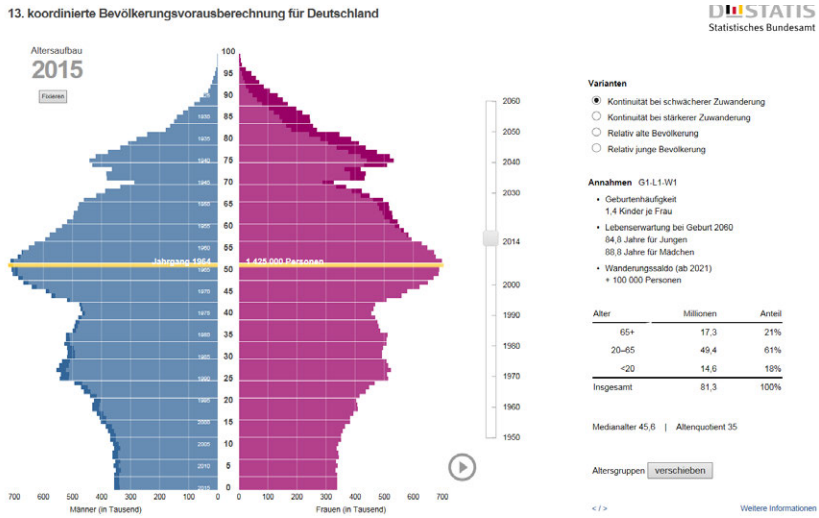


Abb. 4.8: *Visualisierung gruppierter Bevölkerungsdaten für Deutschland im Jahr* 2015 *(Quelle: Statistisches Bundesamt)*

In der Abbildung ist der geburtenstärkste Jahrgang von 1964 betont, dem in 2015 über 1,4 Millionen Menschen zuzurechnen sind – bei einer Gesamtbevölkerung von 81,3 Millionen. Die dunkle Fläche am oberen Rand des rechten Teilhistogramms weist den deutlichen Frauenüberschuss bei den älteren Jahrgängen aus (Altersbereich 55+), während die dunkle Fläche unten am Rand des linken Teilhistogramms den leichten Männerüberschuss bei jüngeren Jahrgängen repräsentiert. Man erkennt noch gut die kriegsbedingten Einschnitte bei den Jahrgängen $1945 - 1947$, die in 2015 das Alter von 70 bis 72 erreichen.

Starker Anstieg des Altersquotienten

Interessant ist das als **Altersquotient** bezeichnete Verhältnis „Anzahl älterer Menschen (ab 65 Jahre) / Anzahl der Menschen im Erwerbsalter (meist definiert als $20 - 64$ Jahre)“. Im Jahre 2015 gehören zur ersten Gruppe etwa 21%, zur zweiten Gruppe etwa 61% der Bevölkerung. Wenn man die interaktive Darstellung im Internet auf das Jahr 2030 stellt, wird die Gesamtbevölkerung bei fortgesetzter Gültigkeit von G1-L1-W1 auf 79,2 Millionen geschrumpft sein (2050: 71,9 Millionen). Der Anteil

[9] Das *Statistische Amt von Großbritannien* hat unter dem Etikett *100 Years of Census* die Bevölkerungsstruktur von England und Wales in Form einer tongestützten Animation aufbereitet. Die Darstellung umfasst alle Zensus-Ergebnisse der letzten 100 Jahre. Wenn man den Anfang der mit dem Jahr 1911 startenden Animation betrachtet, sieht man noch die klassische Pyramidenform.

der Seniorinnen und Senioren an der Gesamtbevölkerung wird sich unter G1-L1-W1 bei Rundung auf volle Prozentwerte auf 28% erhöhen (2050: 32%), während sich der Anteil der Menschen im erwerbsfähigen Alter auf ca. 55% (2050: 52%) vermindert.

Animation „Altersstruktur für Deutschland“

Abbildung 4.9 zeigt die Bevölkerungspyramide auf der Basis der Annahmen G1-L1-W1 für 2030. Die Bevölkerungsstruktur von 2015 aus Abbildung 4.8 ist als Umriss eingeblendet. Auf diese Weise wird die Veränderung der Altersstruktur sehr deutlich.

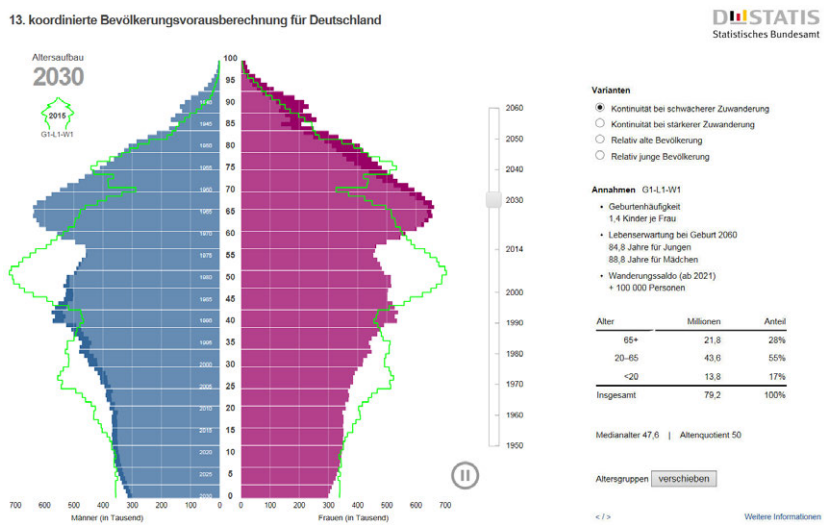


Abb. 4.9: *Visualisierung gruppierter Bevölkerungsdaten für Deutschland im Jahr* 2030 *(Quelle: Statistisches Bundesamt)*

Der Altersquotient wird in der Wirtschafts- und Sozialpolitik zur Quantifizierung von Veränderungen von Bevölkerungsstrukturen verwendet. Für das Jahr 2015 hat dieser Quotient den Wert 0,35 – in Abbildung 4.8 ist er als Prozentwert ausgewiesen. Der angegebene Prozentwert 35 beinhaltet, dass auf etwa 35 Personen im Ruhestand 100 Menschen im erwerbsfähigen Alter entfallen.

Der Altersquotient wird sich, wie Peer Steinbrück in einer Kolumne in der Wochenzeitschrift *Die Zeit* vom 7. Februar 2012 zutreffend feststellte, in den nächsten Jahren spürbar erhöhen. Unter den Annahmen G1-L1-W1 wird er bis 2050 auf ca. 0,60 ansteigen. Man sieht diese Verschiebung schon bei Betrachtung von Abbildung 4.9. Der „Bauch“ des in Umrissen wiedergegebenen Doppel-Histogramms für 2015 wandert bis 2030 nach oben. Der geburtenstärkste Jahrgang ist dann 66 Jahre alt und schon im Ruhestand oder kurz davor. 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 stufenweise Erhöhung des Ruhestand-Eintrittsalters von 65 auf 67

Jahre kann dies nur sehr bedingt auffangen. Am 29. April 2015 widmete sich *Die Welt* sehr ausführlich dem Themenkomplex „Alterssicherung“ und „Demografischer Wandel“.

Exkurs 4.1: Andere Visualisierungen des demografischen Wandels

Alternativen zur dynamischen Bevölkerungspyramide aus Abbildung 4.8 findet man auch im Bereich des Datenjournalismus. So veröffentlichte *Zeit online* ohne Datumsangabe eine Animation, bei der die deutsche Bevölkerung durch eine feste Anzahl stilisierter Figuren repräsentiert wird. Diese sind verschiedenen Altersklassen zugeordnet und in der jeweiligen Klasse alterstypisch dargestellt. Beim Abspielen der Animation ändert sich die Besetzung der Altersklassen und man sieht sehr eindrücklich, wie der Anteil junger Menschen ab- und der Anteil Hochbetagter zunimmt. Die Animation fußt allerdings noch auf den inzwischen revidierten Daten aus der 12. Bevölkerungsvorausberechnung des Statistischen Bundesamts von 2009.

Alternative Animation zum demografischen Wandel

Beispiel 4.3: Sensitivitätsanalysen zum demografischen Wandel

Die in Abbildung 4.8 und Abbildung 4.9 neben den Bevölkerungspyramiden eingeblendeten Annahmen zur Häufigkeit von Geburten, zur Veränderung der Lebenserwartung von Männern und Frauen sowie zur Stärke von Migrationsströmen bestimmen wesentlich das Ergebnis von Bevölkerungsvorausberechnungen. Der Einfluss der Annahmen auf die prognostizierten Daten wird natürlich um so sichtbarer, je weiter der Zielpunkt der Prognose entfernt liegt. Es ist auch klar, dass im Jahre 2060 deutlich andere gesellschaftliche Rahmenbedingungen vorliegen werden und die heute getroffenen Annahmen längst überholt sein können. Die interaktive Animation des *Statistischen Bundesamts* zur Altersstruktur der Bevölkerung ist als Visualisierung verschiedener Wenn-Dann-Szenarien zu verstehen, die es ermöglicht denkbare Entwicklungskorridore frühzeitig zu überblicken. Abbildung 4.10 zeigt solche Korridore für die Entwicklung der Geburtenhäufigkeit und der Lebenserwartung Neugeborener. [10]

Interaktives Lernobjekt „Lebenserwartung Neugeborener in der EU-28“

Das Statistische Bundesamt führt Sensitivitätsanalysen durch, die auch unrealistische Modellvorgaben einbeziehen, z. B. einen Anstieg der Geburtenhäufigkeit auf den Wert 2,1. So lassen sich Einzeleffekte deutlicher zeigen. Von den 8 realitätsnahen Szenarien sind 4 in die interaktive Animation integriert, nämlich die in Abbildung 4.8 und Abbildung 4.9 verwendete Kombination G1 - L1 - W1 sowie die Sets G1 - L1 - W2, G1 - L2 - W1 und G2 - L1 - W2. Vor allem die Annahmen zum Wanderungssaldo und zur Geburtenhäufigkeit können durch politische und wirtschaftliche Ereignisse und Maßnahmen rasch obsolet werden.

[10] Dass die Werte zu den Annahmen G1 und G2 für die Geburtenhäufigkeit im oberen Teil von Abbildung 4.10 etwas oberhalb von 1,4 bzw. 1,6 liegen, beruht darauf, dass bei den veröffentlichten Werten auf eine Dezimalstelle gerundet wurde. Die ungewöhnliche Skalierung der Zeitachse im unteren Teil von Abbildung 4.10 soll verdeutlichen, dass erst die Berechnungen nach 2012 auf Sterbetafeln für Einzeljahre gründen.

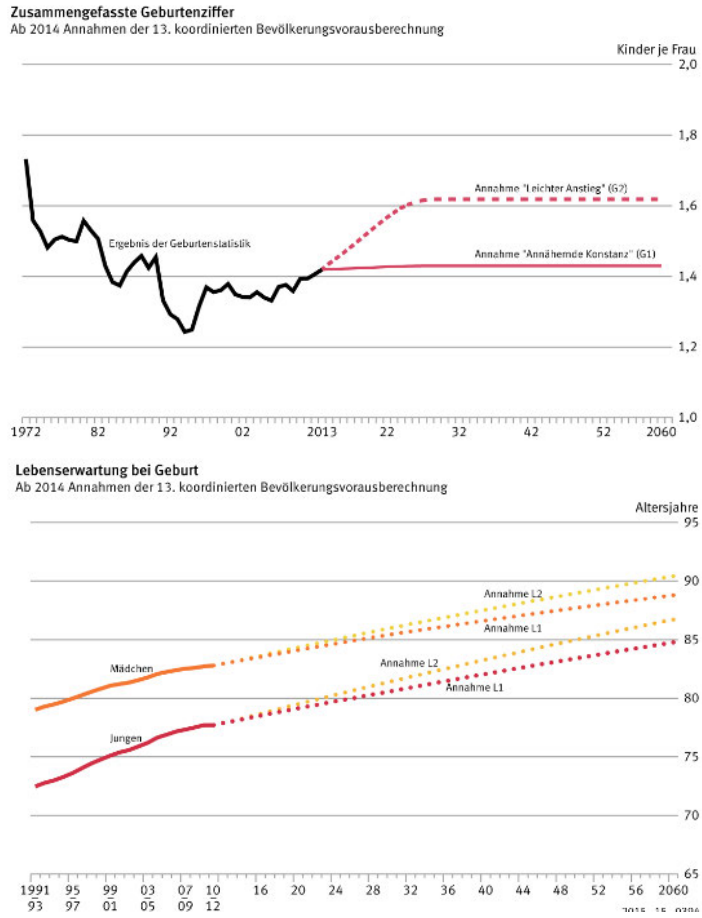


Abb. 4.10: *Geburtenhäufigkeiten und Lebenserwartung Neugeborener bis 2060 (Quelle: Statistisches Bundesamt; Stand: Mai* 2015*)*

Exkurs 4.2: Kontroverse Diskusssion des demografischen Wandels

Das Thema „Demografischer Wandel“ wurde von Thilo Sarrazin (2010) öffentlichkeitswirksam aufgegriffen. Mit seinem sehr kontrovers diskutierten Buch verfolgte das ehemalige Vorstandsmitglied der Deutschen Bundesbank das Ziel, künftige Sozial- und Einwanderungspolitik zu beeinflussen im Sinne einer Veränderung sozialer Anreizsysteme und einer stärkeren Orientierung der Einwanderungsbedingungen an Bildungsvoraussetzungen und Qualifikationsprofilen. Sarrazin brachte den langfristigen Bevölkerungsrückgang in Deutschland in Verbindung mit einem von ihm erwarteten zunehmenden Anteil von Menschen muslimischen Glaubens durch anhaltende Migration aus der Türkei, Nah- und Mittelost und Afrika, mit einem nach seiner Sicht hiermit verbundenen

wachsendem Anteil von Menschen mit niedrigem Bildungsgrad[11] und mit abnehmender Konkurrenzfähigkeit Deutschlands im globalen Wettbewerb. Als Basis für seine eigenen, nicht näher ausgeführten Rechnungen verwendete er die Ergebnisse der 11. Bevölkerungsprognose für Deutschland von 2006.

Der Wirtschaftsstatistiker Hans Wolfgang Brachinger (1951 - 2011) hat sich über die Internet-Plattform *Ökonomenstimme* nach Erscheinen der genannten Publikation kritisch geäußert und bemängelt, dass Sarrazin aus amtlichen Daten Schlussfolgerungen ziehe, ohne diese mit Instrumenten der schließenden Statistik abzusichern, etwa durch die theoretische Analyse vermuteter Ursache-Wirkungs-Zusammenhänge. Zusätzlich ist u. a. anzumerken, dass Sarrazin seine Berechnungen auf die Annahme eines positiven Wanderungssaldos von nur 50 000 Personen pro Jahr stützte. Diese Annahme wurde vom Statistischen Bundesamt zu keinem Zeitpunkt als realistisch erachtet und für Modellrechnungen verwendet. Mit den Annahmen, die Sarrazin seinen Berechnungen zugrunde legte, ergab sich für 2050 eine Gesamtbevölkerung Deutschlands von ca. 66,6 Millionen, mit G1 - L1 - W1 aus der aktuellen 13. Vorausberechnung sind es ca. 71,9 Millionen, bei Annahme von G1 - L1 - W2 hingegen 76,4 Millionen und unter G2 - L1 - W2 sogar 79,0 Millionen.

Histogramme sind grafische Instrumente, mit denen die in sehr umfangreichen Datensätzen enthaltenen Kerninformationen sichtbar werden können – z. B. bei Einkommensdaten die generell zu beobachtende Asymmmetrie 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

Man sollte wissen, dass der optische Eindruck eines Histogramms von der Klasseneinteilung abhängt, also von der Breite und den Anfangspunkten der Klassen. Die in Abbildung 4.7 in Form von Histogrammen veranschaulichten Einkommensverteilungen würden z. B. einen anderen Eindruck vermitteln, wenn man bei der Klassenbildung Intervalle von jeweils 1000 oder 2000 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 eingegangen werden (vgl. aber z. B. Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 2.4.3) oder Toutenburg / Heumann (2009, Abschnitt 2.3.5)). Es sei nur erwähnt, dass die Treppenfunktion, die den oberen Rand eines Histogramms darstellt, bei Kerndichteschätzern durch eine stetige Funktion ersetzt wird.

[11] Der Bildungsgrad wird in amtlichen Statistiken anhand der *I*nternational *S*tandard *C*lassification of *Ed*ucation (ISCED) erfasst, einem von der UNESCO zuletzt im November 2011 überarbeiteten Klassifikationsschema für Bildungssysteme.

4.3 Die empirische Verteilungsfunktion

In Abschnitt 4.1 wurde dargelegt, dass sich ein diskretes Merkmal X mit k Ausprägungen $a_1, \ldots, a_k$ anhand der absoluten oder relativen Häufigkeiten $h(a_1), \ldots, h(a_k)$ bzw. $f(a_1), \ldots, f(a_k)$ 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 4.7 illustriert, anhand eines Histogramms visualisieren.

Wenn die Merkmalswerte metrisch oder zumindest ordinalskaliert sind, also eine natürliche Rangordnung erklärt ist, will man oft auch wissen, wieviele Werte unterhalb oder oberhalb eines Schwellenwertes x liegen. Bei einem Datensatz, der den höchsten erreichten Bildungsabschluss einer Personengruppe beschreibt, kann man z. B. fragen, wieviele Personen einen Abschluss unterhalb eines Hochschulabschlusses haben. Beim n-fachen Würfeln mit einem Würfel kann man etwa an der Häufigkeit von Ergebnissen interessiert sein, die die Augenzahl 5 unterschreiten. Eine Antwort auf solche Fragen liefert die kumulierte Häufigkeitsverteilung.

Übergang zu kumulierten Häufigkeitsverteilungen

Betrachtet sei also ein zumindest ordinalskaliertes Merkmal X mit Ausprägungen $a_1, \ldots, a_k$,die nach aufsteigender Größe geordnet seien.. Für das Merkmal liegen n Beobachtungen x_i vor $(i = 1, 2, \ldots, n)$. 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 diese kumulierte Häufigkeitsverteilung $H(x)$ als Summe der absoluten Häufigkeiten $h(a_i)$, die der Bedingung $a_i \leq x$ genügen. 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. Formal lässt sich $H(x)$ wie folgt schreiben:

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

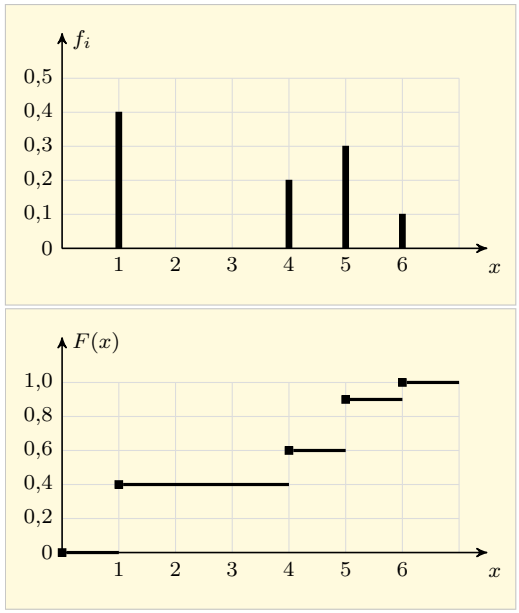
Die **relative kumulierte Häufigkeitsverteilung** $F(x)$ resultiert, wenn man $H(x)$ durch den Umfang n des Datensatzes dividiert:

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

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

$$F(x) = \begin{cases} 0 & \text{für } x < a_1 \\ f_1 & \text{für } a_1 \le x < a_2 \\ \vdots & \vdots \\ f_1 + f_2 + \ldots + f_{k-1} & \text{für } a_{k-1} \le x < a_k \\ 1 & \text{für } x \ge a_k, \end{cases} \tag{4.5}$$

repräsentiert also ebenfalls eine monoton steigende Treppenfunktion, die aber in $x = a_i$ $(i = 1,2,\ldots,k)$ 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.



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. 4.11: *Relative Häufigkeitsverteilung und relative kumulierte Häufigkeitsverteilung (10-faches Würfeln mit einem Würfel)*

Abbildung 4.11 zeigt im oberen Teil die beobachteten relativen Häufigkeiten $f_i = f(a_i)$ für die Augenzahlen $a_i = i$ $(i = 1,2,\ldots,6)$ bei einem Würfelexperiment, bei dem ein Würfel 10-mal geworfen wurde und dabei die 6 ein Mal, die 5 drei Mal, die 4 zwei Mal und die 1 vier Mal erschien.

Der untere Teil der Grafik veranschaulicht den Anteil $F(x)$ der Ausgänge mit $a_i \leq x$, zeigt also die empirische Verteilungsfunktion des Merkmals „Augenzahl X“. Neben der Grafik sind die Werte f_i und die mit F_i bezeichneten Werte $F(a_i)$ der empirischen Verteilungsfunktion an den Stellen $x = a_i$ tabelliert. Man erkennt, dass $F(x)$ an den Stellen $x = a_i = i$ um f_i springt. Für $i = 2$ und $i = 3$ ist die Sprunghöhe f_i allerdings 0, weil die Augenzahlen 2 und 3 hier nicht auftraten. Zwischen zwei benachbarten Ausprägungen von X ändert sich an der Summe der Häufigkeiten grundsätzlich nichts, d. h. die empirische Verteilungsfunktion bleibt hier auf konstantem Niveau.

Interaktives Lernobjekt „Augenzahlen“

Aus Abbildung 4.11 gewinnt man sofort eine grafische Darstellung der absoluten Häufigkeiten und der absoluten kumulierten Häufigkeiten, wenn man die Skalierung der Ordinatenachsen durch Multiplikation mit n abändert. Dort, wo in Abbildung 4.11 auf den Ordinatenachsen die Zahl 1 steht (Summe aller relativen Häufigkeiten f_i), erscheint dann der Wert n (Summe der absoluten Häufigkeiten h_i). Bereits in Abbildung 4.4 wurde anhand des dort dargestellten Stab- und Säulendiagramms verdeutlicht, das sich absolute und relative Häufigkeitsverteilungen nur hinsichtlich der Skalierung der Ordinatenachse unterscheiden.

Der untere Teil von Abbildung 4.11 bezieht sich auf eine empirische Verteilungsfunktion für ein *diskretes* Merkmal mit nur wenigen Merkmalsausprägungen. Da für das Merkmal „Augenzahl“ beim Würfeln mit einem Würfel nur sechs verschiedene Ausprägungen beobachtet werden können, kann die empirische Verteilungsfunktion auch nur höchstens sechs Sprünge aufweisen. Wenn man anstelle nur eines Würfels *zwei* Würfel verwendete und die empirische Verteilung der Augensumme visualisierte, hätte man schon 11 mögliche Ausprägungen, nämlich $2, 3, \ldots, 12$. Folglich kann die empirische Verteilungsfunktion des Merkmals „Augensumme“ hier auch bis zu 11 Sprünge aufweisen.

Interaktives Lernobjekt „Augensummen“

Bei einem Datensatz für ein diskretes Merkmal, bei dem – etwa wie beim Roulettespiel – eine noch größere Anzahl von Ausprägungen möglich ist, kann die empirische Verteilungsfunktion häufiger und in so kurzen Abständen springen, dass sie kaum noch als Sprungfunktion wahrzunehmen ist und als relativ glatter Kurvenzug erscheint.

Aufgabe 4.3

Beispiel 4.4: Empirische Verteilung einer Roulettespielserie

Der Ausgang beim Roulettespiel lässt sich durch ein diskretes Merkmal mit 37 Ausprägungen $a_i = i$ beschreiben ($i = 0, 1, 2, \ldots, 36$). Abbildung 4.12 zeigt die Häufigkeitsverteilung für dieses Merkmal bei einer Serie von $n = 100$ Spielen.

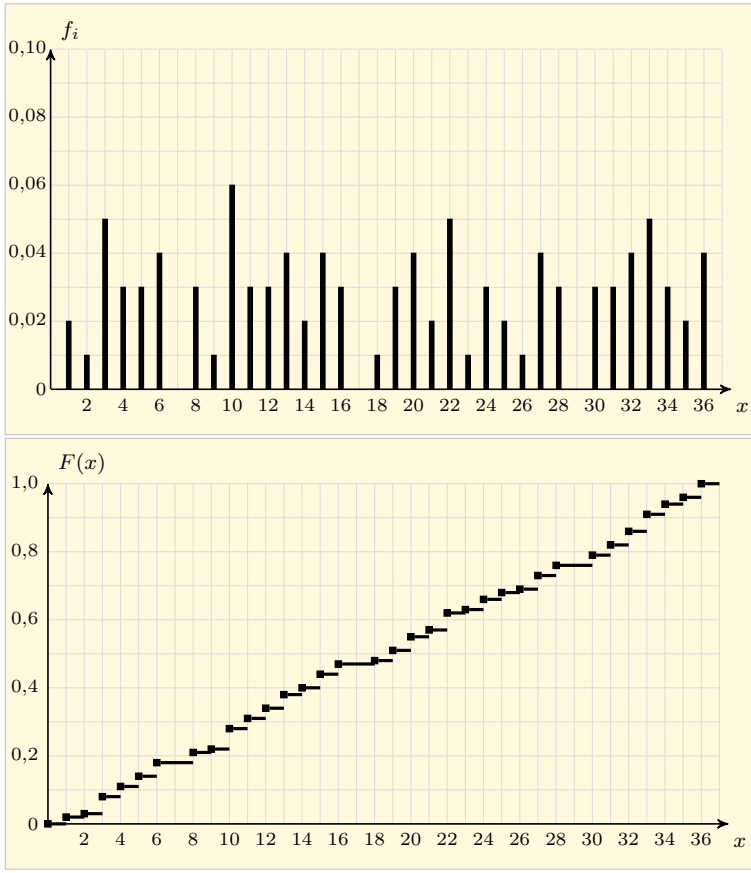


Abb. 4.12: *Relative Häufigkeitsverteilung und relative kumulierte Häufigkeitsverteilung einer Rouletteserie (Serie von $n = 100$ Spielen)*

Die Grafik zeigt im oberen Teil die relativen Häufigkeiten $f_i = f(a_i)$ für die 37 möglichen Ausgänge und im unteren Teil die durch Aufsummieren resultierende empirische Verteilungsfunktion $F(x)$. Letztere besitzt weniger als 36 Sprünge, weil die Ausgänge 0, 7, 17 und 29 bei der Spielserie nicht auftraten.

Bei Datensätzen für *stetige* Merkmale sind i. d. R. sogar alle Elemente des Datensatzes verschieden, so dass auch hier für die empirische Verteilungsfunktion bei größerem Umfang n des Datensatzes ein relativ glatter Kurvenzug resultierte, wenn man Häufigkeiten für die Originaldaten visualisierte. Hier wird man aber zweckmäßigerweise zu einer Klassenbildung übergehen und Klassenbesetzungshäufigkeiten darstellen, also

ein Histogramm wählen. Die beiden Histogramme in Abbildung 4.7, die 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

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 als nichtlineares Streuungsmaß und die empirische Standardabweichung sowie die Spannweite als lineare Streuungsmaße. Welche Kenngröße in Betracht kommt, hängt davon ab, ob die Daten auf einer Nominal- , Ordinal- oder einer metrischen Skala erfasst wurden.

Vorschau auf das Kapitel

Wenn mehrere Kenngrößen anwendbar sind – bei metrisch skalierten Daten z. B. alle drei genannten Lageparameter – muss die Aussagekraft der alternativen Kenngrößen im jeweiligen inhaltlichen Kontext bei der Auswahl berücksichtigt werden. Dies wird an einem Beispiel aus Großbritannien zum Haushaltseinkommen illustriert.

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.

5.1 Lagemaße

Häufigkeitsverteilungen für ungruppierte oder gruppierte Daten vermitteln einen Eindruck von der Gestalt der Verteilung eines Datensatzes. Die Histogramme in Abbildung 4.7 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

Wofür werden Kenngrößen von Verteilungen benötigt?

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_5

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 empirischer Verteilungen sind aber dennoch überaus 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 16. Oktober 2009, 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** $\widetilde{x}$ (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 $x_1, x_2, \ldots, x_n$ für ein solches Merkmal. Um zwischen dem ursprünglichen und dem geordneten Datensatz unterscheiden zu können, sei letzterer mit $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ bezeichnet.[1] 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 $x_{(\frac{n}{2}+1)}$

[1] Man kann auf die Notation $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ verzichten, wenn man von der Annahme ausgeht, dass der Datensatz $x_1, x_2, \ldots, x_n$ schon geordnet vorliegt.

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

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

Der bekannteste Lageparameter ist das **arithmetische Mittel**, das im Folgenden auch als **Mittelwert** angesprochen und mit $\overline{x}$ 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 $x_1, x_2, \ldots, x_n$ eines Datensatzes addiert und die resultierende Summe durch n dividiert:[2]

$$\overline{x} := \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i. \tag{5.2}$$

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, d. h. gegenüber auffällig großen oder kleinen Beobachtungswerten.

Java-Applet „Lageparameter“

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

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

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

Beispiel 5.2: Median und Mittelwert für Energieverbrauchsdaten

In der Wochenzeitung *Die Zeit* vom 11. 4. 2002 fand man in Ergänzung eines mit „Big Oil regiert“ überschriebenen Beitrags die nachstehende Tabelle mit umweltrelevanten Kennzahlen für die USA, Deutschland, Japan, China und

[2] Das Summenzeichen Σ und andere mathematische Symbole sind in Tabelle 22.3 erklärt. Unter dem Summenzeichen wird für den – in (5.2) mit „i“ bezeichneten – ganzzahligen Laufindex der Startwert angegeben, über dem Summenzeichen der letzte zu berücksichtigende Wert des Laufindexes.

Indien. Die Daten beziehen sich auf das Jahr 1999 und stammen von der *Internationalen Energieagentur*.

Land	Erdölverbrauch (in t/Kopf)	Stromverbrauch (in 1000 kWh/Kopf)	CO_2-Emissionen (in t/Kopf)
USA	8,32	13,45	20,46
Deutschland	4,11	6,48	10,01
Japan	4,07	8,13	9,14
China	0,87	0,91	2,40
Indien	0,48	0,42	0,91

Tab. 5.1: *Umweltrelevante Daten für fünf Staaten*

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.

Es seien hier die Daten für das metrisch skalierte Merkmal „Stromverbrauch / Kopf“ (in 1000 kWh) in der mittleren Spalte von Tabelle 5.1 betrachtet. Um den Median zu errechnen, sind die Werte $x_1 = 13{,}45$, $x_2 = 6{,}48$, $x_3 = 8{,}13$, $x_4 = 0{,}91$, $x_5 = 0{,}42$ zunächst nach Größe zu ordnen. Aus der resultierenden Folge $x_{(1)} = 0{,}42$, $x_{(2)} = 0{,}91$, $x_{(3)} = 6{,}48$, $x_{(4)} = 8{,}13$, $x_{(5)} = 13{,}45$ ergibt sich der Median für den hier vorliegenden Fall $n = 5$ nach (5.1) als $\widetilde{x} = x_{(3)} = 6{,}48$. Würde man bei dem ursprünglichen Datensatz den Wert $x_5 = 0{,}42$ für Indien unberücksichtigt lassen, den Median also nur auf der Basis der Datenreihe $x_1, \ldots, x_4$ ermitteln, erhielte man für $\widetilde{x}$ den Wert $\widetilde{x} = \frac{1}{2} \cdot (x_{(2)} + x_{(3)}) = 7{,}305$.

Bestimmt man mit denselben Ausgangsdaten den Mittelwert, so erhält man nach (5.2) den Wert $\overline{x} = \frac{1}{5} \cdot 29{,}39 = 5{,}878$. Würde man für x_1 anstelle von 13,45 den 10-fach größeren Wert 134,50 einsetzen, bliebe der Median bei $\widetilde{x} = 6{,}48$, während sich für den Mittelwert nun $\overline{x} = \frac{1}{5} \cdot 150{,}44 = 30{,}088$ 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 $a_1, \ldots, a_k$ insgesamt n Beobachtungswerte $x_1, \ldots, x_n$ $(n > k)$, 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 $f(a_1), \ldots, f(a_k)$ verwenden und $\overline{x}$ nach

$$\overline{x} := a_1 \cdot f_1 + a_2 \cdot f_2 + \ldots + a_k \cdot f_k = \sum_{i=1}^{k} a_i \cdot f_i \tag{5.4}$$

als Summe von nur k Termen berechnen. Das arithmetische Mittel $\overline{x}$ lässt sich also alternativ als Summe der mit den relativen Häufigkeiten f_i gewichteten Ausprägungen a_i ermitteln ($i = 1{,}2, \ldots, k$).

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.

Beispiel 5.3: Mittelwertbestimmung bei einem Würfelexperiment

In Abbildung 4.10 wurde das Ergebnis eines 10 Würfe umfassenden Würfelexperiments veranschaulicht, bei dem vier Mal die 1, zwei Mal die 4, drei Mal die 5 und einmal die 6 beobachtet wurde. Nach (5.2) erhält man für $\overline{x}$ den Wert

$$\overline{x} = \frac{1}{10} \cdot (1 + 1 + 1 + 1 + 4 + 4 + 5 + 5 + 5 + 6) = \frac{1}{10} \cdot 33 = 3{,}3.$$

Zieht man bei der Berechnung des Mittelwerts (5.4) heran, resultiert mit den neben Abbildung 4.10 tabellierten relativen Häufigkeiten $f_i = f(a_i)$

$$\overline{x} = 1 \cdot 0{,}4 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot 0{,}2 + 5 \cdot 0{,}3 + 6 \cdot 0{,}1 = 3{,}3.$$

Die Vorteile der Formel (5.4) verstärken sich, wenn für n ein noch größerer Wert gewählt wird, z. B. bei einem Würfelexperiment $n = 1000$ Würfe.

Gibt es einen „besten“ Lageparameter?

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, vor allem wenn die Verdienste auf Cent genau ausgewiesen werden. 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 liefern kann. Bei einer Einkommensverteilung kann man z. B. $\overline{x}$ und $\widetilde{x}$ vergleichen und hieraus Aussagen zur Symmetrie oder Asymmetrie der Verteilung ableiten.

Beispiel 5.4: 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 abgenommen habe, zum ersten Mal seit Beginn der 90er Jahre, und zwar um 0,2 % auf nunmehr 408 Britische Pfund. Dieser Befund wurde von der Presse sehr kritisch kommentiert, so dass schließlich Gordon Brown, der damalige Schatzkanzler und spätere Premierminister, Stellung beziehen musste.

Gordon Brown. Quelle: World Economic Forum

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 Britische Pfund 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. Man erkennt dies z. B. anhand von Abbildung 4.7. Diese zeigte zwei Einkommensverteilungen und zusätzlich – oberhalb der Grafiken – den aus den Individualdaten errechneten Mittelwert sowie drei Dezile, von denen eines der dort mit D5 bezeichnete Median war. Bloßes Betrachten der Abbildungen macht schon deutlich, dass das arithmetische Mittel für die betrachteten Grundgesamtheiten weniger repräsentativ als der Median ist. 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.[3]

Dass die Journalisten den Report negativ kommentierten, lag entweder daran, dass sie zwischen dem arithmetischen Mittel und dem Median nicht recht zu unterscheiden wussten oder aber unterstellten, dass dies für die Leser zutrifft. Statistische Methodenkompetenz ist offenbar eine Voraussetzung dafür, besser gegenüber unscharfen oder manipulativen Darstellungen statistischer Sachverhalte in den Medien gefeit zu sein.

[3] Dieser Befund schlug sich im Bericht in einer leichten Abnahme des *Gini-Koeffizienten* nieder, der neben dem Quotienten von Dezilen, etwa $\frac{D9}{D1}$, als Maß für Einkommensungleichheiten Verwendung findet (vgl. hierzu Kapitel 6).

Exkurs 5.1: Weitere Lageparameter

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

$$z = \overline{x}: \qquad \sum_{i=1}^{n} (x_i - z)^2 \to Min.$$

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

$$z = \widetilde{x}: \qquad \sum_{i=1}^{n} |x_i - z| \to Min.$$

Beweise dieser Aussagen findet man z. B. bei Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 2.2.1) oder Schlittgen (2012, Abschnitt 3.1).

Für metrisch skalierte Merkmale gibt es noch weitere Lageparameter. Zu nennen ist hier das **gewichtete arithmetische Mittel**, bei dem die Werte $x_1, x_2, \ldots, x_n$ 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 ein Land jeweils mit dem Gewichtungsfaktor n_i multipliziert.

Zu erwähnen ist auch das **getrimmte arithmetische Mittel**. Dieses lässt einen kleineren Anteil der Randdaten $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eines nach aufsteigender Größe geordneten Datensatzes unberücksichtigt. Wenn dieser Anteil α beträgt, spricht man auch von einem α-getrimmten Mittelwert und kürzt diesen mit $\overline{x}_\alpha$ ab. Bei der Berechnung von $\overline{x}_\alpha$ 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 $\overline{x}_g$ bezeichnete **geometrische Mittel** wird für Datensätze $x_1, x_2, \ldots, x_n$ verwendet, die Veränderungsraten repräsentieren, z. B. Wachstumsraten. Es errechnet sich als

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}.$$

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.

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 $x_1, x_2, \ldots, x_n$ mit Mittelwert $\overline{x}$, so lässt die alleinige Kenntnis von $\overline{x}$ offen, ob die einzelnen Elemente des Datensatzes alle sehr nahe am Mittelwert liegen, mit ihm gar alle übereinstimmen oder von $\overline{x}$ stark abweichen und sich nur „ausmitteln“. Zur Charakterisierung von Merkmalen, für die Abstände zwischen Merkmalsausprägungen erklärt sind, also bei quantitativen Merkmalen, muss man somit 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 $\widetilde{x}$ – den Datensatz zunächst nach aufsteigender Größe. Die Spannweite ergibt sich dann aus dem geordneten Datensatz $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ als Differenz aus dem größten Wert $x_{(n)}$ und dem kleinsten Wert $x_{(1)}$:

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

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 auch massiv auf den Wert von R aus.

Ein häufiger verwendetes Maß für die Streuung eines Datensatzes ist die **Varianz** oder **Stichprobenvarianz** s^2, die auch **empirische Varianz** genannt wird.[4] In die Varianz gehen die Abweichungen $x_i - \overline{x}$ der Merkmalswerte vom Mittelwert $\overline{x}$ ein; $i = 1, 2, \ldots, n$. Wegen (5.3) kommt die Verwendung des Mittelwerts aus allen Abweichungen $x_i - \overline{x}$ nicht als Streuungsmaß in Betracht. Die Varianz bildet statt dessen den Mittelwert

[4] Das Verhalten von Zufallsvariablen wird in den Kapiteln 11 - 12 anhand von Modellen (Wahrscheinlichkeitsverteilungen) charakterisiert. Hier spricht man von *theoretischen Verteilungen*, deren Streuung durch eine *theoretische* Varianz beschrieben wird.

aus den Quadraten der Abweichungen $x_i - \overline{x}$, d. h. es gilt

$$s^2 := \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{5.6}$$

Bei der Varianzberechnung kann die nachstehende Darstellung nützlich sein, bei der $\overline{x^2}$ das arithmetische Mittel der quadrierten Werte $x_1^2, \ldots, x_n^2$ des Datensatzes bezeichnet: [5]

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

Java-Applet „Lage und Streuungsparameter“

Die Darstellung (5.7) geht aus (5.6) hervor, wenn man dort den quadrierten Term $(x_i - \overline{x})^2$ hinter dem Summenzeichen ausmultipliziert (binomische Formel) 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**

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

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

Vorsicht: Uneinheitliche Definition von Varianz und Standardabweichung

Die Bezeichnungen für Varianz und Standardabweichung eines Datensatzes sind in der Lehrbuchliteratur leider nicht einheitlich. Häufig wird für die Varianz anstelle von (5.6) eine Formel verwendet, bei der vor dem Summenterm anstelle von $\frac{1}{n}$ der Term $\frac{1}{n-1}$ steht. Das dann resultierende und hier mit s^{*2} abgekürzte Streuungsmaß

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

wird **korrigierte Varianz** oder **korrigierte Stichprobenvarianz** genannt (vgl. auch Mosler / Schmid (2009, Abschnitt 5.1.4)). Nach Wurzelziehen resultiert die **korrigierte Standardabweichung** s^*.

Die korrigierte Varianz wird beim Schätzen und Testen anstelle von (5.7) bevorzugt verwendet, weil sie – wie mit (14.8) und (14.9) noch gezeigt wird – günstigere Eigenschaften besitzt. Die Division durch $n - 1$

[5] Sind 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.

wird erst im Kontext der schließenden Statistik nachvollziehbar; sie lässt sich im Rahmen der beschreibenden Statistik nicht motivieren. Wichtig ist aber, dass man bei Verwendung eines Taschenrechners oder einer Statistiksoftware 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. (5.7) 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.[6]

Beispiel 5.5: Streuung bei Stromverbrauchsdaten

Geht man erneut vom Datensatz 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 von R zunächst in die geordnete Folge $x_{(1)} = 0{,}42$, $x_{(2)} = 0{,}91$, $x_{(3)} = 6{,}48$, $x_{(4)} = 8{,}13$, $x_{(5)} = 13{,}45$ zu überführen. Es errechnet sich dann $R = 13{,}45 - 0{,}42 = 13{,}03$. Würde man bei dem ursprünglichen Datensatz den Wert 13,45 für die USA z. B. auf den Wert 8,13 von Japan herabsetzen, hätte dies für die Spannweite einen erheblichen Effekt. Es resultierte nun für R der Wert $R = 8{,}13 - 0{,}42 = 7{,}71$.

Bei der Berechnung der empirischen Varianz nach (5.6) werden die Originaldaten um den Mittelwert $\overline{x} = 5{,}878$ vermindert und die resultierenden Mittelwertabweichungen quadriert, aufsummiert und durch $n = 5$ dividiert. Man erhält so bei Rundung auf drei Nachkommastellen

$$s^2 = \frac{1}{5} \cdot \left[7{,}572^2 + 0{,}602^2 + 2{,}252^2 + (-4{,}968)^2 + (-5{,}458)^2\right] \approx 23{,}448.$$

Geht man alternativ von (5.7) aus, erhält man, wenn man auf drei Dezimalstellen rundet und auf den in Beispiel 5.2 errechneten Mittelwert $\overline{x} = 5{,}878$ zurückgreift, die etwas kürzere Rechnung

$$s^2 = \frac{1}{5} \cdot 289{,}9943 - 5{,}878^2 \approx 57{,}999 - 34{,}551 = 23{,}448.$$

Für die Standardabweichung folgt mit (5.8)

$$s = \sqrt{\frac{1}{5} \cdot [7{,}572^2 + 0{,}602^2 + 2{,}252^2 + (-4{,}968)^2 + (-5{,}458)^2]} \approx 4{,}842.$$

[6] In EXCEL wird eine Prozedur zur Berechnung der empirischen Standardabweichung s gemäß (5.7) angeboten und zusätzlich eine für die korrigierte Standardabweichung s^*. Bei der Statistiksoftware SPSS wird hingegen bei der Berechnung von Varianz und Standardabweichung eines Datensatzes stets durch $n - 1$ dividiert. SPSS bezeichnet ein in den *Sozialwissenschaften* und in der *Psychologie* häufig verwendetes Statistik-Softwarepaket (die Abkürzung stand anfangs für *S*tatistical *P*ackage for the *S*ocial *S*ciences). Als Alternative zu kommerzieller Statistiksoftware wird bei der statistischen Analyse von Daten zunehmend *R* eingesetzt – eine kostenfreie und sehr leistungsfähige Statistik-Software und Programmierumgebung.

Die korrigierte empirische Varianz errechnet sich als $s^{*2} = \frac{5}{4} \cdot s^2 \approx 29{,}310$. Der Unterschied zu $s^2 \approx 23{,}448$ 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 $a_1, \ldots, a_k$ eine größere Anzahl n von Beobachtungswerten $x_1, \ldots, x_n$ vor $(n > k)$, so wären bei der Anwendung von (5.7) n Mittelwertabweichungen $x_i - \overline{x}$ zu quadrieren. Statt der Abweichungen $x_i - \overline{x}$ der Urwerte vom Mittelwert kann man alternativ die Abweichungen $a_i - \overline{x}$ der Merkmalsausprägungen vom Mittelwert heranziehen und deren Quadrate mit den Elementen f_i der relativen Häufigkeitsverteilung $f_1 = f(a_1), \ldots, f_k = f(a_k)$ gewichten. Man erhält so für s^2 die zu (5.4) analoge alternative Berechnungsformel

Alternative Berechnung der Varianz

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

bei der sich die Summenbildung auf nur k Terme bezieht. Auch diese Formel lässt sich zur Varianzberechnung bei *gruppierten Daten* heranziehen, wenn man die Ausprägungen a_i durch die Mitte m_i der Klassen ersetzt. Die Häufigkeiten f_i entprechen dann wieder den relativen Besetzungshäufigkeiten der einzelnen Klassen.

Beispiel 5.6: Varianz bei einem Würfelexperiment

Aufgabe 5.1

Es sei noch einmal der Datensatz $\{1, 1, 1, 1, 4, 4, 5, 5, 5, 6\}$ zugrunde gelegt, der den Ausgang des in Abbildung 4.10 veranschaulichten Würfelexperiments beschreibt (Augenzahlen bei 10 Würfen). In Beispiel 5.3 war auf der Basis dieser 10 Werte der Mittelwert $\overline{x} = 3{,}3$ berechnet worden und zwar anhand der Urwerte und alternativ unter Verwendung der relativen Häufigkeiten.

Wenn man die Varianz s^2 unter Rückgriff auf die Urwerte berechnet, kann man (5.6) oder (5.7) verwenden. Bei Verwendung von (5.7) ergibt sich

$$s^2 = \frac{1}{10} \cdot 147 - 3{,}3^2 = 14{,}70 - 10{,}89 = 3{,}81.$$

Zieht man bei der Berechnung der Varianz (5.10) heran, resultiert

$$\begin{aligned} s^2 :=\ & \big[(-2{,}3)^2 \cdot 0{,}4 + (-1{,}3)^2 \cdot 0 + (-0{,}3)^2 \cdot 0 \\ & +0{,}7^2 \cdot 0{,}2 + 1{,}7^2 \cdot 0{,}3 + 2{,}7^2 \cdot 0{,}1\big] \\ =\ & 2{,}116 + 0{,}098 + 0{,}867 + 0{,}729 = 3{,}81. \end{aligned}$$

Standardisierung von Datensätzen

Wenn man Datensätze $x_1, x_2, \ldots, x_n$, die sich auf Messungen in unterschiedlichen Grundgesamtheiten beziehen oder die mit unterschiedlichen Messinstrumenten gewonnen wurden, direkt vergleichbar machen will, kann man von jedem Element eines Datensatzes jeweils dessen Mittelwert $\overline{x}$ subtrahieren und die Differenz noch durch die Standardabweichung s oder die korrigierte Standardabweichung s^* dividieren. Es resultieren neue Datensätze $y_1, y_2, \ldots, y_n$ mit Mittelwert $\overline{y} = 0$ 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 12.2 dieses Manuskripts noch ausführlicher behandelten Transformation (12.11), die zur Standardisierung von Zufallsvariablen herangezogen wird.

Exkurs 5.2: Verhalten der Kenngrößen bei Lineartransformation

Varianz s^2 und Standardabweichung s sind Streuungsmaße, die sich auf Abweichungen $x_i - \overline{x}$ vom *Mittelwert* eines Datensatzes für ein metrisch skaliertes Merkmal beziehen. Ein alternatives Streuungsmaß ist die **mittlere absolute Abweichung vom Median**. Dieses oft mit d abgekürzte Maß basiert auf Abweichungen $x_i - \widetilde{x}$ vom Median, bildet aber nicht den Mittelwert aus den Quadraten, sondern aus den Absolutbeträgen dieser Abweichungen:

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

Wenn man die Daten x_i für ein quantitatives Merkmal einer Transformation $y_i = a + b \cdot x_i$ unterzieht, so werden Median und Mittelwert sowie die Standardabweichung in gleicher Weise transformiert, d. h. es gilt z. B. für den Mittelwert y der transformierten Daten die Beziehung $\overline{y} = a + b \cdot \overline{x}$. Auf die Varianz und die Standardabweichung wirkt sich die Niveauverschiebung a nicht aus; nur der Wert von b ist hier 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 $s_y = |b| \cdot s_x$.

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

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

verwendet. Dieser repräsentiert ein *maßstabsunabhängiges* Streuungsmaß.

5.3 Quantile und Boxplots

Der für ein metrisch oder mindestens ordinalskaliertes Merkmal erklärte Median $\widetilde{x}$ hat die Eigenschaft, dass mindestens 50% der nach Größe geordneten Elemente $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eines Datensatzes kleiner oder gleich und mindestens 50% größer oder gleich $\widetilde{x}$ sind. Bei den 5 Werten in der mittleren Spalte von Tabelle 5.1 war der Median z. B. durch $\widetilde{x} = x_{(3)} = 6{,}48$ gegeben und je 3 der 5 Elemente in dieser Spalte, d. h. 60% der Werte, waren kleiner oder gleich resp. größer oder gleich $\widetilde{x}$. Bei ordinalskaliertem Merkmal ist $\widetilde{x}$ nicht immer eindeutig bestimmt. Bei metrischer Skalierung gilt dies im Prinzip auch; hier lässt sich aber über (5.1) eine eindeutige Festlegung erreichen.

Verallgemeinerung des Medians

Der Median markiert also die „Mitte“ eines Datensatzes. 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 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ kleiner oder gleich und mindestens $(1-p) \cdot 100\%$ größer oder gleich x_p sind.[7] Abbildung 5.1 veranschaulicht diese Definition.

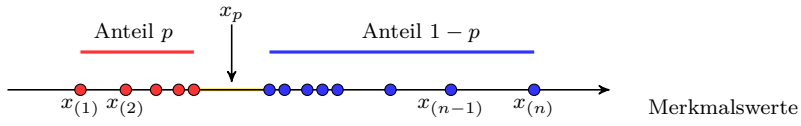


Abb. 5.1: *Veranschaulichung des p-Quantils*

Auch das p-Quantil ist bei einem ordinalskalierten Merkmal i. d. R. nicht eindeutig bestimmt. Bei metrischer Merkmalsskalierung kann, analog zur Definition des Medians, eine Eindeutigkeit erreicht werden, wenn zur Berechnung das arithmetische Mittel derjenigen 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.[8]

Mit dieser Notation kann x_p bei einem metrisch skalierten Merkmal in Verallgemeinerung von (5.1) eindeutig definiert werden durch (vgl. z. B.

[7] Die Notation für Quantile ist in der Literatur nicht einheitlich. Man findet z. B. auch die Schreibweise $\widetilde{x}_p$ anstelle von x_p; vgl. z. B. Steland (2013, Abschnitt 1.6.4) oder Toutenburg / Heumann (2009, Abschnitt 3.1.2).

[8] Die 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. $[3{,}8] = 3$.

Burkschat / Cramer / Kamps (2012))

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

Spezielle Quantile

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

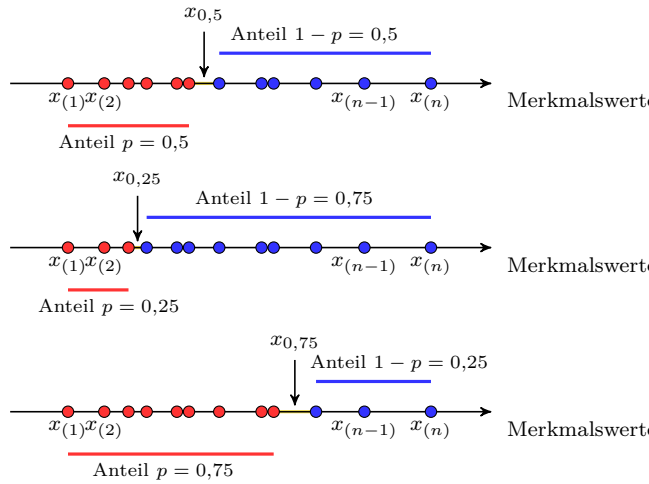


Abb. 5.2: *Median $x_{0,5}$, unteres Quartil $x_{0,25}$ und oberes Quartil $x_{0,75}$*

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

$$Q := x_{0,75} - x_{0,25}, \tag{5.12}$$

wird **Quartilsabstand** genannt. Sie wird in manchen Lehrbüchern auch als **Interquartilsabstand** IQR angesprochen (engl: *i*nter*q*uartile *r*ange). Ferner sind noch die Dezile zu nennen, die sich bei Wahl von $p = 0{,}1, p = 0{,}2, \ldots, p = 0{,}9$ ergeben und oft mit D1, D2, ..., D9 abgekürzt werden. Der Median $\widetilde{x} = x_{0,5}$ stimmt also mit dem Dezil D5 überein.

Wie erkennt man eine asymmetrische Verteilung?

In Abbildung 4.7 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 $= \widetilde{x}$ und der Mittelwert $\overline{x}$ 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 4.7 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 **asymmetrische Verteilung** vor. Die Nicht-Übereinstimmung von Median und Mittelwert einer empirischen Verteilung ist stets ein Indiz für eine Asymmetrie dieser Verteilung.

Boxplots:

- Basisvariante

Ein sehr 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 $x_{0,5}$.

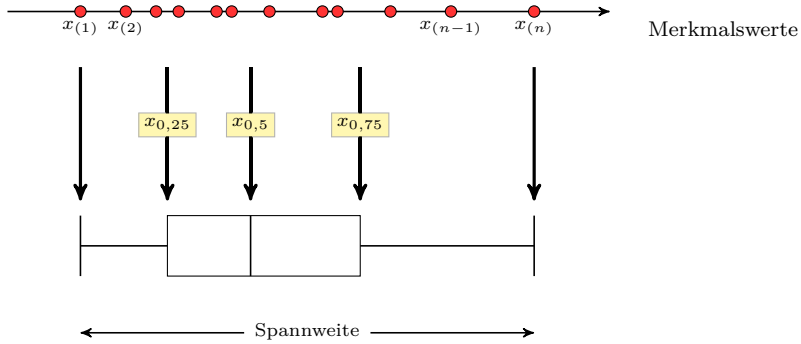


Abb. 5.3: *Aufbau eines Boxplots (Basisversion)*

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. „whisker“, 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. 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.

Aufgabe 5.2-3

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

- Modifikation (Visualisierung von Ausreißern)

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 Quartilabstands 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.

Beispiel 5.7: Boxplots zu Bruttoverdiensten in Europa

Java-Applet „Bruttoverdienste in Europa 2002“ (Boxplots)

Abbildung 4.1 zeigte Bruttostundenverdienste des Europäischen Amts für Statistik (Eurostat) in 27 europäischen Staaten für das Referenzjahr 2002 anhand eines Säulendiagramms. Die Darstellung bezog sich auf den Bereich „Industrie und Dienstleistungen“, in dem 9 Wirtschaftszweige zusammengefasst sind. Die in Abbildung 4.1 veranschaulichten Werte sind Mittelwerte aus den Verdiensten in diesen Branchen (gewichtete Mittel mit der Anzahl der in einem Wirtschaftszweig Beschäftigten als Gewichte). Wenn man ein etwas differenziertes Bild gewinnen will und z. B. auf einen Blick erfassen möchte, wie die Verdienste in den einzelnen Ländern von Branche zu Branche streuen, kann man für jedes Land einen Boxplot heranziehen, der den aus 9 Branchenverdiensten bestehenden Datensatz für jedes Land zu 5 Charakteristika aggregiert.

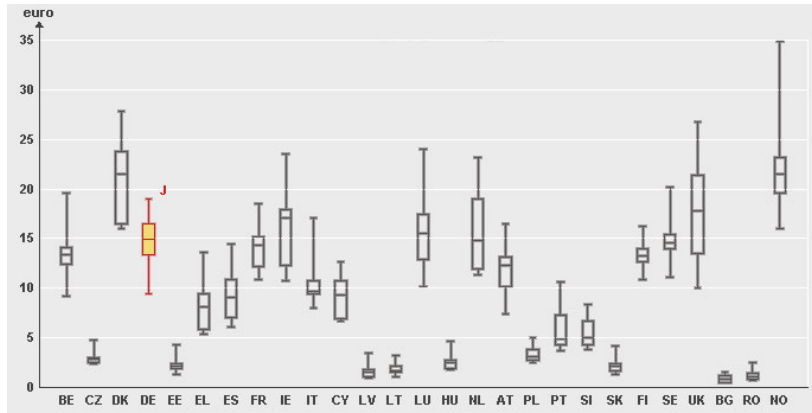


Abb. 5.4: *Streuung von Bruttoverdiensten zwischen Wirtschaftszweigen*

Der Boxplot für Deutschland ist in der Grafik betont. Der die obere Begrenzung des Boxplots definierende maximale Wert des Datensatzes, also die Branche, in der in Deutschland die Verdienste am höchsten sind, ist ebenfalls hervorgehoben. Es ist dies der Finanzsektor „Kreditinstitute und Versicherungen“, der nach der „*n*omenclature générale des *ac*tivités *é*conomique“ (amtliche Klassifikation NACE für Wirtschaftszweige; Stand 2006) mit “J“ codiert wurde.

Man erkennt anhand des Niveaus der Mediane, wie extrem im Jahr 2002 das mittlere Verdienstniveau zwischen den Staaten variierte – mit sehr niedrigen Niveaus in Bulgarien (BG) und Rumänien (RO) und hohen Niveaus in Dänemark (DK) oder Norwegen (NO). Die Grafik kann zum Verständnis der

fortschreitenden Arbeitsplatzverlagerungen in Niedriglohnländer im Zuge der Globalisierung beitragen. Starke Verdienstniveauunterschiede in Europa ließen sich allerdings schon aus Abbildung 4.1 ableiten. Die Boxplots liefern aber ein wesentlich differenzierteres Bild als Abbildung 4.1. Man erkennt nämlich hier auch, dass die Spannweite zwischen den Branchen mit minimalen und maximalen Verdiensten von Land zu Land recht unterschiedlich ausfällt (z. B. kleinere Spannweite für Dänemark im Vergleich zu Norwegen). Boxplots mit großer Spannweite und kleinem Quartilsabstand (kürzere Boxen) weisen auf wenig ausgeglichene Einkommensverteilungen hin. Abbildung 5.4, hinter der Individualdaten von Millionen europäischer Arbeitnehmer stehen, illustriert, dass man mit geeigneten Visualisierungsinstrumenten zentrale „Botschaften“ und Auffälligkeiten sichtbar machen kann, die sich aus unüberschaubaren „Zahlenfriedhöfen“ alleine nicht ohne weiteres erschließen lassen.

Exkurs 5.3: 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 Entwickungsländern beobachtet und durch Hilfsprogramme bekämpft, z. B. durch das mit den *UN Millennium Development Goals* beschriebenem Entwicklungsprogramm der Vereinten Nationen. Dessen erstes Ziel lautet „Beseitigung von Hunger und extremer Armut“.

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 *Armutsgefährdungsgrenze*. Eine in einem Ein-Personen-Haushalt lebende Person wird als „armutsgefährdet“ angesehen, wenn ihr weniger als 60 % des Medians der nationalen Einkommensverteilung zur Verfügung stehen. Mit „Einkommen“ ist hier das Nettoeinkommen unter Einbezug staatlicher Transferleistungen gemeint. Da sich die Einkommensverteilung eines Landes im Zeitverlauf ändert, ändert sich auch der Wert der als Bezugsgröße verwendeten Quote von 60 % des Medianeinkommens. Im Jahr 2009 lag der Grenzwert bei 940 Euro / Monat und 2013 bei 979 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 Armutsgefährdung in Armut übergeht, ist nicht einheitlich definiert, bestimmt sich aber wieder über einen Prozentsatz des Medians der

nationalen Einkommensverteilung. Eurostat und nationale Statistikämter in Europa verwenden 40 % des Medians als Schwellenwert, der die Kategorien „arm / nicht-arm“ trennt.

Die Heranziehung der nationalen Einkommensverteilung bei der Definition von Armut und Armutsgefährdung für Deutschland impliziert, dass regionale Einkommensunterschiede, etwa solche zwischen Bundesländern, unberücksichtigt bleiben. 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 Armutsquote kann hier deutlich nach unten verfälscht sein. Die *Süddeutsche Zeitung* weist in ihrer Ausgabe vom 4. November 2011 darauf hin, dass 15,6 % der Deutschen im Jahr 2009 als „armutsgefährdet“ galten, in München aber nur 10,7 %. Würde man für München aber nicht 60 % des Medians der gesamtdeutschen Einkommensverteilung als Grenzwert verwenden, sondern 60 % des Medians der Einkommensverteilung von Bayern oder gar nur von München, müssten schon 13,6 % resp. ca. 18,0 % der Einwohner Münchens in 2009 den Status „armutsgefährdet“ erhalten“. Quantitative Informationen zum Anteil der Armutsgefährdeten sind also nur sinnvoll interpretierbar, wenn man die Bezugsgröße kennt. Sie liefern – ähnlich wie in Abbildung 4.7 die Quantilsquotienten D9/D1 – nur eine Aussage über Einkommensungleichheit.

Es gibt auch Ansätze zur Berechnung von Armutsgefährdungsquoten auf der Basis regionaler Einkommensdaten. Die dabei erhobenen Armutsgefährdungsquoten informieren über Armutsunterschiede *innerhalb* der betreffenden Region (*vertikale* Armut), erfassen also keine Unterschiede *zwischen* Regionen (*horizontale* Armut). Bei Verwendung nationaler Einkommensverteilungen wird vertikale und horizontale Armutsidentifikation vermischt. Beide Ansätze messen etwas anderes und ermöglichen zusammen differenziertere Analysen.

Die Verwendung nationaler Einkommensverteilungen bei der Definition von Armut und Armutsgefährdung hat zur Folge, dass eine Person, die in Deutschland als armutsgefährdet 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 armutsgefährdeten Personen an der Gesamtbevölkerung in Tschechien bei nur 9,0 %, also deutlich unter der für Deutschland ermittelten Quote von 15,6 %.

Die Daten zur Armutsgefährdung werden im Rahmen einer Erhebung über Einkommen und Lebensbedingungen in Europa gewonnen. Die Erhebung ist unter dem Kürzel *EU-SILC* bekannt (*E*uropean *U*nion *S*tatistics on *I*ncome and *L*iving *C*onditions), in Deutschland unter „Leben in Europa“.

6 Konzentration von Merkmalswerten

Bei metrisch skalierten Merkmalen mit nicht-negativen Ausprägungen – etwa Einkommen von Arbeitnehmern oder Marktanteile 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 Merkmalssumme auf sich vereint.

Vorschau auf das Kapitel

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.

6.1 Die Lorenzkurve

Bei metrisch skalierten Merkmalen mit nicht-negativen Ausprägungen – z. B. Umsätze oder Marktanteile 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 Merkmalssumme 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?

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_6

- 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 6.1: Energieverbrauch und CO_2-Emissionen

In Tabelle 5.1 waren Daten der *Internationalen Energieagentur* zum Pro-Kopf-Verbrauch von Erdöl und Strom sowie zu den CO_2-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 5.4 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 CO_2-Emission aller fünf Länder innerhalb der 5 Elemente umfassenden Grundgesamtheit verteilt.

Flash-Animation „Merkmalskonzentration“

Ein wichtiges Instrument für die grafische 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 $x_1 \ldots, x_n$. Wenn man deren Elemente nach zunehmender Größe sortiert, resultiert eine geordnete Liste $x_{(1)} \ldots, x_{(n)}$. 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

$$u_i := \frac{i}{n}; \qquad i = 1, \ldots, n, \tag{6.1}$$

so 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

$$p_i := x_{(1)} + x_{(2)} + \ldots + x_{(i)}; \qquad i = 1, \ldots, n \tag{6.2}$$

und den Anteil der zugehörigen Merkmalsträger an der Merkmalssumme p_n mit

$$v_i := \frac{p_i}{p_n}; \qquad i = 1, \ldots, n, \tag{6.3}$$

Java-Applet „Lorenzkurve“

so ist die Lorenzkurve ein aus n Teilstrecken bestehender monoton steigender Polygonzug, der den Punkt $(0;0)$ mit den Punkten $(u_1;v_1), \ldots, (u_n;v_n)$ 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.

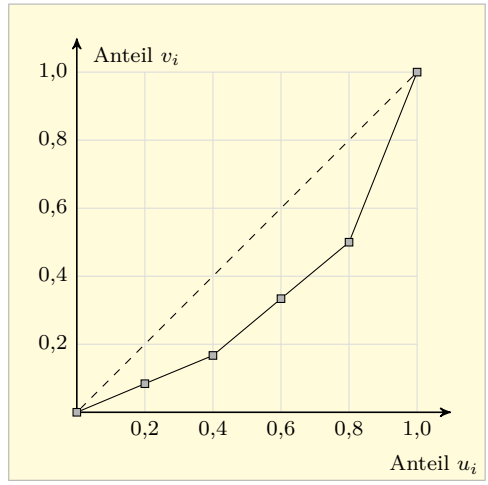


Abb. 6.1: *Beispiel einer Lorenzkurve im Falle* $n = 5$

Abbildung 6.1 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 (6.1) - (6.3), 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 6.1 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 / Künstler / Pigeot / Tutz (2010, Abschnitt 2.3.1) oder Toutenburg / Heumann (2009, Abschnitt 3.5.1).

6.2 Konzentrationsmaße

Corrado Gini

Die Lorenzkurve visualisiert Konzentrationsphänomene, repräsentiert aber noch kein Maß für die Stärke von Konzentration. Da sie sich mit zunehmender Konzentration immer mehr von der im konzentrationsfreien Fall 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 6.2 weist erneut die Lorenzkurve aus Abbildung 6.1 für die Urwerte 20, 20, 40, 40 und 120 aus, nun mit Hervorhebung der Fläche $A = \frac{G}{2}$.

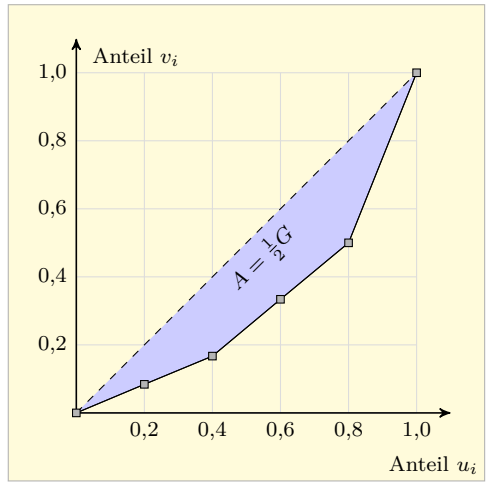


Abb. 6.2: *Veranschaulichung von* $\frac{G}{2}$ *im Falle* $n = 5$

Um $G = 2 \cdot A$ zu berechnen, ist es zweckmäßig, den Inhalt B der in Abbildung 6.3 betonten Fläche zu betrachten, die aus einem Dreieck mit dem Flächeninhalt 0,5 und einer Fläche mit dem Inhalt A besteht. Es gilt also $B = \frac{G}{2} + 0{,}5$, d. h. $G = 2B - 1$. Die Fläche mit dem

Flächeninhalt B lässt sich nun, wie in Abbildung 6.3 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 $G = 2 \cdot$ (Summe der Inhalte der n Teilflächen) $- 1$.

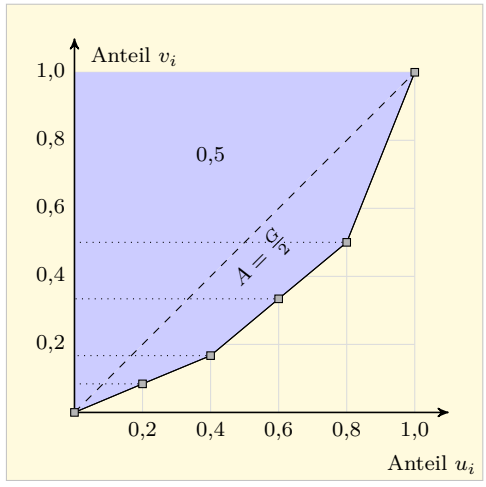


Abb. 6.3: *Geometriegestützte Herleitung einer Formel für G*

Man erhält bei Anwendung elementarer Flächeninhaltsformeln mit p_n aus (6.2) und mit der gewichteten Merkmalssumme

$$q_n := 1 \cdot x_{(1)} + 2 \cdot x_{(2)} + \ldots + n \cdot x_{(n)} \tag{6.4}$$

nach einigen Umformungen für den Gini-Koeffizienten die Darstellung[1]

$$G = \frac{2 \cdot q_n}{n \cdot p_n} - \frac{n+1}{n} = \frac{1}{n}\left(\frac{2 \cdot q_n}{p_n} - 1\right) - 1. \tag{6.5}$$

Für die Urliste mit den Elementen 20, 20, 40, 40 und 120, deren Lorenzkurve in Abbildung 6.1 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 6.4 ist der Fall maximaler Konzentration dargestellt. Zugrunde 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.

[1] Vgl. etwa Bamberg / Baur / Krapp (2012, Abschnitt 3.4.2).

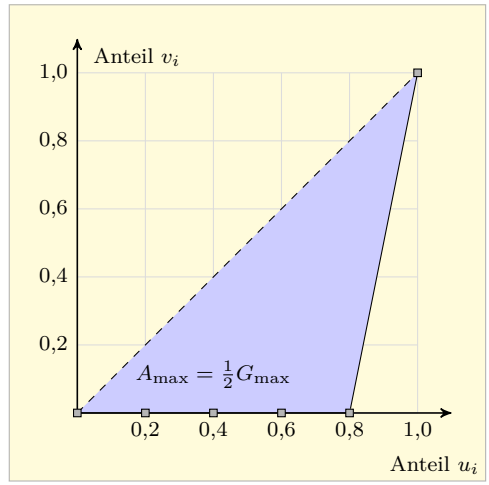


Abb. 6.4: *Veranschaulichung von $\frac{G}{2}$ bei maximaler Konzentration ($n = 5$)*

Java-Applet „Gini-Koeffizient“

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. Für den Gini-Koeffizienten gilt also stets

$$0 \leq G \leq \frac{n-1}{n}. \tag{6.6}$$

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**

$$G^* := \frac{G}{G_{max}} = \frac{n}{n-1} \cdot G \tag{6.7}$$

beheben. Für den normierten Gini-Koeffizienten hat man also

$$0 \leq G^* \leq 1, \tag{6.8}$$

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.

Quantifizierung von Einkommensungleichheit

Besondere Bedeutung kommt dem Gini-Koeffizienten bei der Quantifizierung von Einkommensungleichheiten zu (s. hierzu den Exkurs 6.1). Für diese Zielsetzung kann man alternativ auch Quantilsquotienten empirischer Einkommensverteilungen heranziehen. Wenn man z. B., wie in Abbildung 4.7 illustriert, das Verhältnis $\frac{D9}{D1}$ 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.

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.

Relative und absolute Konzentration

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

$$H := \sum_{i=1}^{n} \left(\frac{x_i}{p_n}\right)^2 = \frac{1}{p_n^2} \cdot \sum_{i=1}^{n} x_i^2, \tag{6.9}$$

also als Summe der quadrierten Anteile $\frac{x_i}{p_n}$ der einzelnen Elemente der Urliste. 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

$$\frac{1}{n} \leq H \leq 1. \tag{6.10}$$

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 0,2), bereits deutliche Konzentration beinhalten. Bei größeren Werten von n wird schon ein Indexwert H von ca. 0,18 als Indiz für deutliche Konzentration angesehen.

Anwendungsfeld für den Herfindahl-Index

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.

Beispiel 6.2: Konzentrationsmessung bei Stromverbrauchsdaten

Will man für die Daten zum Stromverbrauch in Tabelle 5.1 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 (6.1) durch $u_i = \frac{i}{5}$ gegeben, also durch 0,2, 0,4, ..., 1,0; die Ordinatenwerte v_i errechnen sich nach (6.3). Für die Ermittlung des Gini-Koeffizienten G benötigt man noch die in (6.4) 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	13,45	0,42	0,42	0,014	0,42	0,176
2	6,48	0,91	1,33	0,045	1,82	0,828
3	8,13	6,48	7,81	0,266	19,44	41,990
4	0,91	8,13	15,94	0,542	32,52	66,097
5	0,42	13,45	29,39	1,0	67,25	180,902
Summe			$p_5 = 29{,}39$		$q_5 = 121{,}45$	289,993

Tab. 6.1: *Berechnung des Gini-Koeffizienten (Stromverbrauchsdaten)*

Stellt man sich, analog zu Beispiel 5.1, 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 $(u_2; v_2) = (0{,}4; 0{,}045)$ der Lorenzkurve, dass 40 % der Gruppe (die beiden Personen aus Indien und China mit dem niedrigsten Stromverbrauch) nur für etwa 4,5 % des Gesamtstromverbrauchs der Gruppe verantwortlich sind, d. h. die restlichen 60 % der Gruppe verbrauchen 95,5 %. Entsprechend

lässt sich aus $(u_4; v_4) = (0{,}8; 0{,}542)$ ableiten, dass die USA allein bereits 45,8 % des Gesamtstromverbrauchs verursachen. Für den normierten Gini-Koeffizienten G^* sollte man also hier einen Wert erwarten, der eine deutliche Merkmalskonzentration beinhaltet. In der Tat ergibt sich mit (6.5) und den Werten p_5 und q_5 aus Tabelle 6.1

$$G = \frac{1}{5}\left(\frac{2 \cdot 121{,}45}{29{,}39} - 1\right) - 1 \approx 0{,}453$$

und hieraus nach (6.7)

$$G^* = \frac{5}{4} \cdot G \approx 0{,}566.$$

Aufgabe 6.1-2

Auch für den Herfindahl-Index erhält man nach (6.9) einen Wert, der auf eine nennenswerte Konzentration verweist:

$$H = \frac{1}{29{,}39^2} \cdot 289{,}993 \approx 0{,}336.$$

Exkurs 6.1: Messung und Bewertung von Einkommensungleichheit

Die *Weltbank* veröffentlicht zur Charakterisierung von Einkommensungleichheit in den Ländern Listen mit Gini-Koeffizienten. Für 2010 waren die Werte für die skandinavischen Länder durchweg niedrig (im Bereich von 0,26 bis 0,27) und auch Deutschland hatte mit ca. 0,31 noch einen vergleichsweise niedrigen Gini-Koeffizienten. Werte über 0,50 wurden hingegen für etliche afrikanische Staaten beobachtet (z. B. Zambia 0,57 oder Lesotho 0,54), während die Werte für China und die USA im oberen Mittelfeld lagen (Bereich von 0,41 bis 0,42). Auch von den *Vereinten Nationen* und von der *Central Intelligence Agency (CIA)* der USA werden Länderlisten mit Gini-Koeffizienten veröffentlicht, von der CIA im Rahmen des von ihr herausgegebenen *World Factbook*. Neben den Gini-Koeffizienten werden 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 (2009) 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 Schulabbrecher 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 – z. B. Länder in Skandinavien. Auch in

Deutschland wird das Thema „Einkommensungleichheit“ neuerdings stärker diskutiert, z. B. in der Wochenzeitschrift *Die Zeit* vom 19. August 2011 oder in der *Süddeutschen Zeitung* vom 6. September 2013.

7 Indikatoren

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.

Vorschau auf das Kapitel

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.

7.1 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 vergliche, 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.

Arten von Verhältniszahlen

Sehr anschauliche Verhältniszahlen sind die in Abschnitt 4.1 bereits ausführlicher 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 ist der Anteil p der im SS 2015 an der Fakultät „Kultur- und Sozialwissenschaften“ der FernUniversität Hagen eingeschriebenen Studierenden (Teilgesamtheit) an der Zahl aller im SS 2015 in Hagen eingeschriebenen Studierenden (Grundgesamtheit). Auch die Erwerbslosenquote p ist eine Gliederungszahl; sie verknüpft die Anzahl der Erwerbslosen mit der Anzahl aller Personen im

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_7

erwerbsfähigen Alter. Eine Gliederungszahl p wird meist als Prozentwert ausgewiesen (Multiplikation mit 100).

EZB-Grafiken „Staatsschulden in der EU“

Es gibt Verhältniszahlen, die durch Quotientenbildung eine Verbindung zwischen *zwei* unterschiedlichen *Merkmalen* 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) oder die Verschuldung eines EU-Mitgliedstaats (Maßzahl: Euro / BIP oder Euro / Einwohner).

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. momentan Preise für den privaten Verbrauch auf das Jahr 2010. Der Preis x_t für Diesel-Kraftstoff im Jahr $t = 2013$ wird also nicht direkt, sondern in Form des Quotienten $I_t := \frac{x_t}{x_0}$ ausgewiesen, wobei x_0 den Preis im Referenzjahr 2010 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.

Erfassung komplexer Entwicklungen anhand von Indikatoren

Geeignete Maß- und Verhältniszahlen werden oft als **Indikatoren** herangezogen, um komplexe Entwicklungen, etwa die Veränderung von objektiven Lebensbedingungen und subjektivem Wohlbefinden oder von 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:

Interaktives Lernobjekt „Erwerbstätigkeit“

- *Gesundheit*: Lebenserwartung Neugeborener, Anteil von Personen mit Fettleibigkeit, Anteil der Gesundheitskosten am BIP; Ärztedichte; Anzahl der HIV-Fälle pro Million Einwohner;
- *Wohlstand*: BIP pro Kopf; Bruttoeinkommen von Arbeitnehmern pro Stunde; Erwerbstätigenquote; Anteil der nach amtlicher Definition als „arm“ geltenden Personen;
- *Bildung*: Abiturientenquote eines Jahrgangs; Anteil der Ausgaben für öffentliche und private Bildungseinrichtungen am BIP;
- *Umwelt*: Anteil erneuerbarer Energien am Primärenergieverbrauch; Treibhausemissionen in CO_2-Äquivalenten; Energieproduktivität (BIP / Primärenergieverbrauch);

- *Öffentliche Sicherheit*: Polizeidichte; Aufklärungsquote bei Gewaltkriminalität; inhaftierter Bevölkerungsanteil;
- *Innovationskraft*: Anzahl der Patente pro Einwohner; Bevölkerungsanteil mit Hochschulabschluss; Anteil der Staatsausgaben für Forschung und Entwicklung.

Wo findet man Informationen über Indikatoren?

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“.

Erstellung von Ranglisten

In den Medien werden Maß- oder Verhältniszahlen häufig im Zusammenhang mit Vergleichen zwischen Ländern, Regionen oder auch Institutionen herangezogen. In der Wochenzeitschrift *The Economist* wurden z. B. in einem Artikel vom 4. April 2015 die Militärausgaben von Großbritannien im Jahr 2013 mit denen einiger anderer Staaten verglichen und zwar in absoluten Zahlen und zusätzlich als Anteil des Bruttoinlandsprodukts des betreffenden Jahres. Dabei resultierten ganz unterschiedliche Ranglisten. Die verwendeten Daten stammten von *SIPRI* (*S*tockholm *I*nternational *P*eace *R*esearch *I*nstitute), einem unabhängigen Institut für Friedensforschung und Rüstungskontrolle, dessen Zahlen bei Abrüstungsverhandlungen eine wichtige Rolle spielten.

Bei *SIPRI* findet man neben den in US-Dollar ausgewiesenen absoluten Werten für Militärausgaben und deren Anteil am BIP auch Daten zur Beziehungszahl „Militärausgaben pro Kopf“. Nimmt man diese als Vergleichsbasis, erhält man eine weitere Rangliste. Man muss daher bei Ranglisten aller Art schauen, auf welcher Vergleichsbasis sie fußen, ob der verwendete Indikator sachadäquat ist und ob eventuell die parallele Verwendung mehrerer Indikatoren ein differenzierteres Bild liefert.

Beispiel 7.1: Vergleich der Militärausgaben von Ländern

Tabelle 7.1 weist für zwölf Länder die von *SIPRI* veröffentlichten Militärausgaben für 2014 aus. Die Tabelle zeigt in den letzten drei Spalten die absoluten Werte für Militärausgaben (in Milliarden US-Dollar, Wechselkurs im Referenzjahr), den Anteil dieser Ausgaben am BIP (in Prozent) sowie die Ausgaben pro Kopf (in vollen US-Dollar). Die Länder sind in der Tabelle nach absteigender Größe der absoluten Werte für die Miltärausgaben geordnet. Ordnet

man hingegen nach dem BIP-Anteil oder den Pro-Kopf-Ausgaben, ergeben sich andere Rangfolgen. Bei diesen ist in Tabelle 7.1 der erste Rangplatz jeweils durch normalen, der letzte Rangplatz durch kursiven Fettdruck betont.

Rang	Nation		Militärausgaben absolut (US-Dollar)	in % des BIP	pro Kopf
1.		USA	609,9	3,5	1891
2.		China	216,4	2,1	155
3.		Russland	84,5	4,5	593
4.		Saudi-Arabien	80,8	**10,4**	**2747**
5.		Frankreich	62,3	2,2	964
6		Großbritannien	60,5	2,2	952
7.		Indien	50,0	2,4	***39***
8.		Deutschland	46,5	1,2	562
9.		Japan	45,8	***1,0***	360
10.		Brasilien	31,7	1,5	157
11.		Israel	15,9	5,2	2040
12.		Singapur	9,8	3,7	1789

Tab. 7.1: *Militärausgaben für* 2014 *im Ländervergleich (Quelle: SIPRI; Datenextraktion: April* 2015*)*

Aufgabe 7.1

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, anders als in der vorletzten Spalte von Tabelle 7.1, mit dieser verknüpft.

7.2 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 7.2: Medaillenspiegel bei der Olympiade

Tabelle 7.2 zeigt die ersten zehn Platzierungen beim offiziellen Medaillenspiegel der Sommerolympiade 2008. Dieser orientiert sich primär an der *Anzahl der Goldmedaillen*; nur bei Gleichstand wirkt sich die Anzahl der Silber- und Bronzemedaillen auf die Platzierung aus.

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. 7.2: *Offizieller Medaillenspiegel der Sommerolympiade* 2008 *(Auszug)*

Der offizielle Medaillenspiegel der Olympiade 2008 erschien in allen europäischen Zeitungen. Er wurde nach Abschluss der Olympiade in verschiedenen Internet-Foren kontrovers diskutiert. Es kursierten mehrere alternative Varianten. Im *Guardian* und auch in der *Süddeutschen Zeitung* erschienen Beiträge, die die Fragwürdigkeit des offiziellen Rankings thematisierten. In amerikanischen Zeitungen, z. B. in der *New York Times* fand man von Anfang an einen anderen Medaillenspiegel, bei dem die *Gesamtzahl der Medaillen* als Indikator für den sportlichen Erfolg einer Nation fungierte. Die bei diesem Ansatz resultierenden ersten zehn Platzierungen sind in Tabelle 7.3 wiedergegeben.

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

Tab. 7.3: *US-amerikanischer Medaillenspiegel der Sommerolympiade* 2008

Es wurden weitere Versionen des Medaillenspiegels vorgeschlagen, etwa eine Bewertung nach der *Anzahl der Goldmedaillen pro Kopf*. Bei Verwendung dieses Ansatzes lagen Jamaica und Bahrain ganz vorne und die führenden Länder der amtlichen Liste rückten auf weit hinten liegende Plätze. Es gab Vorschläge, auch die Wirtschaftskraft eines Landes einzubeziehen, weil diese die Trainingschancen von Sportlern beeinflussen kann.

Aufgabe 7.2

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. Hier wäre zu klären, wie die Gewichte festgelegt werden sollen. 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 – Anfang 2015 war es noch das Jahr 2010 – für die der Index auf 100 gesetzt ist.

Beispiel 7.3: Der amtliche Verbraucherpreisindex

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.

Flash-Animation „Warenkorb“

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 7.1 ist es die Güterklasse „Brennstoffe“.

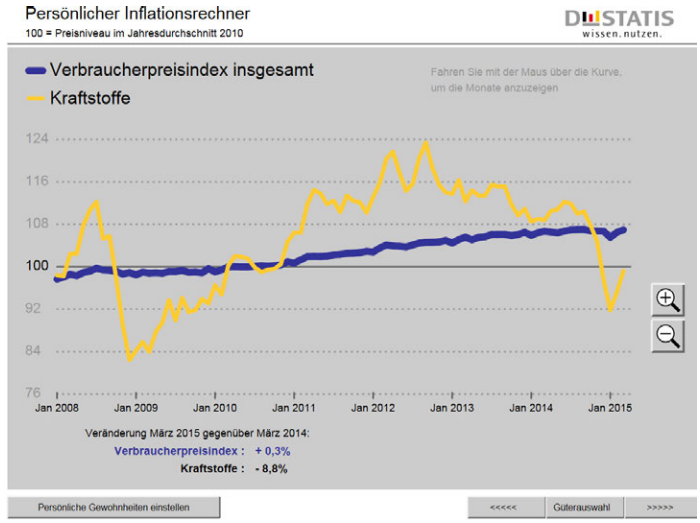


Abb. 7.1: *Inflationsrechner (Statistisches Bundesamt; März* 2015*)*

Fährt man mit der Maus 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. Wenn man die – in Abbildung 7.1 nicht wiedergegebene – Preisentwicklung bei Pauschalreisen visualisiert, sieht man deutliche zyklische Schwankungen. Die höchsten Werte werden in Perioden mit hoher Nachfrage beobachtet, z. B. am Jahresende.

Inflationsrechner

Aufgabe 7.3

Abbildung 7.1 zeigt neben dem allgemeinen Verbraucherpreisindex die Preisentwicklung für PKW-Kraftstoffe. Auffälligkeiten beim Entwicklungspfad bei den Kraftstoffen sind vor allem auf Schwankungen der Preise für Rohöl zurückzuführen. Mit Beginn der Finanzkrise im Sommer 2008 gingen industrielles Wachstum und damit auch die Ölnachfrage zurück. Dies erklärt den scharfen Preisrückgang für Kraftstoffe zu diesem Zeitpunkt.

Preiskaleidoskop

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.

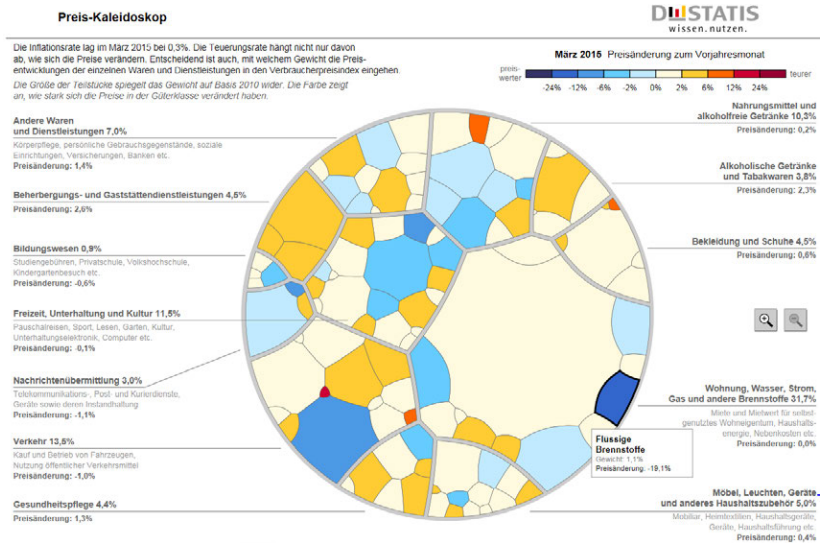


Abb. 7.2: *Preiskaleidoskop (Statistisches Bundesamt; März* 2015*)*

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 7.2 ist die Komponente „Flüssige Brennstoffe“ der Ausgabengruppe „Wohnung, Wasser, Strom, Gas und andere Brennstoffe“ betont. Die Ausgaben hierfür gingen mit einem Gewicht von 1,1 % in den Warenkorb ein und lagen etwa 19,1 % unter dem Vorjahresniveau.

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 das **European Innovation Scoreboard** der Generaldirektion „Unternehmen und Industrie“ der EU-Kommission sowie der **Global Innovation Index**, an dem u. a. die US-amerikanische Cornell University und das ebenfalls zur EU-Kommission gehörende Joint Research Center in Ispra beteiligt sind. Die beiden letztgenannten zusammengesetzten Indexzahlen vermitteln Informationen über europäische Länder bezüglich der Verwendung moderner Kommunikationstechnologien in Geschäftsprozessen bzw. zur Innovationskraft der Länder.

Global Innovation Index, Ranking 2013

Vor- und Nachteile zusammengesetzter 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 Indikatoren mit gleichem Gewicht verknüpft, weil man keine Informationen

Kritische Anmerkung zu Ranglisten

hat, die eine unterschiedliche Gewichtung motivieren. Es ist nachvollziehbar, dass Eurostat, das Europäische Amt für Statistik, der Verwendung von zusammengesetzten Indikatoren, bei den sich das Gewichtungsschema nicht – wie beim Verbraucherpreisindex – auf natürliche Weise aus Daten ergibt, eher zurückhaltend gegenüber steht, obwohl andere Generaldirektionen der EU-Kommission und auch mehrere supranationale Institutionen (OECD, UN) sie breit einsetzen. 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.

Exkurs 7.1: Der Human Development Index der UN

Video der UN zum HDI-Report 2013

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. Die Veröffentlichung der Werte erfolgt im Rahmen der *Human Development Reports* der Vereinten Nationen. Aufgrund einer 2010 erfolgten Änderung der Operationalisierung des HDI-Indexes sind ältere und aktuelle HDI-Werte nicht direkt vergleichbar.

Tabelle 7.4 zeigt für das Jahr 2013 die besten sechs HDI-Werte und die beiden niedrigsten Werte. Die besten und schlechtesten Werte für die vier Indikatoren, aus denen sich der HDI zusammensetzt, sind ebenfalls ausgewiesen. 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 der Niederlande und der USA – oft kaum unterscheiden. Kleinste Messfehler oder minimale Veränderungen des Gewichtungsschemas können eine andere Rangfolge liefern.

	HDI (Gesamtindex)		**Gesundheit** (Indikator *Lebenserwartung*)
1.	Norwegen (0,944)	1.	Japan (83,6)
2.	Australien (0,933)	2.	Hongkong, China (83,4)
3.	Schweiz (0,917)	3.	Schweiz (82,6)
4.	Niederlande (0,915)	4.	Australien (82,5)
5.	USA (0,914)	5.	Italien (82,4)
6.	Deutschland (0,911)	6.	Singapur (82,3)
	⋮		⋮
186.	Demokr. Rep. Kongo (0,338)	186.	Swaziland (49,0)
187.	Niger (0,337)	187.	Sierra Leone (45,6)

	Bildungsstand (Indikatoren B1 / B2)		**Lebensstandard** (Indikator „Kaufkraft“)
1.	USA (12,9) / Australien (19,9)	1.	Qatar (119029)
2.	Deutschland (12,9) / Neuseeland (19,4)	2.	Liechtenstein (87085)
3.	Australien (12,5) / Island (18,7)	3.	Kuwait (85820)
4.	Norwegen (12,6) / Irland (18,6)	4.	Singapur (72371)
5.	Neuseeland (12,5) / Niederlande (17,9)	5.	Brunei (70883)
6.	Israel (12,5) / Norwegen (17,7)	6.	Norwegen (63909)
	⋮		⋮
186.	Niger (1,4) / Niger (5,4)	186.	Zentralafrik. Republik (588)
187.	Burkina Faso (1,3) / Eritrea (4,1)	187.	Demokr. Rep. Kongo (444)

Tab. 7.4: *HDI-Werte und Sub-Indikatoren ausgewählter Länder für* 2013 *(Quelle: Human Development Report der UN,* 2014*)*

Sind unterschiedliche Indikatorensysteme vergleichbar?

Welche Indikatoren oder Indikatorensysteme für die Erfassung einer gesellschaftsrelevanten Dimension, etwa „Wirtschaftswachstum“, besonders aussagekräftig sind, ist nicht immer leicht zu beantworten. In der Wochenzeitung *Die Zeit* vom 28. März 2009 wurden z. B. 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. So unterscheiden sich der Well-Being-Index des Statistischen Amts von Großbritannien und der von den amerikanischen Firmen Gallup und Heal-

thways geführte Well-Being-Index u. a. hinsichtlich der Sub-Indikatoren, die in den Index eingehen. Manchmal ändern sich auch Operationalisierungen von Variablen. Die seit September 2014 angewendete Neufassung des Europäischen Systems Volkswirtschaftlicher Gesamtrechnungen (ESVG) – Ersatz des bisherigen ESVG 1995 durch das ESVG 2010 – hat zur Folge, dass das BIP um ca. 3 % allein aufgrund geänderter Messvorschriften steigt.

Exkurs 7.2: 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 jeweils neuesten Ergebnisse werden anhand einer interaktiven Karte zugänglich macht.

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. Auch außerhalb der amtlichen Statistik gibt es 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, Deutschland eingeschlossen. 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 im Glückatlas 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 geringen Stichprobenumfänge. Da die regionalen Zufriedenheitsunterschiede sehr klein sind, könnte schon der unvermeidliche *Stichprobenfehler* das Ranking determinieren.

Erwähnt sei auch der *World Values Survey*, der Wohlfahrt, subjektives Wohlbefinden und soziokulturelle Wertemuster zu erfassen sucht und sich auf persönliche Interviews stützt (mindestens 1000 pro Land).

---

8 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. aller Werte 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.

8.1 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

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_8

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 $a_1, \ldots, a_k$, das Merkmal Y die Ausprägungen $b_1, \ldots, b_m$ auf. Die Merkmalswerte $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ repräsentieren eine **bivariate Urliste**. Diese lässt sich z. B. in der Form $(x_1, y_1), \ldots, (x_n, y_n)$ 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

$$h_{ij} := h(a_i, b_j) \qquad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m \tag{8.1}$$

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

$$f_{ij} := f(a_i, b_j) \qquad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, m \tag{8.2}$$

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 besonders übersichtlich in tabellarischer Form wiedergeben. Die resultierende Tabelle heißt **Kontingenztafel** oder **Kontingenztabelle**, gelegentlich auch **Kreuztabelle**. Sie definiert die gemeinsame **empirische Verteilung** der beiden Merkmale.

Tabellen für bivariate Häufigkeits -verteilungen

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. Nachstehend ist diese für den Fall absoluter Häufigkeiten wiedergegeben. 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. 8.1: $(k \times m)$-*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

$$h_{i\cdot} := h_{i1} + h_{i2} + \ldots + h_{im} = \sum_{j=1}^{m} h_{ij} \qquad i = 1, 2, \ldots, k \tag{8.3}$$

und analog bei relativen Häufigkeiten die Summen

$$f_{i\cdot} := f_{i1} + f_{i2} + \ldots + f_{im} = \sum_{j=1}^{m} f_{ij} \qquad i = 1, 2, \ldots, k \tag{8.4}$$

ausweist (lies: *h-i-Punkt* resp. *f-i-Punkt*). Die Summe (8.3) bzw. (8.4) entspricht der absoluten bzw. relativen Häufigkeit derjenigen Merkmalskombinationen, bei denen X die Ausprägung a_i und Y eine beliebige der m Ausprägungen $b_1, \ldots, b_m$ hat. Letzteres bedeutet, dass Y nicht berücksichtigt wird. Die Häufigkeiten $h_{1\cdot}, h_{2\cdot}, \ldots, h_{k\cdot}$ werden **absolute Randhäufigkeiten** von X genannt, die Häufigkeiten $f_{1\cdot}, f_{2\cdot}, \ldots, f_{k\cdot}$ **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

$$h_{\cdot j} := h_{1j} + h_{2j} + \ldots + h_{kj} = \sum_{i=1}^{k} h_{ij} \qquad j = 1, 2, \ldots, m \tag{8.5}$$

resp.

$$f_{\cdot j} := f_{1j} + f_{2j} + \ldots + f_{kj} = \sum_{i=1}^{k} f_{ij} \qquad j = 1, 2, \ldots, m. \tag{8.6}$$

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

Aufgabe 8.1

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

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

Randverteilung von Y

Tab. 8.2: *Vollständige* $(k \times m)$*-Kontingenztafel für absolute Häufigkeiten*

Durch die Randverteilungen wird eine Verbindung zwischen uni- und bivariaten Häufigkeitsverteilungen hergestellt. Aus den gemeinsamen Häufigkeiten (8.1) bzw. (8.2) zweier Merkmale X und Y lassen sich stets gemäß (8.3) und (8.5) bzw. (8.4) und (8.6) 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 8.1: Geschlechtspezifische Ergebnisse beim Politbarometer

Flash-Animation „Kontingenztafel (Politbarometer)“

In Tabelle 4.1 waren Ergebnisse des ZDF-Politbarometers vom 16. Oktober 2009 für 1021 befragte Personen wiedergegeben. Tabelle 8.3 zeigt die Häufigkeiten zur „Sonntagsfrage“ erneut, nun differenziert nach Geschlecht. Im oberen Teil ist die absolute und darunter die relative Häufigkeitsverteilung für das Merkmal „Parteipräferenz“ wiedergegeben, jeweils mit Angabe der beiden Randverteilungen. 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 X			Ausprägungen von Y ♂ b_1	♀ b_2	Randverteilung von X
	CDU/CSU	a_1	179	204	383
	SPD	a_2	100	117	217
	FDP	a_3	80	59	139
	DIE LINKE.	a_4	67	50	117
	BÜNDNIS 90/DIE GRÜNEN	a_5	54	62	116
	Sonstige	a_6	21	28	49
			501	520	1021
			Randverteilung von Y		

Ausprägungen von X			Ausprägungen von Y ♂ b_1	♀ b_2	Randverteilung von X
	CDU/CSU	a_1	0,175	0,200	0,375
	SPD	a_2	0,098	0,115	0,213
	FDP	a_3	0,078	0,058	0,136
	DIE LINKE.	a_4	0,066	0,049	0,115
	BÜNDNIS 90/DIE GRÜNEN	a_5	0,053	0,061	0,114
	Sonstige	a_6	0,021	0,027	0,048
			0,491	0,509	1
			Randverteilung von Y		

Tab. 8.3: (6×2)-*Kontingenztafel für absolute und für relative Häufigkeiten*

Spezialfall: (2×2)-Kontingenztafel

Ein Spezialfall einer Kontingenztabelle ist die **Vierfeldertafel**, die sich für $k = m = 2$ ergibt und in Tabelle 8.4 für den Fall absoluter Häufigkeiten wiedergegeben ist. Vierfeldertafeln werden im Zusammenhang mit 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 „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. 8.4: *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 8.4 könnte man z. B. das folgende Baumdiagramm wählen:

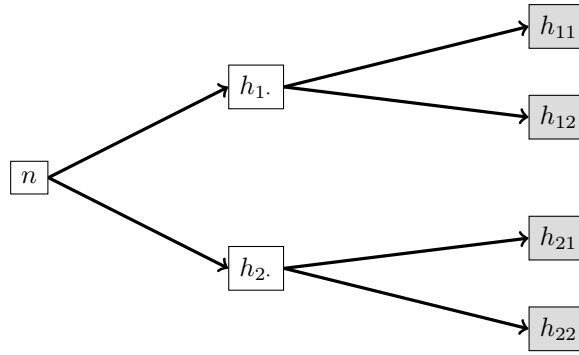


Abb. 8.1: *Baumdiagramm als Alternative zu einer Vierfeldertafel*

Beispiel 8.2: Baumdiagramm zu amtlichen Bevölkerungsdaten

In einer Pressemitteilung des Statistischen Bundesamtes vom 24. Juni 2004 hieß es, dass 51,1 % der damals mit 82,5 Millionen veranschlagten Bevölkerung Deutschlands Frauen sind. Ferner wurde mitgeteilt, dass der Anteil der Erwerbstätigen bei den Frauen bei 42,4 % und bei den Männern bei 55,3 % lag. Kinder sind hier jeweils einbezogen und der Kategorie „nicht erwerbstätig“ zugeordnet. Aus dieser Verlautbarung lässt sich z. B. nicht unmittelbar ablesen, wieviele Männer und Frauen ohne Erwerbstätigkeit waren.

Bevor man eine Vierfeldertafel für absolute Häufigkeiten ableitet, ist es hilfreich, die in der Pressenotiz enthaltene Substanz erst einmal in ein Baumdiagramm zu übertragen. Dieses ist in Abbildung 8.2 wiedergegeben, wobei die vom

Statistischen Bundesamt direkt kommunizierte Information durch Fettdruck betont ist.

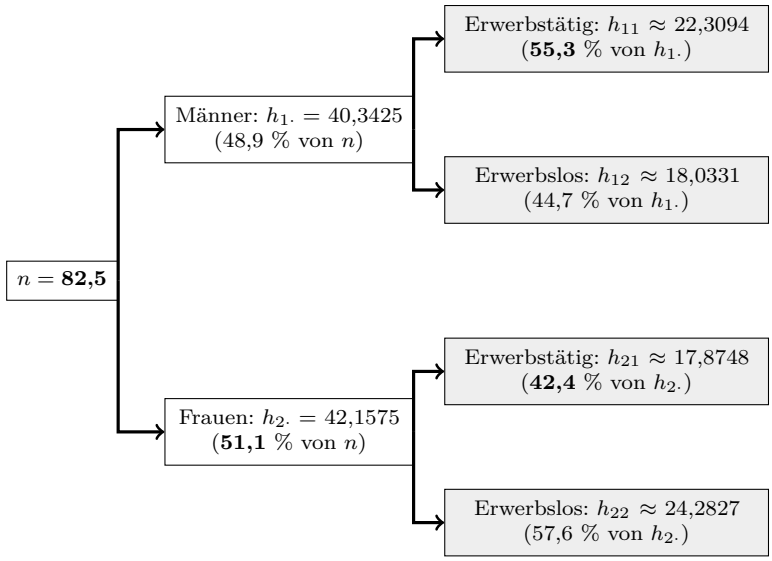


Abb. 8.2: *Baumdiagramm für Bevölkerungsdaten*

Hieraus ergibt sich dann für die beiden Merkmale bei Rundung auf volle Hunderter die in Tabelle 8.5 wiedergegebene Vierfeldertafel für absolute Häufigkeiten, bei der – anders als bei der Kontingenztabelle aus Beispiel 8.1 – die Ausprägungen des Merkmals „Geschlecht“ vertikal aufgelistet sind. Wenn man bei dem obigen Baumdiagramm auf der zweiten Ebene nach dem Erwerbsstatus und auf der dritten Ebene nach Geschlecht unterteilte und dann die Häufigkeiten wieder in eine Vierfeldertafel übertrüge, würden die Ausprägungen des Merkmals „Geschlecht“ im Tabellenkopf stehen.

	Erwerbstätige	Erwerbslose	Zeilensummen
Männer	22,3094	18,0331	40,3425
Frauen	17,8748	24,2827	42,1575
Spaltensummen	40,1842	42,3158	82,5

Tab. 8.5: *Vierfeldertafel für absolute Häufigkeiten*

8.2 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 16. Oktober 2009 insgesamt 11,5 % der Personen der Stichprobe weibliche Wähler mit SPD-Präferenz waren (117 von 1021 Personen), lässt sich noch keine Aussage über eine geschlechtsspezifische Präferenz dieser Partei gewinnen. Zur Herleitung einer solchen Aussage benötigt man auch die Information, wie oft die SPD insgesamt favorisiert wurde, d.h. wie groß die Teilmenge aller Befragten in der Stichprobe war, die sich für die SPD aussprach. Diese Information wird durch eine Randhäufigkeit vermittelt (hier: $h_{2\cdot} = 217$). Eine geeignete Verknüpfung der gemeinsamen Häufigkeiten für zwei diskrete Merkmale X und Y mit den Randhäufigkeiten führt zu **bedingten relativen Häufigkeiten**. Diese sind dann der Ausgangspunkt für die Untersuchung von Zusammenhängen zwischen zwei diskreten Merkmalen.

Ausprägung von Y

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

Randverteilung von Y

Tab. 8.6: *Absolute Häufigkeiten für die Ausprägungen von Y unter der Bedingung $X = a_i$*

Bedingte Häufigkeitsverteilung für Y

Um das Konzept der bedingten Häufigkeiten verständlich zu machen, ist in Tabelle 8.6 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.

Dividiert man nun jedes der m Elemente $h_{i1}, h_{i2}, \ldots, h_{im}$ durch die Randhäufigkeit $h_{i\cdot}$, so erhält man die relativen Häufigkeiten für das

Auftreten der Ausprägungen b_1, b_2, …, 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

$$f_Y(b_j|a_i) := \frac{h_{ij}}{h_{i\cdot}} \qquad j = 1, 2, \ldots, m. \tag{8.7}$$

Die m bedingten relativen Häufigkeiten $f_Y(b_1|a_i)$, $f_Y(b_2|a_i)$, …, $f_Y(b_m|a_i)$ definieren die **bedingte Häufigkeitsverteilung** für Y unter der Bedingung $X = a_i$.

Bedingte Häufigkeitsverteilung für X

Analog kann man, wie in Tabelle 8.7 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}, …,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.

Ausprägung von Y

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

Randverteilung von Y

Tab. 8.7: *Absolute Häufigkeiten für die Ausprägungen von X unter der Bedingung $Y = b_j$*

Teilt man jedes der k Elemente h_{1j}, h_{2j}, …, h_{kj} durch die Randhäufigkeit $h_{\cdot j}$, so erhält man die relativen Häufigkeiten für das Auftreten der Ausprägungen a_1, a_2, …, 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

$$f_X(a_i|b_j) := \frac{h_{ij}}{h_{\cdot j}} \qquad i = 1, 2, \ldots, k. \tag{8.8}$$

Die k bedingten relativen Häufigkeiten $f_X(a_1|b_j)$, $f_X(a_2|b_j)$, …, $f_X(a_k|b_j)$ konstituieren die **bedingte Häufigkeitsverteilung** für X unter der Bedingung $Y = b_j$.

Abbildung 8.3 fasst zusammen, wie man aus einer Kontingenztabelle für zwei diskrete oder gruppierte stetige Merkmale X und Y unter Verwendung der Randhäufigkeiten die bedingten Häufigkeitsverteilungen für beide Merkmale gewinnt. Eine bedingte Häufigkeitsverteilung resultiert, wenn man jedes Element einer Zeile oder Spalte einer Kontingenztabelle durch die zur jeweiligen Zeile oder Spalte gehörende Randhäufigkeit teilt. Bedingte Häufigkeitsverteilungen sind univariate Verteilungen, weil nur die Merkmalsausprägungen für ein einziges Merkmal variieren.

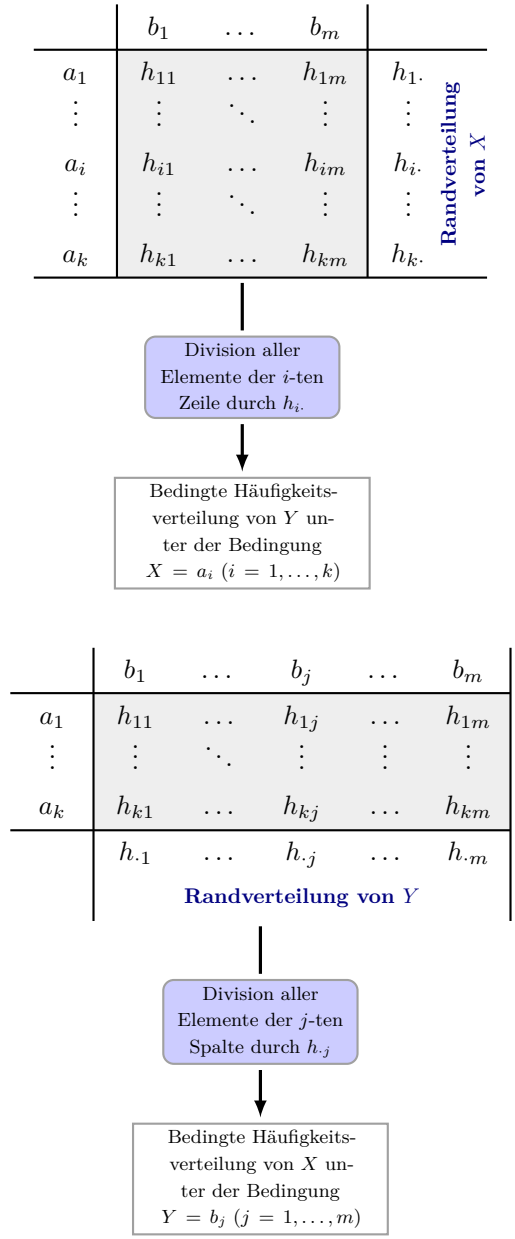


Abb. 8.3: *Bestimmung bedingter Häufigkeitsverteilungen*

Beispiel 8.3: Bedingte Häufigkeiten beim ZDF-Politbarometer

Die Bestimmung und Interpretation bedingter relativer Häufigkeiten sei anhand der Daten zum ZDF-Politbarometer illustriert. Bei diesem Datensatz ist z. B. die Frage von Interesse, ob zwischen der Parteipräferenz X und dem Geschlecht Y der befragten Personen in der Stichprobe ein Zusammenhang besteht.

Flash-Animation „Bedingte Häufigkeiten (Politbarometer)“

Ausprägungen von X			Ausprägungen von Y: ♂ b_1	♀ b_2	Randverteilung von X
	CDU/CSU	a_1	179	204	383
	SPD	a_2	100	117	217
	FDP	a_3	80	59	139
	DIE LINKE.	a_4	67	50	117
	BÜNDNIS 90/DIE GRÜNEN	a_5	54	62	116
	Sonstige	a_6	21	28	49
			501	520	1021
			Randverteilung von Y		

Ausprägungen von X			Ausprägungen von Y: ♂ b_1	♀ b_2	Randverteilung von X
	CDU/CSU	a_1	179	204	383
	SPD	a_2	100	117	217
	FDP	a_3	80	59	139
	DIE LINKE.	a_4	67	50	117
	BÜNDNIS 90/DIE GRÜNEN	a_5	54	62	116
	Sonstige	a_6	21	28	49
			501	520	1021
			Randverteilung von Y		

Tab. 8.8: *Berechnung bedingter Häufigkeiten (ZDF-Politbarometer)*

Tabelle 8.8 zeigt zweifach die Kontingenztafel für absolute Häufigkeiten aus Beispiel 8.1. 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 8.8 betont, dass von den $n = 1021$ Personen der Stichprobe insgesamt 501 Befragte männlich waren und dass innerhalb dieser Teilstichprobe 179 Männer die CDU/CSU, 100 die SPD, 80 die FDP, 67 die Linken, 54 die Grünen und 21 sonstige Parteien favorisiert hatten. Die bedingten relativen Häufigkeiten $f_X(a_1|b_1)$, $f_X(a_2|b_1)$, ..., $f_X(a_6|b_1)$, die die bedingte Häufigkeitsverteilung für X unter der Bedingung $Y = b_1$ definieren, bestimmen sich nach (8.8) als

$$f_X(a_1|b_1) = \frac{h_{11}}{h_{\cdot 1}} = \frac{179}{501} \approx 0{,}357 \qquad f_X(a_2|b_1) = \frac{h_{21}}{h_{\cdot 1}} = \frac{100}{501} \approx 0{,}200$$

$$f_X(a_3|b_1) = \frac{h_{31}}{h_{\cdot 1}} = \frac{80}{501} \approx 0{,}160 \qquad f_X(a_4|b_1) = \frac{h_{41}}{h_{\cdot 1}} = \frac{67}{501} \approx 0{,}138$$

$$f_X(a_5|b_1) = \frac{h_{51}}{h_{\cdot 1}} = \frac{54}{501} \approx 0{,}108 \qquad f_X(a_6|b_1) = \frac{h_{61}}{h_{\cdot 1}} = \frac{21}{501} \approx 0{,}042.$$

Aufgabe 8.2

Der Wert 0,357 sagt z. B. aus, dass von den Personen in der Stichprobe, die männlichen Geschlechts waren, ca. 35,7 % bei der „Sonntagsfrage“ vom 16. Oktober 2009 die CDU/CSU favorisiert hatten.

Im unteren Teil von Tabelle 8.8 ist hervorgehoben, dass von den $n = 1021$ Personen der Stichprobe 217 Personen die SPD favorisiert hatten, nämlich 100 Männer ($Y = b_1$) und 117 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 (8.7) als

$$f_Y(b_1|a_2) = \frac{h_{21}}{h_{2\cdot}} = \frac{100}{217} \approx 0{,}461 \qquad f_Y(b_2|a_2) = \frac{h_{22}}{h_{2\cdot}} = \frac{117}{217} \approx 0{,}539.$$

Das Ergebnis 0,539 beinhaltet, dass von den Personen in der Stichprobe, die sich für die SPD entschieden hatten, 53,9 % 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.

Flash-Animation „Empirische Unabhängigkeit“

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 $b_1, b_2, \ldots, b_m$ als Bedingung gewählt wird, d. h. die m bedingten Häufigkeitsverteilungen $f_X(a_1|b_j)$, $f_X(a_2|b_j)$, ..., $f_X(a_k|b_j)$ müssten übereinstimmen ($j = 1, 2, \ldots, m$). Insbesondere müssten dann die i-ten Elemente dieser m bedingten Verteilungen identisch sein, d. h. es würde bei Unabhängigkeit gelten

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

Äquivalent ist wegen (8.8) die Darstellung

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

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 $h_{i\cdot}$, während die zweite Summe mit dem Stichprobenumfang n übereinstimmt (vgl. Tabelle 8.6). Es gilt also bei Unabhängigkeit von X und Y für jede der gemeinsamen Häufigkeiten h_{ij} der Kontingenztafel

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

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

$$\widetilde{h}_{ij} := \frac{h_{i\cdot} \cdot h_{\cdot j}}{n} \tag{8.9}$$

(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

Formale Definition der Unabhängigkeit zweier Merkmale

$$h_{ij} \begin{cases} = \widetilde{h}_{ij} & \text{bei empirischer Unabhängigkeit der Merkmale} \\ \neq \widetilde{h}_{ij} & \text{bei empirischer Abhängigkeit der Merkmale} \end{cases} \tag{8.10}$$

gilt. Zwei Merkmale X und Y, deren gemeinsame Häufigkeitsverteilung durch Tabelle 8.2 gegeben ist, sind also genau dann unabhängig, wenn für jedes der $k \cdot m$ Elemente h_{ij} der Kontingenztafel $h_{ij} = \widetilde{h}_{ij}$ ist mit $\widetilde{h}_{ij}$ aus (8.9). 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 $\widetilde{h}_{ij}$ für die gemeinsamen Häufigkeiten sind nicht notwendigerweise ganzzahlig.

Die Aussage (8.10) 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 $h_{ij} - \widetilde{h}_{ij}$. Diese sind der Ausgangspunkt für die Konstruktion von Zusammenhangsmaßen für nominalskalierte Merkmale (s. Abschnitt 9.1).

Beispiel 8.4: Parteipräferenz und Geschlecht

Zur Illustration der Vorgehensweise bei der Untersuchung der empirischen Unabhängkeit bzw. Abhängigkeit von Merkmalen werde erneut die im oberen Teil von Tabelle 8.3 wiedergegebene Kontingenztafel für absolute Häufigkeiten herangezogen. Diese zeigte die Ergebnisse des Politbarometers vom 16. Oktober 2009 in Form der gemeinsamen Häufigkeiten für die Merkmale „Parteipräferenz X“ und „Geschlecht Y“. Interessant ist hier die Fragestellung, ob sich das Wählerverhalten von Frauen und Männern unterscheidet.

Flash-Animation „Parteipräferenz und Geschlecht“

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 (8.9) zu errechnenden Werten zu vergleichen, die bei empirischer Unabhängigkeit gelten müssten. Die vier Werte $h_{11} = 179$, $h_{12} = 204$, $h_{21} = 100$, $h_{22} = 117$ der ersten beiden Zeilen der Kontingenztafel sind also z. B. zu vergleichen mit

$$\widetilde{h}_{11} = \frac{h_{1\cdot} \cdot h_{\cdot 1}}{n} = \frac{383 \cdot 501}{1021} \approx 187{,}9 \qquad \widetilde{h}_{12} = \frac{h_{1\cdot} \cdot h_{\cdot 2}}{n} = \frac{383 \cdot 520}{1021} \approx 195{,}1$$

$$\widetilde{h}_{21} = \frac{h_{2\cdot} \cdot h_{\cdot 1}}{n} = \frac{217 \cdot 501}{1021} \approx 106{,}5 \qquad \widetilde{h}_{22} = \frac{h_{2\cdot} \cdot h_{\cdot 2}}{n} = \frac{217 \cdot 520}{1021} \approx 110{,}5.$$

Die anderen 8 Werte $\widetilde{h}_{ij}$ sind analog zu bestimmen. Man erhält, wenn man wieder auf eine Dezimalstelle rundet

$$\widetilde{h}_{31} \approx 68{,}2 \qquad \widetilde{h}_{32} \approx 70{,}8 \qquad \widetilde{h}_{41} \approx 57{,}4 \qquad \widetilde{h}_{42} \approx 59{,}6$$
$$\widetilde{h}_{51} \approx 56{,}9 \qquad \widetilde{h}_{52} \approx 59{,}1 \qquad \widetilde{h}_{61} \approx 24{,}0 \qquad \widetilde{h}_{62} \approx 25{,}0.$$

In Tabelle 8.9 sind die beobachteten absoluten Häufigkeiten h_{ij} (auf grauem Raster) und die bei Unabhängigkeit zu erwartenden fiktiven Werte $\widetilde{h}_{ij}$ nebeneinander gestellt.

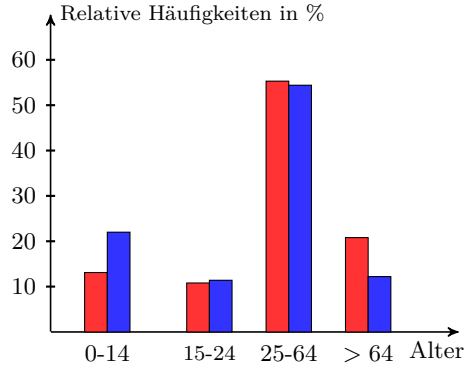
			Ausprägungen von Y		**Ausprägungen von Y**	
			♂ b_1	♀ b_2	♂ b_1	♀ b_2
Ausprägungen von X	CDU/CSU	a_1	179	204	187,9	195,1
	SPD	a_2	100	117	106,5	110,5
	FDP	a_3	80	59	68,2	70,8
	DIE LINKE.	a_4	67	50	57,4	59,6
	Bündnis 90/Die Grünen	a_5	54	62	56,9	59,1
	Sonstige	a_6	21	28	24,0	25,0

Tab. 8.9: *Absolute Häufigkeiten (Politbarometer) – beobachtete Werte h_{ij} (gerasterter Teil) und Werte $\widetilde{h}_{ij}$ bei empirischer Unabhängigkeit*

Man erkennt, dass die sich entsprechenden Werte h_{ij} und $\widetilde{h}_{ij}$ zwar nicht extrem, aber doch in nicht vernachlässigbarem Umfang differieren – es ist z. B. $h_{32} = 59$ und $\widetilde{h}_{32} = 70{,}8$. Die Daten sprechen *nicht* für eine empirische Unabhängigkeit der beiden Merkmale „Parteipräferenz X“ und „Geschlecht Y“.

Visualisierung empirischer Verteilungen zweier diskreter Merkmale

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 8.4 visualisiert auf der Basis von Eurostat-Daten für 2014 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“.



Altersklasse	DE	IE
0-14	13,1	22,0
15-24	10,8	11,4
25-64	55,3	54,4
≥ 65	20,8	12,2

Abb. 8.4: *Bevölkerungsstrukturen 2014 in Deutschland und Irland (Doppel-Säulendiagramm; linke Teilbalken: Deutschland)*

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 13,1 % in Deutschland viel niedriger (Irland: 22,0 %) und der Anteil der über 64-jährigen mit 20,8 % viel höher lag (Irland: 12,2 %).

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.

8.3 Empirische Verteilungen stetiger Merkmale

Wenn man an n Untersuchungseinheiten die Ausprägungen zweier *stetiger* Merkmale X und Y ermittelt, wird man bei der resultierenden **bivariaten Urliste** $(x_1, y_1), \ldots, (x_n, y_n)$ 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.

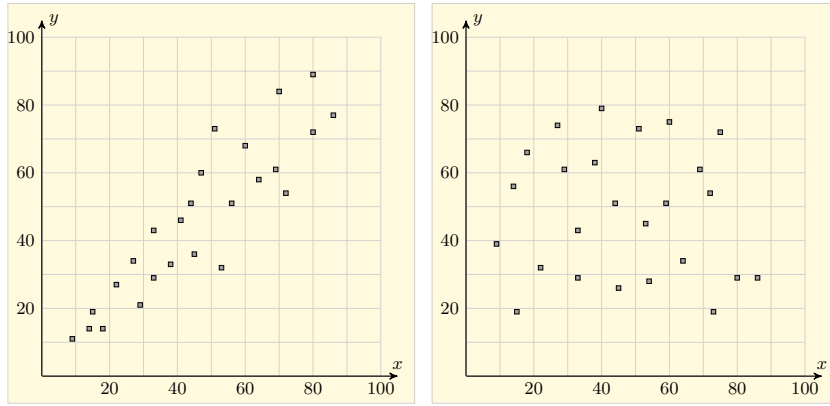


Abb. 8.5: *Zwei Streudiagramme*

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 $(x_1, y_1), \ldots, (x_n, y_n)$ durch Punkte repräsentiert. Dabei resultiert ein **Streudiagramm**. Abbildung 8.5 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 9.2 abgeleitet.

Wenn in dem betrachteten Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$ Punkte mehrfach auftreten, wird dies in einem klassischen Streudiagramm nicht sicht-

bar. Man kann dies aber z. B. sichtbar machen, indem man die mehrfach belegten Punkte größer gestaltet. Die Punkte des Streudiagramms sind dann Kreise mit Mittelpunkt (x_i, y_i), deren Flächeninhalt z_i davon abhängt, wie oft der jeweilige Punkt im Datensatz auftritt. Die resultierende Grafik wird als **Blasendiagramm** (engl.: *bubble chart*) bezeichnet. Dieses ist hier durch Kreise definiert, deren Mittelpunkte (x_i, y_i) Ausprägungen stetiger Merkmale X und Y sind und deren Fläche z_i durch eine dritte Variable Z determiniert ist. Die Variable Z kann diskret sein und die Anzahl widerspiegeln, mit der jedes Wertepaar (x_i, y_i) auftritt. Blasendiagramme lassen sich aber auch für Datensätze $(x_1, y_1), \ldots, (x_n, y_n)$ heranziehen, bei denen kein Wert mehrfach auftritt. Die Variable Z kann dabei ein weiteres stetiges Merkmal repräsentieren.

Blasendiagramme lassen sich auch zur Veranschaulichung univariater Häufigkeitsverteilungen verwenden – dort ist aber die Position der Kreismittelpunkte nicht durch Vorgaben determiniert.

Beispiel 8.5: 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 2014. 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 2014 erzielte Umsatz (in Millionen Euro). Stellt man die Werte $(x_1, y_1), \ldots, (x_4, y_4)$ als Punkte dar und zeichnet um die

	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. 8.10: *Marktanteil, Veränderungsrate und Umsatz für vier Tablets*

Punkte Kreise, deren Fläche zum Wert des Merkmals „Umsatz Z“ proportional ist, resultiert ein Blasendiagramm. Das in Abbildung 8.6 wiedergegebene Diagramm kann für die Sortimentsplanung des kommenden Jahres herangezogen werden. Man sieht z. B., dass das Tablet C im Jahr 2014 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.

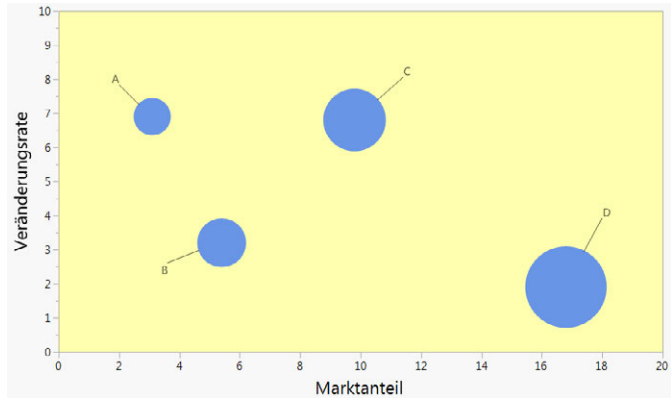


Abb. 8.6: *Analyse eines Produktportfolios anhand eines Blasendiagramms*

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

Wenn man ein solches Blasendiagramm für mehrere aufeinanderfolgende Jahre t_0, t_1, ... erstellt und diese in einer Animation hintereinander schaltet, wird sogar noch 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 $t = 2013$. Zur Erstellung dynamischer Blasendiagramme kann man z. B. *JMP* oder *R* heranziehen.

Exkurs 8.1: Big Data

Wir sind heute mit einer stetig wachsender Flut von Daten konfrontiert, die aus ganz unterschiedlichen Quellen stammen – 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, meist unter dem Etikett **„Big Data“** firmierend, 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 (2013) 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).

9 Zusammenhangsmaße

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.

Vorschau auf das Kapitel

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.

9.1 Nominalskalierte Merkmale

In Abschnitt 8.2 wurde mit (8.10) 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 $\widetilde{h}_{ij}$ aus (8.9) 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 $h_{ij} - \widetilde{h}_{ij}$ 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 $\widetilde{h}_{ij}$ bezogen, d.h. man bildet die Summe der $k \times m$ Terme $\frac{(h_{ij}-\widetilde{h}_{ij})^2}{\widetilde{h}_{ij}}$.

Ein nicht-normiertes Zusammenhangsmaß

Wenn man diese Terme analog zu Tabelle 8.1 (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 $\frac{(h_{ij}-\widetilde{h}_{ij})^2}{\widetilde{h}_{ij}}$ in jeder der k Zeilen addiert und dann die k Zeilensummen aufsummiert.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_9

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

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

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]

$$\chi^2 := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \widetilde{h}_{ij})^2}{\widetilde{h}_{ij}}. \tag{9.1}$$

Flash-Animation „χ^2-Koeffizient“

Der χ^2-Koeffizient ist ein **Zusammenhangsmaß** für zwei nominalskalierte Merkmale, das nach Konstruktion Null ist, wenn die Merkmale empirisch unabhängig sind. Bei einem schwachen Merkmalszusammenhang nimmt (9.1) 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

$$0 \leq \chi^2 \leq \chi^2_{max} = n \cdot (M - 1) \qquad M := \min(k; m). \tag{9.2}$$

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 (9.1) abgeleitetes Zusammenhangsmaß, dessen Wert nicht mehr von n abhängt, ist der durch

$$\Phi := \sqrt{\frac{\chi^2}{n}} \tag{9.3}$$

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

[1] Das Zusammenhangsmaß (9.1) wird in der induktiven Statistik u. a. verwendet, um Hypothesen über Merkmalszusammenhänge zu testen (sog. χ^2-Unabhängigkeitstest; vgl. (15.33)).

oben beschränkt, d. h. es gilt mit M aus (9.2):

$$0 \leq \Phi \leq \Phi_{max} := \sqrt{M-1}. \tag{9.4}$$

Harald Cramér

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

$$V := \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \cdot (M-1)}}. \tag{9.5}$$

Flash-Animation „Cramér's V“

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

$$0 \leq V \leq 1. \tag{9.6}$$

Mit (9.6) 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 9.1: Parteipräferenz und Geschlecht

Auf der Basis der (6×2)-Kontingenztabelle mit den Daten vom Politbarometer vom 16. Oktober 2009 wurde in Beispiel 8 4 festgestellt (vgl. Tabelle 8.9), 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.

Flash-Animation „Parteipräferenz und Geschlecht (Zusammenhangsmessung)“

Zur Quantifizierung der Zusammenhangsstärke lassen sich nun die Zusammenhangsmaße (9.1), (9.3) und (9.5) heranziehen. Die Berechnung des χ^2-Koeffizienten (9.1) besteht hier aus der Bestimmung von $6 \cdot 2 = 12$ Termen $\frac{(h_{ij}-\widetilde{h}_{ij})^2}{\widetilde{h}_{ij}}$. Für den ersten Term errechnet man z. B.

$$\frac{(h_{11}-\widetilde{h}_{11})^2}{\widetilde{h}_{11}} = \frac{(179-187{,}9)^2}{187{,}9} \approx 0{,}42.$$

Analog ermittelt man unter Rückgriff auf die in Beispiel 8.3 bestimmten Werte $\widetilde{h}_{ij}$ die übrigen 11 Terme. Man erhält bei Rundung auf 2 Dezimalstellen[2]

$$\chi^2 \approx 0{,}42 + 0{,}40 + 2{,}04 + 1{,}61 + 0{,}15 + 0{,}38 \\ + 0{,}41 + 0{,}38 + 1{,}97 + 1{,}55 + 0{,}14 + 0{,}36 = 9{,}79.$$

Da hier $n = 1021$ sowie $k = 6$, $m = 2$ und damit $M = \min(6; 2) = 2$ ist, folgt für die kleinste obere Schranke χ^2_{max} des χ^2-Koeffizienten nach (9.2)

$$\chi^2_{max} = 1021 \cdot 1 = 1021.$$

Der Wert $\chi^2 \approx 9{,}79$ liegt deutlich näher an der unteren Schranke 0, was für einen nur schwach ausgeprägten Merkmalszusammenhang spricht. Für den Φ-Koeffizienten (9.3) gilt

$$\Phi = \sqrt{\frac{9{,}79}{1021}} \approx 0{,}098.$$

Dieser Wert und der für das Cramérsche Zusammenhangsmaß V aus (9.5) stimmen hier überein.

Man kann die vorstehenden Berechnungen natürlich auch unter Heranziehung geeigneter Statistiksoftware durchführen, wie der folgende SPSS-Ausdruck illustriert. Die letzte Spalte des Ausdrucks ist hier irrelevant und wird daher nicht weiter kommentiert.

		Wert	Näherungswe ise Signifikanz
Nominal- bzgl. Nominalmaß	Cramer-V	,098	,081
	Kontingenzkoeffizient	,098	,081
	Anzahl der gültigen Fälle	1021	

Tab. 9.1: *SPSS-Computerausdruck zur Berechnung der Zusammenhangsmaße Φ und V (Politbarometer-Daten aus Beispiel 8.1)*

Zusammenhangsmessung bei Merkmalen mit nur zwei Ausprägungen

Im Spezialfall der in Tabelle 8.4 wiedergegebenen **Vierfeldertafel** hat man für den χ^2-Koeffizienten (9.1) zunächst die aus nur vier Summanden bestehende Doppelsumme

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

Aus dieser Darstellung gewinnt man nach Einsetzen von 8.9 und einigen – hier nicht wiedergegebenen – Umformungen die nachstende Formel, bei

[2] Hier sind die 12 Summanden zwecks übersichtlicherer Präsentation gerundet wiedergegeben. Normalerweise wird man aber erst am Schluss der Rechnung runden, um die Rundungsfehler klein zu halten. Wenn man so verfährt, resultiert 9,81 anstelle von 9,79. In beiden Fällen ergibt sich aber für den Φ-Koeffizienten bei Berücksichtigung von drei Dezimalstellen der Wert 0,098.

der im Nenner das Produkt der Randhäufigkeiten steht:

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1.}h_{2.}h_{.1}h_{.2}}. \tag{9.7}$$

Aufgabe 9.1

Bei einer Vierfeldertafel, oder – allgemeiner – im Falle $M = 2$ stimmen der Phi-Koeffizient Φ aus (9.3) und das Kontentingenzmaß (9.5) von Cramér stets überein. Es gilt dann offenbar

$$\Phi = V = \frac{|h_{11}h_{22} - h_{12}h_{21}|}{\sqrt{h_{1.}h_{2.}h_{.1}h_{.2}}}. \tag{9.8}$$

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

Exkurs 9.1: Weitere Zusammenhangsmaße

Es gibt noch weitere Ansätze zur Messung von Zusammenhängen bei nominalskalierten Merkmalen, die wie der Φ-Koeffizient und Cramér's V Modifikationen von (9.1) darstellen. Erwähnt sei hier ein Zusammenhangsmaß von Karl Pearson, das in der Literatur meist mit K oder mit C abgekürzt wird und sich vom Φ-Koeffizienten nur dadurch unterscheidet, dass unter dem Wurzelzeichen im Nenner von (9.3) statt n der Term $\chi^2 + n$ steht:

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

Auch dieses Maß ist noch dimensionsabhängig. Analog zu (9.4) gilt

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

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 (9.5), der vom χ^2-Koeffizient in einem Schritt zu einem normierten Zusammenhangsmaß führt, ist allerdings transparenter und weniger umständlich.

9.2 Metrische Merkmale

Bei Merkmalen mit metrischer Skalierung sind, anders als bei nominalskalierten 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**

$$\begin{aligned} s_{xy} :&= \frac{1}{n} \cdot [(x_1 - \overline{x})(y_1 - \overline{y}) + \ldots + (x_n - \overline{x})(y_n - \overline{y})] \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \end{aligned} \tag{9.9}$$

Ein nicht-normiertes Zusammenhangsmaß: die empirische Kovarianz

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

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

Diese Formel verifiziert man, ähnlich wie (5.7), wenn man den in (9.9) hinter dem Summenzeichen stehenden Produktterm ausmultipliziert und dann die Summierung gliedweise vornimmt.

Was die Kovarianz inhaltlich bezeichnet wird verständlich, wenn man die Datenpaare $(x_1, y_1), \ldots, (x_n, y_n)$ für X und Y in einem Streudiagramm präsentiert, in das man – parallel zum ersten – noch ein zweites Koordinatensystem einzeichnet, dessen Ursprung im Punkt $(\overline{x}, \overline{y})$ liegt. Durch das zweite Bezugssystem sind, wie in Abbildung 9.1 dargestellt, vier Quadranten definiert. Jeder Punkt (x_i, y_i) definiert zusammen mit den auf den Achsen des zweiten Koordinatensystems liegenden Punkten $(x_i, \overline{y})$ und $(\overline{x}, y_i)$ sowie dem neuen Ursprung $(\overline{x}, \overline{y})$ (lies: *x-quer-y-quer*) ein Rechteck mit Flächeninhalt A_i.

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

$$p_i := (x_i - \overline{x})(y_i - \overline{y}) \qquad i = 1, \ldots, n,$$

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 $(x_i - \overline{x})$ und $(y_i - \overline{y})$ 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.

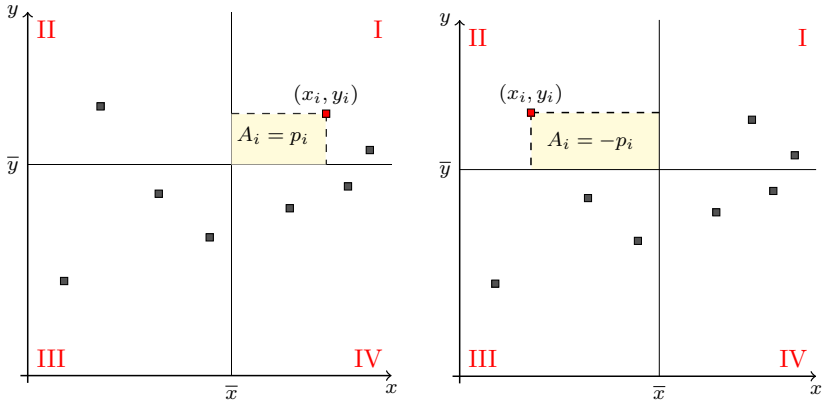


Abb. 9.1: *Interpretation der Summanden in der Kovarianzformel*

Abbildung 9.1 veranschaulicht 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 $(\overline{x}, \overline{y})$ 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.

Flash-Animation „Vorzeichen der Kovarianz“

Wenn alle Punkte auf einer steigenden Geraden durch $(\overline{x}, \overline{y})$ 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 $(\overline{x}, \overline{y})$ 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.

Ein normiertes Zusammenhangsmaß

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 resp. s_y dividiert. Das resultierende Zusammenhangsmaß

Karl Pearson

$$r := \frac{s_{xy}}{s_x \cdot s_y} \tag{9.10}$$

wird **Korrelationskoeffizient** genannt. Da der Ansatz (9.10) 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 (9.10) ersieht man, dass die Merkmale X und Y symmetrisch eingehen. Eine Vertauschung der Merkmalsbezeichnungen ändert nichts am Wert von r.

Wenn man in (9.10) für den Zähler den Summenterm aus (9.9), 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

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{9.11}$$

Mit (9.9) und (5.8) gewinnt man aus (9.11) noch als weitere Darstellung

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}}. \tag{9.12}$$

Flash-Animation „Schranken für r“

Da die im Nenner von (9.10) 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 $+1$:

$$-1 \leq r \leq 1. \tag{9.13}$$

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 $(\overline{x}, \overline{y})$.

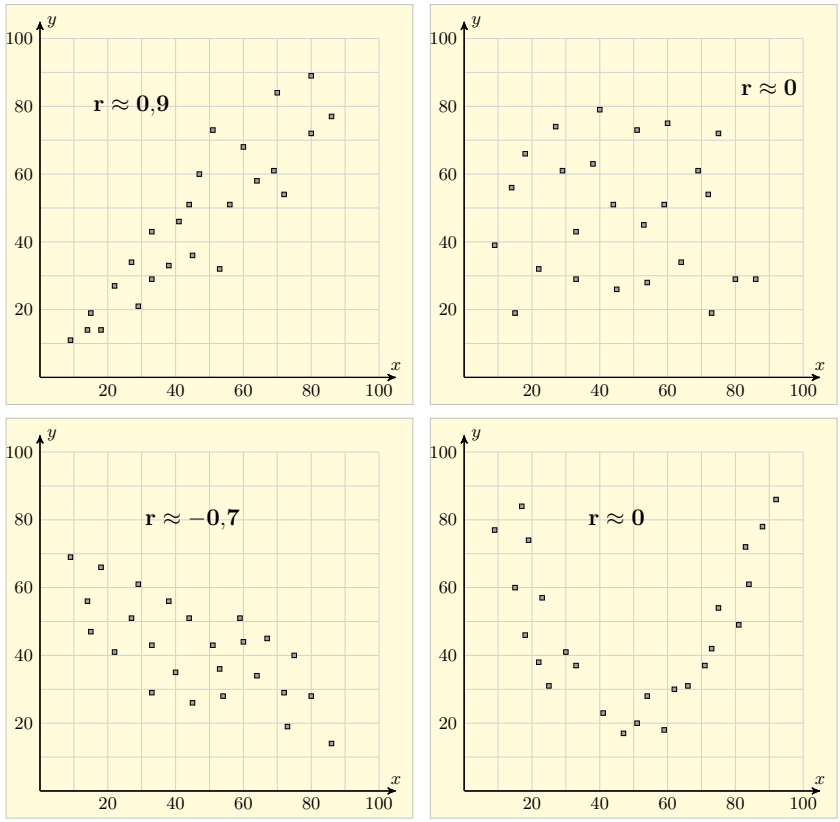


Abb. 9.2: *Korrelationskoeffizienten für verschiedene Streudiagramme*

Abbildung 9.2 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 8.5, 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 $(x_1, y_1), \ldots, (x_n, y_n)$ sind im letztgenannten Fall so auf die vier Quadranten des in Abbildung 9.1 wiedergegebenen Koordinatensystems mit Bezugspunkt $(\overline{x}, \overline{y})$ verteilt, dass sich die Beiträge p_i beim Aufsummieren gerade aufheben.

Interaktives Lernobjekt „Korrelation“

Das dritte Streudiagramm – untere Hälfte von Abbildung 9.2 – 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 $r \neq 0$

Korrelationskoeffizient r: Maß für linearen Zusammenhang

lässt also nur auf 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*, bei Werten $0{,}5 \leq |r| < 1$ von *mäßiger bis starker* Korrelation.

Beispiel 9.2: Wie gut waren die Vorhersagen 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 $(x_1, y_1), \ldots, (x_n, y_n)$ lägen dann auf einer steigenden Geraden ($r = 1$). Ein hoher Wert für r wäre also ein Ausweis hoher Prognosegüte.

Tabelle 9.2 weist in den Spalten $1 - 3$ für 15 Perioden i (Jahre 1983, . . . ,1997) 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	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1983	1,00	1,20	1,20	1,00	1,44
1984	2,50	2,60	6,50	6,25	6,76
1985	3,00	2,50	7,50	9,00	6,25
1986	3,00	2,50	7,50	9,00	6,25
1987	2,00	1,70	3,40	4,00	2,89
1988	1,50	3,40	5,10	2,25	11,56
1989	2,50	4,00	10,00	6,25	16,00
1990	3,00	4,60	13,80	9,00	21,16
1991	3,50	3,40	11,90	12,25	11,56
1992	2,50	1,50	3,75	6,25	2,25
1993	0,00	-1,90	0,00	0,00	3,61
1994	0,00	2,30	0,00	0,00	5,29
1995	3,00	1,90	5,70	9,00	3,61
1996	2,00	1,40	2,80	4,00	1,96
1997	2,50	2,20	5,50	6,25	4,84
Summe	32,00	33,30	84,65	84,50	105,43
Mittelwert	2,133	2,220	5,643	5,633	7,029

Tab. 9.2: *Wachstumsprognosen der „Fünf Weisen“ und wahre Werte (linker Teil: Datentabelle; rechter Teil: Arbeitstabelle)*

Der Wert $r = 1$ ist natürlich in der Realität nie 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 Krise im Finanz- und Immobiliensektor, die seit 2008 zu weltweiten Restrukturierungen innerhalb des Bankensystems führte. Prognosen werden daher während des Prognosezeitraums bei Bedarf noch korrigiert.

Wenn man für den Datensatz aus Tabelle 9.2 den Bravais-Pearsonschen Korrelationskoeffizient r bestimmen will, wird man in der Praxis eine Statistiksoftware anwenden, weil die Berechnung anhand von (9.11) oder (9.12) etwas mühsam ist. Es sei aber dennoch hier einmal exemplarisch vorgeführt, wie man bei der manuellen Berechnung vorgehen kann.

Um die Formel (9.12) anzuwenden, benötigt man zunächst die Mittelwerte $\overline{x}$ und $\overline{y}$, die man durch Aufsummieren der zweiten resp. dritten Spalte von Tabelle 9.2 und nachfolgende Division durch $n = 15$ gewinnt. Ferner benötigt man den Mittelwert $\overline{xy}$ der Produkte aus x_i und y_i sowie die Mittelwerte $\overline{x^2}$ und $\overline{y^2}$ aus den Quadraten der Werte x_i resp. y_i. Diese Mittelwerte erhält man durch Erweiterung der Tabelle um drei Hilfsspalten, die im rechten Teil von Tabelle 9.2 ausgewiesen sind. Die Tabelle weist in der unteren Zeile alle Mittelwerte auf drei Dezimalstellen gerundet aus. Einsetzen der Mittelwerte in (9.12) liefert für r

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

Aufgabe 9.2

Das Ergebnis ist nicht überraschend – die prognostizierten und die beobachteten realen Wachstumsraten sind positiv korreliert. Dasselbe Ergebnis erhält man natürlich auch bei Verwendung jeder marktgängigen Statistiksoftware.

		x	y
x	Korrelation nach Pearson	1	,602*
	Signifikanz (2-seitig)		,018
	N	15	15
y	Korrelation nach Pearson	,602*	1
	Signifikanz (2-seitig)	,018	
	N	15	15

```
R> x <- c(1, 2.5, 3, 3, 2, 1.5, 2.5, 3, 3.5, 2.5, 0, 3, 2, 2.5)
R> 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)
R> cor(x, y)
[1] 0.6018267
```

Tab. 9.3: *Computerausdruck zur Berechnung des Korrelationskoeffizienten r mit SPSS (links) und mit R (rechter Teil)*

Die beiden Screenshots zeigen das Ergebnis einer Berechnung von r mit SPSS sowie bei Verwendung der freien Statistiksoftware *R*. Die Signifikanzangaben im SPSS-Output in Tabelle 9.3 werden hier nicht thematisiert. Der R-Output umfasst auch die Dateneingabe, die beim SPSS-Screenshot ausgeblendet ist.

Korrelation impliziert nicht zwingend einen sachlogischen Zusammenhang

Der Korrelationskoeffizient r kann Aufschluss darüber geben, ob es einen mehr oder weniger ausgeprägten *linearen* Zusammenhang zwischen zwei metrischen Merkmalen gibt. Da eine Vertauschung von X und Y den Wert von r nicht berührt, also keines der Merkmale ausgezeichnet ist, sagt r nichts aus über die Richtung eines Zusammenhangs im Sinne eines direkten Einflusses eines Merkmals auf das andere Merkmal aus. Ein hoher Absolutbetrag $|r|$ besagt lediglich, dass die Daten für die in Rede stehenden Merkmale entweder eine gleichgerichtete Tendenz ausweisen (im Falle $r > 0$) oder eine gegenläufige Tendenz (im Falle $r < 0$). Ein anhand eines großen Werts $|r|$ festgestellter Zusammenhang muss nicht zwingend bedeuten, dass zwischen den Merkmalen ein Kausalzusammenhang besteht, also eine sachlogische Verbindung.

Betrachtet man etwa Zeitreihendaten für zwei Merkmale X und Y, für die kein kausaler Zusammenhang erkennbar ist – etwa die in den Jahren 1995 - 2009 von OPEC-Staaten geförderte Erdölmenge X und die Anzahl Y der in den gleichen Jahren in Deutschland eingeschulten Kinder – so könnte es sein, dass beide Merkmale in den 15 Jahren eine gleichgerichtete Entwicklung genommen haben und für r ein Wert $r > 0{,}5$ errechnet wird. Die 15 Datenpunkte $(x_i; y_i)$ weisen dann einen empirischen Zusammenhang aus, der sich allein auf die Daten bezieht und nicht als Wirkzusammenhang interpretiert werden darf.

Denkbar ist auch der Fall, dass zwischen 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. Man bezeichnet diesen speziellen Fall eines fehlenden direkten sachlogischen Zusammenhangs gelegentlich auch als **Scheinkorrelation**.

Beispiel 9.3: Scheinkorrelation

Betrachten wir die vom Deutschen Wetterdienst für Düsseldorf in den Jahren $1970 - 1994$ registrierten Sturmtage x_i und die im selben Zeitraum durch Naturkatastrophen weltweit verursachten volkswirtschaftlichen Schäden y_i in Milliarden US-Dollar. Der Index i bezeichnet das jeweilige Jahr. Die 25 Datenpaare (x_i, y_i) sind im linken Teil von Abbildung 9.3 in Form eines Streudiagramms wiedergegeben. Aus den hier nicht numerisch wiedergegebenen Daten errechnet man einen Korrelationskoeffizienten ($r = 0{,}324$), der eine schwach ausgeprägte positive Korrelation zwischen den Merkmalen zu stützen scheint. Der rechte Teil von Abbildung 9.3 zeigt die Daten für jedes Merkmal einzeln in Form je eines Zeitreihengraphen. Hier erkennt man die zeitliche Entwicklung der beiden Einzelmerkmale X und Y, die im Streudiagramm nicht mehr zu sehen ist. Beide Merkmale weisen tatsächlich einen nach oben gerichteten Trend auf.

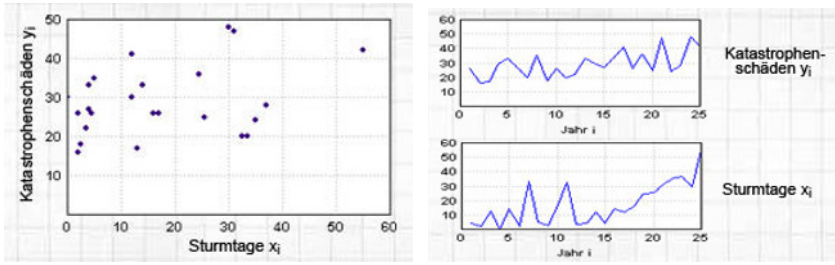


Abb. 9.3: *Streudiagramm und Zeitreihen für die Merkmale „Anzahl X der Sturmtage in Düsseldorf“ und „Weltweite Schäden Y durch Naturkatastrophen“ (1970 – 1994)*

Nur durch sachlogische Überlegungen – nicht aus den Daten und den Grafiken alleine – wird man darauf stoßen, dass ein drittes Merkmal Z für die Veränderung bei den Merkmalen X und Y verantwortlich sein könnte. Als Merkmal Z käme die Variable „Weltweite CO_2-Emission“ in Betracht, die von der Größe der Weltbevölkerung und dem industriellen Entwicklungsstand beeinflusst wird. Wäre Z für die gleichgerichtete Änderung bei den Merkmalen X und Y verantwortlich, müsste man den zwischen den Merkmalen X und Y anhand des Korrelationskoeffizienten r festgestellten Zusammenhang als Scheinkorrelation bewerten.

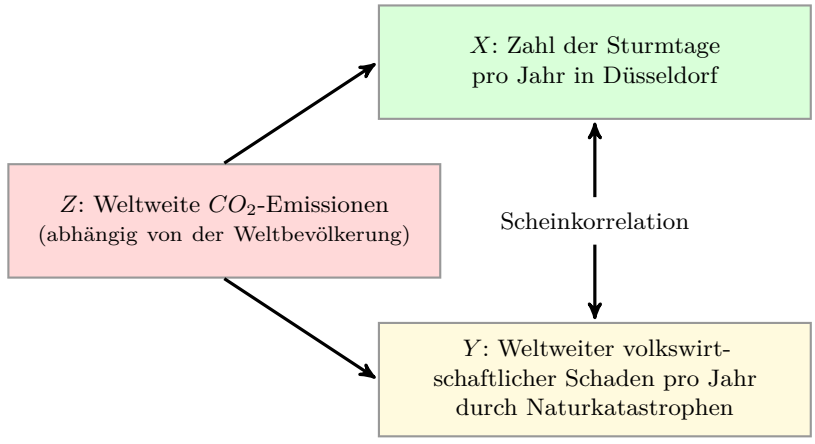


Abb. 9.4: *Scheinkorrelation zwischen den Merkmalen „Anzahl X der Sturmtage in Düsseldorf“ und „Weltweite Schäden Y durch Naturkatastrophen“ (1970 – 1994)*

Exkurs 9.2: 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 ausgeblendet würde. Die von Drittvariableneinflüssen bereinigte Korrelation heißt auch **partielle Korrelation** oder **Partialkorrelation**. Durch eine solche Bereinigung lassen sich auch Scheinkorrelationen aufdecken.

Ein bei Sedlmeier / Renkewitz (2013, Abschnitt 7.6) aufgeführtes Beispiel für die Anwendung des Konzepts der Partialkorrelation bezieht sich auf den Zusammenhang von Kirchgangshäufigkeit X und Ausländerfeindlichkeit Y, bei der das Alter Z als Drittvariable in Betracht kommt. In der *Psychologie* wird eine Drittvariable, die den Zusammenhang zwischen einer unabhängigen Variablen X und einer abhängigen Variablen Y beinflusst, auch **Moderatorvariable** genannt.

9.3 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 (9.11) 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 (9.11) x_i und y_i durch $rg(x_i)$ bzw. $rg(y_i)$ sowie $\overline{x}$ und $\overline{y}$ durch die Mittelwerte $\overline{rg}_x$ resp. $\overline{rg}_y$ der Rangplätze ersetzt. Man erhält so den mit r_{SP} (lies: *r-s-p*)

abgekürzten **Rangkorrelationskoeffizienten nach Spearman**:

$$r_{SP} = \frac{\sum_{i=1}^{n} (rg(x_i) - \overline{rg}_x)(rg(y_i) - \overline{rg}_y)}{\sqrt{\sum_{i=1}^{n} (rg(x_i) - \overline{rg}_x)^2} \cdot \sqrt{\sum_{i=1}^{n} (rg(y_i) - \overline{rg}_y)^2}}. \tag{9.14}$$

Da r_{SP} sich als Anwendung des Korrelationskoeffizenten nach Bravais-Pearson auf Paare $(rg(x_i), rg(y_i))$ von Rangpositionen interpretieren lässt, gelten die Schranken aus (9.13) auch für den Rangkorrelationskoeffizienten, d. h. es gilt

$$-1 \leq r_{SP} \leq 1. \tag{9.15}$$

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 gleichsinniger - oder ein 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 (9.14). Die Mittelwerte $\overline{rg}_x$ resp. $\overline{rg}_y$ der Rangplätze sind dann jeweils identisch mit dem Mittelwert aus den ersten n natürlichen Zahlen, also der Zahlen $1, 2, \ldots, n$. Man kann zeigen, dass die Summe der Zahlen $1, 2, \ldots, n$ durch $\frac{n(n+1)}{2}$ gegeben ist, ihr Mittelwert also durch $\frac{n+1}{2}$. Einsetzen in (9.14) liefert bei Verwendung der Abkürzung d_i für die Differenz der Rangpositionen $rg(x_i)$ und $rg(y_i)$ nach elementaren Umformungen

Aufgabe 9.3

$$r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \qquad d_i := rg(x_i) - rg(y_i). \tag{9.16}$$

Beispiel 9.4: 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 9.4 ausgewiesen.

	Sachbearbeiter A		Sachbearbeiter B		
Kunden-Nr. i	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. 9.4: *Bonitätsbewertung von Sparkassenkunden*

Ausgangspunkt für die Berechnung von r_{SP} ist die Rechenformel (9.16). Um die Formel anwenden zu können, müssen die Rangplätze der Ausgangsdaten $x_1, x_2, \ldots, x_5$ und $y_1, y_2, \ldots, y_5$ 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 (9.16) ein, resultiert der Wert

$$r_{SP} = 1 - \frac{6 \cdot \left[1^2 + 1^2 + (-1)^2 + 0^2 + (-1)^2\right]}{5 \cdot (25 - 1)} = 0{,}8.$$

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 Gütefunktion 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.

10 Zufall und Wahrscheinlichkeit

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 werden noch der Begriff der bedingten Wahrscheinlichkeit und der der Unabhängigkeit von Ereignissen definiert.

Vorschau auf das Kapitel

10.1 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

$$\Omega = \{\omega : \ \omega \text{ ist Elementarereignis}\} \tag{10.1}$$

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_10

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 $\overline{A}$ (lies: *Komplementärmenge* zu A) bezeichnete **Komplementärereignis** zu A ist das Ereignis, das genau dann eintritt, wenn A nicht eintritt. Die Menge $\overline{A}$ umfasst alle Elementarereignisse, die zu Ω, nicht aber zu A gehören. Man schreibt hierfür auch $\overline{A} = \Omega \setminus A$ (lies: $\overline{A}$ 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 $\overline{\Omega}$ 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, ...) als Kreise oder Ellipsen dargestellt sind. Das Rechteck repräsentiert die Ergebnismenge Ω, von dem die eingezeichneten Mengen Teilmengen sind.

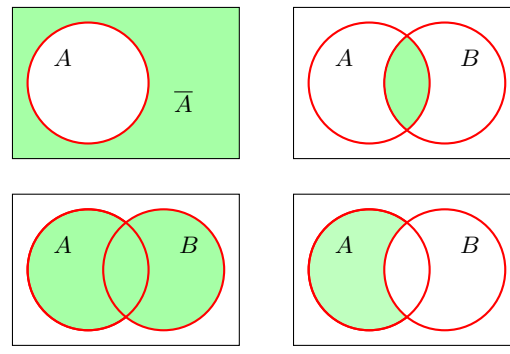


Abb. 10.1: *Venn-Diagramme für* $\overline{A}$, $A \cap B$, $A \cup B$ *und* $A \setminus B$

Abbildung 10.1 zeigt vier Venn-Diagramme. Die oberen beiden Teilgrafiken veranschaulichen das Komplementärereignis $\overline{A} = \Omega \setminus A$ und die Schnittmenge $A \cap B$. Die unteren zwei Teilgrafiken zeigen die Vereinigungsmenge $A \cup B$ resp. die Differenzmenge $A \setminus B = A \cap \overline{B}$. Die dargestellten Ereignisse bzw. Mengen sind innerhalb des Ω symbolisierenden Rechtecks jeweils durch dunklere Färbung ausgewiesen.

Aufgabe 10.1

Beispiel 10.1: 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 Ω = {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.

Aufgabe 10.2

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 $\overline{A} = \Omega \setminus A$ zu A ist hier $\overline{A} = \{2, 4, 6\}$ (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

$$A \cap B = \{5\} = \{\text{Augenzahl ist ungerade und größer als } 4\}.$$

Beim *Würfeln mit zwei Würfeln* umfasst Ω schon 36 Elementarereignisse, nämlich die zu Paaren

$$\Omega = \{(1;1), (1;2), \ldots, (1;6), (2;1), (2;2), \ldots, (2;6), \ldots, (6;1), (6;2), \ldots, (6;6)\}$$

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 $\overline{A} = \{\text{Augensumme ist größer als } 3\}$ 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 $\Omega = \{1, 2, 3, \ldots\} = \mathbb{N}$ 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.

Heutiger Wahrscheinlichkeitsbegriff

In der Statistik wird der Wahrscheinlichkeitsbegriff hingegen objektiv quantifiziert. Dabei stützt man sich, wie inzwischen jeder Teilbereich der modernen Mathematik, auf eine axiomatische Fundierung.

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).

[1] Der Buchstabe „P“ steht für „probability“, das englische Wort für „Wahrscheinlichkeit“. Man findet anstelle der Notation $P(..)$ in der Literatur auch die Notationen $Pr(..)$ oder $W(..)$.

[2] Das dritte Axiom ist hier für den Fall formuliert, dass die Ereignismenge 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 (vgl. hierzu z. B. Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 4.8)).

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 $M = \{a_1, a_2, \ldots, a_k\}$ beschrieben sind, die relativen Häufigkeiten für je zwei disjunkte Teilmengen von M und die Summe aller relativen Häufigkeiten ist stets 1.

Rechenregeln für Wahrscheinlichkeiten

Aus dem Axiomensystem von Kolmogoroff lassen sich einige elementare Rechenregeln für Wahrscheinlichkeiten ableiten. Unter Heranziehung der Venn-Diagramme aus Abbildung 10.1 verifiziert man die Gleichungen

$$P(\overline{A}) = 1 - P(A) \tag{10.2}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{10.3}$$

$$P(A \setminus B) = P(A) - P(A \cap B). \tag{10.4}$$

Gleichung (10.3) wird z. B. anhand des dritten Venn-Diagramms aus Abbildung 10.1 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.

Berechnung von Wahrscheinlichkeiten erfordert zusätzliche Information

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:

Bedingungen für Laplace-Experimente

L1: Die Ergebnismenge ist endlich, also $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$.

L2: Die Wahrscheinlichkeiten für die n Elementarereignisse sind gleich.

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

Simon Pierre Laplace

der Anzahl aller möglichen Ergebnisse des Zufallsexperiments:[3]

$$P(A) = \frac{\text{Anzahl der für } A \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}. \tag{10.5}$$

Im Zähler steht also die Anzahl der Elementarereignisse, für die das Ereignis A als eingetreten gilt, im Nenner die Anzahl aller zu Ω gehörenden Elementarereignisse.

Eine nach (10.5) 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 (10.5) erfüllt sind, folgt z. B. sofort daraus, dass der Zähler in (10.5) stets einen Wert besitzt, der zwischen 0 und dem Wert des Nenners liegt.

Gleichung (10.5) liefert für viele Anwendungen – etwa bei Glücksspielen – eine leicht handhabbare und sehr nützliche Rechenformel. Eine Definition des Begriffs „Wahrscheinlichkeit“ stellt (10.5) in Verbindung mit L1 und L2 aber nicht dar, weil der zu erklärende Begriff der Wahrscheinlichkeit schon in die Annahme L2 eingeht.

Beispiel 10.2: 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 Lernobjekt „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 $\overline{A} = \Omega \setminus A$ lässt sich nach (10.5) ermitteln als $P(\overline{A}) = \frac{4}{6} = \frac{2}{3}$ oder anhand von (10.2) gemäß $P(\overline{A}) = 1 - \frac{1}{3} = \frac{2}{3} \approx 0{,}667$. 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.

[3] Die 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“.

Beim *dreifachen Münzwurf* kann man die Wahrscheinlichkeit für

$$A = \{\text{Bei den drei Münzwürfen tritt zweimal „Zahl“ auf}\}$$

ebenfalls anhand des Laplace-Ansatzes (10.5) berechnen. Die Ergebnismenge Ω ist beim dreifachen Münzwurf bei erneuter Verwendung von „Z“ für „Zahl“ und „K“ für „Kopf“ durch die acht Tripel

Aufgabe 10.3

$$\Omega = \{(Z,Z,Z),(Z,Z,K),(Z,K,Z),(K,Z,Z),$$
$$(K,K,Z),(K,Z,K),(Z,K,K),(K,K,K)\}$$

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.

Wahrscheinlichkeiten als Grenzwert relativer Häufigkeiten

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 $f_j(Zahl)$ für das Eintreten von „Zahl“ während dieses Zufallsexperiments verfolgt $(j = 1, 2, ..., n)$, so stellt sich am Ende ein Wert $f_n(Zahl)$ ein, der sich tendenziell dem Wert 0,5 nähert und zwar um so deutlicher, je größer man n wählt.

Abbildung 10.2 zeigt den Verlauf eines virtuellen Münzwurfexperiments mit $n = 1000$. Bei einem solchen Simulationsexperiment kann man, anders als in der Realität, die Wahrscheinlichkeit $p := P(Zahl)$ für „Zahl“ und damit auch die von $1-p = P(Kopf)$ für „Kopf“ bei der Programmierung festlegen und folglich als bekannt voraussetzen.

Interaktives Lernobjekt „Münzwurf“

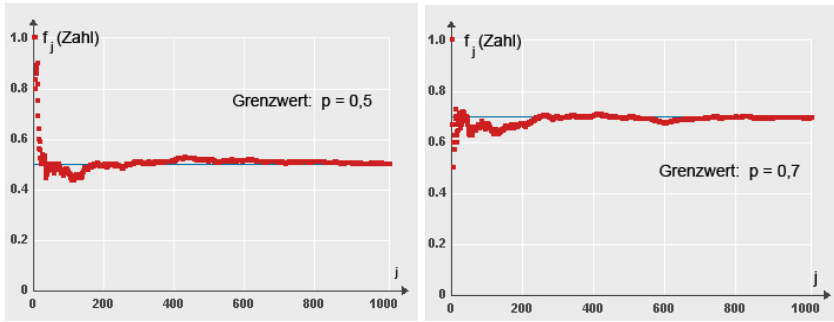


Abb. 10.2: *Relative Häufigkeiten für „Zahl“ bei 1000-fachem Wurf einer fairen Münze (linke Teilgrafik) und einer unfairen Münze (rechts)*

Im linken Teil von Abbildung 10.2 ist der Verlauf eines „fairen“, im rechten Teil der eines „unfairen“ Münzwurfexperiments zu sehen ($p = 0{,}5$ resp. $p = 0{,}7$). 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. Der rechte Teil von Abbildung 10.2 zeigt, dass die Bestimmung von Wahrscheinlichkeiten als Grenzwert[4]

$$f_n(A) \overset{n\to\infty}{\longrightarrow} P(A) \tag{10.6}$$

relativer Häufigkeiten bei Zufallsexperimenten mit endlicher Ergebnismenge, anders als der Laplace-Ansatz (10.5), nicht an die Bedingung L2 der Gleichwahrscheinlichkeit der Elementarereignisse gebunden ist. Wirft man bei einem realen Münzwurfexperiment eine Münze n-mal ohne zu wissen, ob es sich um eine „faire“ Münze handelt, so kann der Wert, gegen den die relative Häufigkeit $f_n(Zahl)$ mit zunehmendem n konvergiert, Aufschluss darüber gegeben, ob die Münze „fair“ ist. Wenn sich die relative Häufigkeit dem Grenzwert 0,5 annähert, ist von einer fairen Münze auszugehen.

10.2 Zufallsstichproben und Kombinatorik

Um die Wahrscheinlichkeit $P(A)$ von Ereignissen A bei Laplace-Experimenten nach (10.5) 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:

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

[4] Die Konvergenz in (10.6) 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).

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 vom **Urnenmodell mit Zurücklegen** oder von einer **Stichprobenziehung mit Zurücklegen**. Der zweite Fall charakterisiert das **Urnenmodell ohne Zurücklegen** bzw. eine **Stichprobenziehung ohne Zurücklegen**.

- Ziehen mit und 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**. 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, 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.

Die Wahrscheinlichkeiten, die man nach (10.5) bestimmt, hängen also davon ab, welches Modell 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

Anzahl der Möglichkeiten einer geordneten Auswahl von n Elementen:

- beim Ziehen ohne Zurücklegen

einer Zufallsstichprobe des Umfangs n aus N Elementen beträgt somit $N \cdot (N-1) \cdot \ldots \cdot (N-n+1)$. Dieser Produktterm lässt sich kompakter schreiben, wenn man auf die Kurzschreibweise $N!$ (lies: *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

$$\begin{aligned} N \cdot (N-1) \cdot \ldots \cdot (N-n+1) &= \frac{N \cdot (N-1) \cdot \ldots \cdot 1}{(N-n) \cdot (N-n-1) \cdot \ldots \cdot 1} \\ &= \frac{N!}{(N-n)!} \end{aligned} \tag{10.7}$$

- beim Ziehen mit Zurücklegen

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

$$\underbrace{N \cdot N \cdot \ldots \cdot N}_{n-mal} = N^n. \tag{10.8}$$

Anzahl der Möglichkeiten einer ungeordneten Auswahl von n Elementen:

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 anzuordnen. 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 (10.7), der unterschiedliche Anordnungen der n Stichprobenelemente berücksichtigt, ist also durch $n!$ zu dividieren, wenn die Reihenfolge der Elemente keine Rolle spielt. Man erhält so

$$\frac{\frac{N!}{(N-n)!}}{n!} = \frac{N!}{(N-n)! \cdot n!}.$$

[5] Ist k eine natürliche Zahl, so bezeichnet $k! := 1 \cdot 2 \cdot \ldots \cdot k$ das Produkt aus allen natürlichen Zahlen von 1 bis k. Für 0 ist die Fakultät durch $0! = 1$ definiert.

[6] Die 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 Ausage 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.

Der rechtsstehende Term wird **Binomialkoeffizient** genannt und mit $\binom{N}{n}$ (lies: *N über n*) abgekürzt.[7] Es gilt also

- beim Ziehen ohne Zurücklegen

$$\binom{N}{n} := \frac{N!}{(N-n)! \cdot n!}. \tag{10.9}$$

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

$$\binom{N+n-1}{n} = \frac{(N+n-1)!}{(N-1)! \cdot n!}. \tag{10.10}$$

Tabelle 10.1 fasst die Ergebnisse für die vier betrachteten Fälle zusammen.

Art der Stichprobe	Ziehen *ohne* Zurücklegen	Ziehen *mit* Zurücklegen
Ziehen *mit* Berücksichtigung der Reihenfolge	$\frac{N!}{(N-n)!}$	N^n
Ziehen *ohne* Berücksichtigung der Reihenfolge	$\binom{N}{n}$	$\binom{N+n-1}{n}$

Tab. 10.1: *Anzahl der Möglichkeiten der Ziehung einer Stichprobe des Umfangs n aus einer Grundgesamtheit mit N Elementen*

Beispiel 10.3: Varianten von Stichprobenziehungen

Für die Formeln (10.7) - (10.10) sei je ein Anwendungsbeispiel genannt und durchgerechnet. Als Beispiel für die Anwendung von (10.7) 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

$$\frac{8!}{(8-3)!} = \frac{8 \cdot 7 \cdot \ldots \cdot 1}{5 \cdot 4 \cdot \ldots \cdot 1} = 8 \cdot 7 \cdot 6 = 336.$$

Zur Illustration der Anwendung von (10.8) kann das Würfeln mit zwei Würfeln herangezogen werden, etwa das simultane Werfen je eines roten und eines grünen Würfels. In Beispiel 10.1 wurde bereits die Ereignismenge Ω dieses

[7] Der Binomialkoeffizient $\binom{N}{n}$ gibt die Anzahl der Möglichkeiten an, eine Stichprobe ohne Zurücklegen des Umfangs n aus einer Menge mit N Elementen ohne Berücksichtigung der Reihenfolge zu ziehen. Es ist $\binom{N}{0} = 1$, $\binom{N}{1} = N$ und $\binom{N}{N} = 1$.

Zufallsexperiments wiedergegeben. Die Menge Ω umfasst 36 Zahlenpaare $(i;j)$, wobei i die Augenzahl des ersten und j die des zweiten Würfels darstellt $(i = 1, 2, ..., 6;\ j = 1, 2, .. 6)$. 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 (10.8) als $6^2 = 36$.

Die Anzahl der möglichen Ausgänge beim deutschen Zahlenlotto lässt sich anhand von (10.9) ermitteln, weil es hier um eine Stichprobenziehung *ohne Zurücklegen* und *ohne Berücksichtigung der Anordnung* geht. Es resultiert

$$\binom{49}{6} = \frac{49!}{43! \cdot 6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13983816.$$

Von diesen fast 14 Millionen Möglichkeiten wird nur eine einzige realisiert. Die Wahrscheinlichkeit dafür, 6 Richtige zu erzielen, ist also extrem gering. Man errechnet mit (10.5) den Wert

$$\frac{1}{13\,983\,816} \approx 0{,}0000000715 = 7{,}15 \cdot 10^{-8}.$$

Aufgabe 10.4

Als Beispiel für die Anwendung von (10.10) sei die Wahl eines Unternehmensvorstands genannt, bei der 3 Bewerber $B1$, $B2$ und $B3$ zur Auswahl stehen. 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 (10.10) 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

$$\Omega = \{(B1, B1), (B1, B2), (B2, B2), (B1, B3), (B2, B3), (B3, B3)\}.$$

10.3 Bedingte Wahrscheinlichkeiten

In Abschnitt 8.1 wurden Häufigkeiten auch auf Teilmengen einer Population bezogen. So wurde in den Beispielen 8.1 - 8.2 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 (10.5) 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 (10.5) 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 (10.5) 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

$$P(A|B) = \frac{\text{Anzahl der für } A \cap B \text{ günstigen Ergebnisse}}{\text{Anzahl der für } B \text{ günstigen Ergebnisse}}. \tag{10.11}$$

Da die Wahrscheinlichkeit $P(A)$ für den Eintritt von A durch (10.5) erklärt ist, gilt analog für die Wahrscheinlichkeiten $P(A \cap B)$ und $P(B)$

$$P(A \cap B) = \frac{\text{Anzahl der für } A \cap B \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}$$

$$P(B) = \frac{\text{Anzahl der für } B \text{ günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}.$$

Zusammenhang zwischen bedingten Wahrscheinlichkeiten

Multipliziert man den Bruchterm in der Formel für $P(A \cap B)$ mit dem *Kehrwert* des Bruchterms der letzten Gleichung, also mit $\frac{1}{P(B)}$, resultiert der Bruchterm aus (10.11). Mit (10.11) gilt also auch

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{10.12}$$

Analog gilt für die bedingte Wahrscheinlichkeit $P(B|A)$ die Darstellung

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{10.13}$$

Die Formeln (10.12) und (10.13) 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

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A). \tag{10.14}$$

Für die bedingten Wahrscheinlichkeiten $P(A|B)$ und $P(B|A)$ gilt also

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \tag{10.15}$$

die nach dem englischen Mathematiker und Pfarrer Thomas Bayes (1702 - 1761) auch **Satz von Bayes** genannt wird.[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 $P(A|\overline{B})$ beide mit $P(A)$ identisch sind. Man kann in diesem Falle in (10.12) den Term $P(A|B)$ durch $P(A)$ ersetzen und erhält dann nach Multiplikation mit $P(B)$

Unabhängigkeit von Ereignissen

$$P(A \cap B) = P(A) \cdot P(B). \tag{10.16}$$

Zwei zufällige Ereignisse A und B sind also genau dann unabhängig, wenn sie der Bedingung (10.16) genügen. Unabhängig sind z. B. die Ergebnisse zweier aufeinanderfolgender Roulettespiele oder Münzwürfe.

Es sei angemerkt, dass die Verwendung der Formeln (10.12) - (10.15) den Nachteil hat, oft eher Verwirrung zu stiften als eine Lösungshilfe bei konkreten Problemen zu bieten. Die Anzahlen, die bei diesen Gleichungen im Zähler und Nenner eingehen, lassen sich alternativ auch über Baumdiagramme oder Kontingenztabellen für absolute Häufigkeiten mit Randverteilungen erschließen. Man kann viele Fragestellungen, bei denen bedingte Wahrscheinlichkeiten im Spiel sind, auf diese Weise lösen ohne den Satz von Bayes zu kennen (vgl. hierzu das nachstehende Beispiel einschließlich der Aufgaben 10.5 - 10.7).

Beispiel 10.4: Bedingte Wahrscheinlichkeiten bei Drogentherapien

Es sei eine Gruppe von 60 drogenabhängigen Personen betrachtet, die stationär (Ereignis A) oder ambulant (Ereignis $\overline{A}$) behandelt werden.[9] Alle Personen werden einem HIV-Test unterzogen. Bei 15 Personen fällt der Test positiv aus

[8] Der Satz von Bayes existiert auch in einer allgemeineren Fassung für Zusammenhänge zwischen *mehr als zwei* bedingten Wahrscheinlichkeiten – vgl. z. B. Mosler / Schmid (2011, Abschnitt 1.3.3).

[9] Dieses Beispiel ist adaptiert aus Caputo / Fahrmeir / Künstler / Lang / Pigeot / Tutz (2009, Kapitel 4).

(Ereignis B), bei den anderen 45 negativ (Ereignis $\overline{B}$). 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 $P(\overline{B}) = \frac{45}{60} = 0{,}75$ die Wahrscheinlichkeiten dafür, dass diese Person HIV-positiv resp. HIV-negativ ist;
- $P(A|B) = 0{,}8$ und $P(A|\overline{B}) = 0{,}4$ 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 Person stationär therapiert wird und auch HIV-positiv ist, lässt sich aus (10.14) 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 $P(A \cap \overline{B})$, dass die ausgewählte Person stationär therapiert wird und HIV-negativ ist, den Wert

$$P(A \cap \overline{B}) = P(A|\overline{B}) \cdot P(\overline{B}) = 0{,}4 \cdot 0{,}75 = 0{,}3.$$

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 $P(A \cap \overline{B})$ zusammen. Dies folgt aus dem Axiom K3 von Kolmogoroff. Dieses ist anwendbar, weil – vgl. die rechte Hälfte der vierteiligen Abbildung 10.1 – die Mengen $A \cap B$ und $A \cap \overline{B}$ disjunkt sind und ihre Vereinigung A ergibt. Es gilt also

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) = 0{,}3 + 0{,}2 = 0{,}5.$$

Bei Unabhängigkeit der Ereignisse A und B dürfte die Wahrscheinlichkeit für A nicht davon abhängen, ob der Fall B oder $\overline{B}$ vorliegt. Bei obigem Beispiel liegt aber eine solche Abhängigkeit vor.

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 8.4):

	Test positiv (B)	Test negativ ($\overline{B}$)	Zeilensummen
stationär (A)	*12*	*18*	30
ambulant ($\overline{A}$)	3	27	30
Spaltensummen	*15*	*45*	*60*

Tab. 10.2: *Kontingenztabelle für vier Kategorien von Suchtpatienten*

Aufgabe 10.5 - 8

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 10.2 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.

Exkurs 10.1: Das „Ziegenproblem“

Eine interessante Denkaufgabe, die einen direkten Bezug zum Thema „Bedingte Wahrscheinlichkeiten“ hat und in den Medien hitzig diskutiert wurde, ist das sog. „Ziegenproblem“, im angelsächsischen Sprachraum auch „Monty Hall Problem“ genannt. Das Problem, dem der Wissenschaftsjournalist Gero von Randow sogar ein ganzes Buch (Randow (1992)) widmete, 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 meckernde 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 und glotzt erstaunt ins Publikum. 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. Er vermittelt dem Kandidaten mit seinem Einschreiten eine möglicherweise wahlbeeinflussende Zusatzinformation. Dass die vom Kandidaten gewählte Tür die Nummer 1 erhält, stellt keine Beschränkung der Allgemeinheit dar. 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.

Das Problem wird übersichtlicher, wenn man die Situation unter Verwendung von Wahrscheinlichkeiten und bedingten Wahrscheinlichkeiten darstellt. Möge A_i das Ereignis bezeichnen, 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 (10.5)

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}.$$

Der Kandidat hat sich zunächst für Tür 1 entschieden und diese Entscheidung führt, wenn auf ihr beharrt wird, mit der Wahrscheinlichkeit $P(A_1) = \frac{1}{3}$ zum Gewinn des Autos. Man könnte meinen, dass die Intervention des Showmasters, die das Ausscheiden der Wahl von Tor 3 impliziert, dazu führt, dass die Chancen für die richtige Entscheidung zwischen den beiden verbleibenden Türen gleich groß sind, also je mit der Wahrscheinlichkeit $\frac{1}{2}$ verknüpft sind. Letzteres würde bedeuten, dass eine Änderung der ursprünglichen Entscheidung für Tür 1 keine Verbesserung der Gewinnwahrscheinlichkeit mit sich brächte.

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 beiden Summanden $P(A_2)$ und $P(A_3)$ den Wert 0 hat. Ein Wechsel der ursprünglichen gewählten Tür verdoppelt also die Gewinnchancen. Man kann sich diesen Sachverhalt auch so verdeutlichen: Trifft A_1 zu, so führt ein Festhalten an der ursprünglichen Entscheidung für Tür 1 zum Gewinn des Autos. Wäre A_2 zutreffend, hätte der Showmaster Tür 3 geöffnet und ein Wechsel der ursprünglichen Entscheidung 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 ratsam.

Die Konsequenzen einer Korrektur der ursprünglichen Entscheidung für Tür 1 lässt sich formalisieren. Die Strategie des Showmasters beinhaltet, dass

- $P(S_2|A_3) = P(S_3|A_2) = 1$ (Strategie des Showmasters, wenn der Kandidat mit Tür 1 eine Tür mit Ziege gewählt hat)
- $P(S_2|A_1) = P(S_3|A_1) = 0{,}5$ (Strategie des Showmasters, wenn der Kandidat mit Tür 1 die Tür mit dem Auto gewählt hat).

Die Wahrscheinlichkeit für den Gewinn des Autos bei Korrektur der ursprünglichen Entscheidung für Tür 1 setzt sich additiv zusammen aus der Wahrscheinlichkeit $P(S_3 \cap A_2)$ dafür, dass der Showmaster Tür 3 öffnet und gleichzeitig A_2 zutrifft und der Wahrscheinlichkeit $P(S_2 \cap A_3)$ dafür, dass er Tür 2 öffnet und A_3 zutrifft. Bei sukzessiver Anwendung von (10.14) und dem Satz von Bayes (10.15) erhält man für die genannten beiden Wahrscheinlichkeiten

$$\begin{aligned} P(S_3 \cap A_2) = P(A_2|S_3) \cdot P(S_3) &= \frac{P(S_3|A_2) \cdot P(A_2)}{P(S_3)} \cdot P(S_3) \\ &= \underbrace{P(S_3|A_2)}_{1} \cdot \underbrace{P(A_2)}_{\frac{1}{3}} = \frac{1}{3}. \\ P(S_2 \cap A_3) = P(A_3|S_2) \cdot P(S_2) &= \frac{P(S_2|A_3) \cdot P(A_3)}{P(S_2)} \cdot P(S_2) \\ &= \underbrace{P(S_2|A_3)}_{1} \cdot \underbrace{P(A_3)}_{\frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

Für die Summe aus $P(S_3 \cap A_2)$ und $P(S_2 \cap A_3)$, die die Gewinnwahrscheinlichkeit bei Wechsel von Tür 1 auf die noch nicht vom Showmaster ausgeschlossene alternative Tür darstellt, ergibt sich also auch auf diesem Wege der Wert $\frac{2}{3}$.

10.4 Wahrscheinlichkeitsverteilungen

Interpretation von Daten als Ergebnis von Zufallsvorgängen

Eine Kernaufgabe der beschreibenden Statistik besteht darin, Ausprägungen von Merkmalen, also Daten, anhand ausssagekräftiger 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 (s. Abbildung 4.10). Erzeugt man durch erneutes n-maliges Würfeln neue Daten, so resultiert ein anderer Datensatz, dessen Elemente wieder über ein Zufallsexperiment generiert wurden. Bei einem Datensatz zu Bruttoverdiensten im EU-Land Spanien bietet es sich an, die Daten zu Einkommensklassen zu gruppieren und anhand eines Histogramms zu präsentieren (s. Abbildung 4.7). Zieht man aus der Grundgesamtheit aller Beschäftigten in Spanien eine Zufallsstichprobe, erhält man ebenfalls einen Datensatz, dessen Elemente Ergebnis eines Zufallsvorgangs sind. Ein Vorgang, bei dem Zufallseinflüsse ins Spiel kommen, nennt man *stochastisch* – im Gegensatz zu einem *deterministischen* Vorgang, bei dem der Ausgang exakt vorhersagbar ist.

Stochastische und deterministische Modelle

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, also mit stochastischen Modellen, und der Berechnung von Wahrscheinlichkeiten auf der Basis solcher Modelle oder anhand kombinatorischer Überlegungen befasst, nennt man **Wahrscheinlichkeitsrechnung**. Für die schließende Statistik ist die Verwendung *stochastischer* Modelle typisch; sie benötigt die Wahrscheinlichkeitsrechnung als Fundament.

Diskrete und stetige Zufallsvariablen

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.

Zufallsvariablen als Abbildungen

Eine Zufallsvariable lässt sich als eine Abbildung interpretieren, die jedem Ereignis, also jeder Teilmenge der Ereignismenge eines Zufallsvorgangs, eine Wahrscheinlichkeit zuordnet. Bei einer stetigen Zufallsvariablen interessiert man sich weniger für die Eintrittswahrscheinlichkeit dafür, dass die Variable einen bestimmten Einzelwert annimmt. Vielmehr geht es hier eher 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“ für Hersteller und Verbraucher unerheblich, mit welcher Wahrscheinlichkeit eine Brenndauer von genau a Stunden erreicht wird. Relevanter ist es Informationen darüber zu haben, mit welcher Wahrscheinlichkeit eine Brenndauer von mindestens a Stunden erreicht wird oder eine Brenndauer, die zwischen a und b Stunden liegt, etwa mit $a = 1000$ und $b = 2000$.

Einsatzbeispiele für Wahrscheinlichkeitsmodelle

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 groß postuliert werden. Dieses Modell, das auch als **diskrete Gleichverteilung** angesprochen wird, genauer als Spezialfall der diskreten Gleichverteilung (6 Realisationen), ist in Kapitel 11 näher beschrieben. Es ist ein stochastisches Modell, weil es Zufallseinflüsse anhand der Eintrittswahrscheinlichkeiten berücksichtigt.

Interaktives Lernobjekt „Augenzahlen“ (mit Modell)

Es gibt viele weitere Beispiele, bei denen der Einsatz von Verteilungsmodellen sinnvoll ist. Erfasst man etwa die Körpergröße aller 30-jährigen Männer in Deutschland und präsentiert diese 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 12) – 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, das Zentrum und die Streuung der als Modell 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

$$F(x) := P(X \leq x) \tag{10.17}$$

- anhand der Verteilungsfunktion

zuordnet. Zwecks Unterscheidung von der empirischen Verteilungsfunktion (4.5) nennt man die Funktion (10.17) auch präziser **theoretische Verteilungsfunktion**.[10] Zur Charakterisierung einer *empirischen* Verteilung wurde neben der empirischen Verteilungsfunktion – in Abschnitt 4.2 auch relative kumulierte Häufigkeitsverteilung genannt – die relative Häufigkeitsverteilung herangezogen (vgl. Abbildung 4.10 oder Abbildung 4.11). Zur Beschreibung einer *theoretischen* Verteilung kann man außer der Verteilungsfunktion (10.17) ebenfalls eine zweite Funktion heranziehen. Der in der beschreibenden Statistik verwendeten relativen Häufigkeitsverteilung 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.

- anhand von Wahrscheinlichkeits- oder Dichtefunktion

Kenngrößen theoretischer Verteilungen

Wie bei empirischen Verteilungen lassen sich 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.

Zwischen empirischen Verteilungen von Merkmalen (Häufigkeitsverteilungen) und theoretischen Verteilungen von Zufallsvariablen gibt es jedenfalls

[10] Nur um die Notation nicht zu sehr zu komplizieren, wird in diesem Manuskript sowohl für die empirische als auch für die theoretische Verteilungsfunktion dieselbe Bezeichnung $F(x)$ verwendet.

auffällige Analogien. Wichtig ist aber eine Unterscheidung beider Konzepte, also von Daten- und Modellebene. Verteilungen von Zufallsvariablen sind als Modelle zu verstehen, die oft gut geeignet sind, Strukturen und Gesetzmäßigkeiten, die großen Datenmengen zugrunde liegen können, zu approximieren. Kenngrößen theoretischer Verteilungen – von den Kenngrößen empirischer Verteilungen klar abzugrenzen – 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 14 - 15).

Tabelle 10.3 betont einige der erwähnten Analogien zwischen Häufigkeitsverteilungen und Verteilungen für Zufallsvariablen:

	Beschreibende Statistik	Wahrscheinlichkeitsrechnung
Bezugsrahmen	Menge aller untersuchungsrelevanten Merkmalsträger	Menge der möglichen Ausprägungen einer Zufallsvariablen
Verteilungen	Empirische Verteilung eines Merkmals, festgelegt durch	Theoretische Verteilung einer Zufallsvariablen, festgelegt durch
	- relative Häufigkeiten	- Wahrscheinlichkeits- oder Dichtefunktion
	- empirische Verteilungsfunktion	- theoretische Verteilungsfunktion
Kenngrößen	Mittelwert, Median, empirische Quantile, empirische Varianz	Erwartungswert, theoretische Quantile, theoretische Varianz

Tab. 10.3: *Analogien zwischen empirischen und theoretischen Verteilungen*

Wichtige Verteilungen (mit Anwendungsbeispielen)

In Kapitel 11 werden einige diskrete und in Kapitel 12 einige stetige Verteilungen und ihre Kenngrößen ausführlich vorgestellt. An dieser Stelle sollen nur beispielhafte Anwendungsfelder der dort behandelten Verteilungsmodelle genannt werden.

Diskrete Verteilungen:

- Diskrete Gleichverteilung: Glücksspiele (Würfeln, Roulette);
- Binomialverteilung (Bernoulli-Verteilung als Spezialfall): Glücksspiele (z. B. Münzwurfexperimente), 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 der hier nur erwähnten diskreten und stetigen Verteilungen findet man bei Fahrmeir / Künstler / Pigeot / Tutz (2010, Kapitel 5 - 6). Ob ein bestimmtes Verteilungsmodell, etwa das Modell 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.

11 Diskrete Zufallsvariablen

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.

Vorschau auf das Kapitel

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.

11.1 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.

Beschreibung diskreter Zufallsvariablen:

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 $x_1, \ldots, x_k$ 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; $i = 1, \ldots, k$. Die Funktion f, die jeder Ausprägung x_i eine Eintrittswahrscheinlichkeit p_i zuordnet, heißt **Wahrscheinlichkeitsfunktion** von X (engl: *probability density function*, kurz *pdf*). Damit die Wahrscheinlichkeitsfunktion nicht

- anhand der Wahrscheinlichkeitsfunktion

[1] Zum Begriff „abzählbar unendlich“ vgl. erneut die Fußnote in Abschnitt 2.2.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_11

nur auf der Trägermenge $\{x_1, \ldots, x_k\}$, sondern für alle reellen Zahlen x erklärt ist, setzt man sie Null für alle x mit $x \neq x_i$:

$$f(x) = \begin{cases} p_i & \text{für } x = x_i;\ i = 1, 2, \ldots, k \\ 0 & \text{für alle sonstigen } x. \end{cases} \tag{11.1}$$

Die Wahrscheinlichkeitsfunktion $f(x)$ lässt sich anhand eines Stab- oder Säulendiagramms mit k Stäben bzw. Säulen der Länge $p_1, p_2, \ldots, p_k$ darstellen. Sie kann nur nicht-negative Werte annehmen. Ferner muss die Summe der Eintrittswahrscheinlichkeiten $p_1, p_2, \ldots, p_k$ in (11.1) 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 schon in (10.17) eingeführten **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. Offenbar hat $F(x)$ 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. Anstelle von (11.2) kann man demnach hier auch schreiben:

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ p_1 & \text{für } x_1 \leq x < x_2 \\ \vdots & \vdots \\ p_1 + p_2 + \ldots + p_{k-1} & \text{für } x_{k-1} \leq x < x_k \\ 1 & \text{für } x \geq x_k. \end{cases} \tag{11.2}$$

Es gibt eine weitere Parallele zwischen den relativen Häufigkeitsverteilungen der beschreibenden Statistik und den Verteilungen diskreter Zufallsvariablen. Durch Aufsummieren relativer Häufigkeiten kommt man

[2] Wenn 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 also z. B. präziser $f_X(x)$ und $F_X(x)$ anstelle von $f(x)$ und $F(x)$ schreiben.

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.

Die diskrete Gleichverteilung

Besonders einfach ist der Fall einer diskreten Verteilung, bei der in (11.1) 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 $x_1, x_1, \ldots, x_k$ gehen aus (11.1) und (11.2) als Spezialfall hervor, wenn dort für alle Eintrittswahrscheinlichkeiten p_i der Wert $p = \frac{1}{k}$ eingesetzt wird. Aus (11.2) wird

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ \frac{1}{k} & \text{für } x_1 \le x < x_2 \\ \vdots & \vdots \\ \frac{k-1}{k} & \text{für } x_{k-1} \le x < x_k \\ 1 & \text{für } x \ge x_k. \end{cases} \tag{11.3}$$

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 $x_1 = 1, x_2 = 2, \ldots, x_6 = 6$, die alle die Eintrittswahrscheinlichkeit $p = \frac{1}{6}$ aufweisen. Die Wahrscheinlichkeitsfunktion $f(x)$ der zugehörigen Gleichverteilung ist im linken Teil von Abbildung 11.1 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 $x_1 = 1, x_2 = 2, \ldots, x_6 = 6$ jeweils Sprünge der Höhe $\frac{1}{6}$ auf.

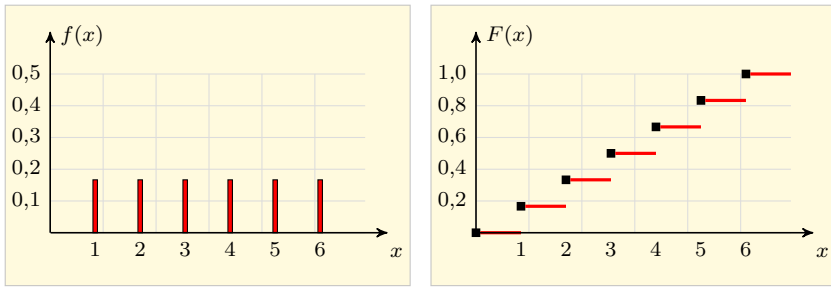
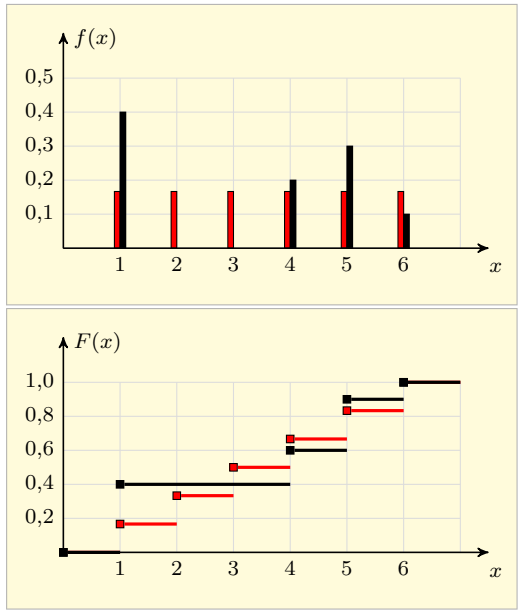


Abb. 11.1: *Wahrscheinlichkeits- und Verteilungsfunktion der diskreten Gleichverteilung mit $p = \frac{1}{6}$ (Würfeln mit einem Würfel)*

Beispiel 11.1: Daten- und Modellebene beim Würfelexperiment

Zusammenhänge zwischen Verteilungen diskreter Zufallsvariablen (theoretische Verteilungen) und relativen Häufigkeitsverteilungen (empirische Verteilungen) lassen sich anhand des statistischen Experiments „Würfeln mit einem Würfel“ gut sichtbar machen, wenn man das Experiment n-mal durchführt mit hinreichend groß gewähltem n.

Interaktives Lernobjekt „Augenzahlen“ (mit Modell)

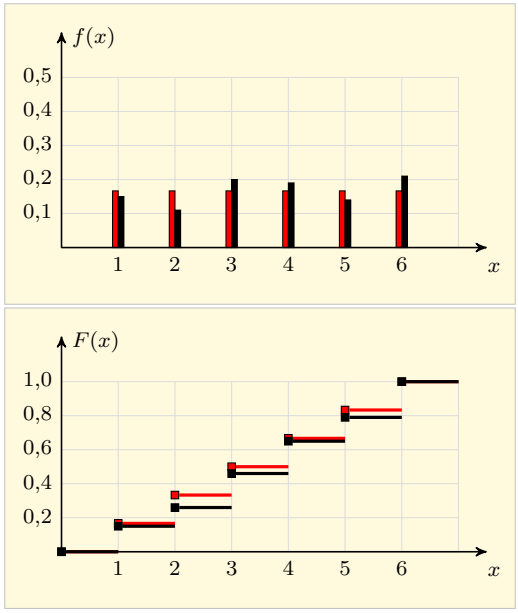


i	f_i	F_i
1	0,4	0,4
2	0	0,4
3	0	0,4
4	0,2	0,6
5	0,3	0,9
6	0,1	1,0

Abb. 11.2: *Relative Häufigkeiten für die Augenzahlen bei 10-fachem Würfeln und Modell (diskrete Gleichverteilung mit $p = \frac{1}{6}$)*

Abbildung 11.2 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 11.1 dargestellte Modell der diskreten Gleichverteilung mit dem Parameter $p = \frac{1}{6}$ eingezeichnet (graue Säulen; im e-Buch rot). Neben der Abbildung sind in einer Tabelle die beobachteten relativen Häufigkeiten f_i für die einzelnen Augenzahlen und die wieder mit F_i abgekürzten Werte der empirischen Verteilungsfunktion an den Stellen $x = x_i$ aufgeführt ($i = 1, 2, \ldots, 6$). 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 11.3 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



i	f_i	F_i
1	0,15	0,15
2	0,11	0,26
3	0,20	0,46
4	0,19	0,65
5	0,14	0,79
6	0,21	1,00

Abb. 11.3: *Relative Häufigkeiten für die Augenzahlen bei* 100*-fachem Würfeln und Modell (diskrete Gleichverteilung mit* $p = \frac{1}{6}$*)*

Abbildung 11.3 mit Abbildung 11.2, 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.

Der Wert $f_1 = 0{,}4$ in der Tabelle neben Abbildung 11.2 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 11.3, 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 11.1: 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 (*p*ersönliche *I*dentifikations*n*ummern), 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 Pins 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 bekannt gewordener 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 offenbar sehr deutlich von einer diskreten Gleichverteilung. Es überrascht daher auch nicht, dass Banken ihren Kunden die Pins von Bankkarten zuteilen. Die Pin-Codes 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 **Zweipunkt-Verteilung** findet. Diese Verteilung liegt vor, wenn eine Zufallsvariable X nur zwei Ausprägungen aufweist, etwa x_1 und x_2 oder A und $\overline{A}$. 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 (11.1) hat dann die spezielle Gestalt

$$f(x) = \begin{cases} p & \text{für } x = x_1; \\ 1-p & \text{für } x = x_2; \\ 0 & \text{für alle sonstigen } x. \end{cases} \tag{11.4}$$

Charakterisierung der Bernoulli-Verteilung

Durch (11.4) oder die Verteilungsfunktion

$$F(x) = P(X \le x) = \begin{cases} 0 & \text{für } x < x_1; \\ p & \text{für } x_1 \le x < x_2; \\ 1 & \text{für } x \ge x_2 \end{cases} \tag{11.5}$$

ist eine Bernoulli-Verteilung vollständig definiert. Ihre Gestalt hängt vom Parameter p ab. Eine mit dem Parameter p bernoulli-verteilte Zufallsvariable X bezeichnet man als $Be(p)$-verteilt und verwendet hierfür die Notation $X \sim Be(p)$ (lies: X ist *bernoulli-verteilt* mit dem Parameter p). Das Merkmal „Ergebnis eines Münzwurfexperiments“ (einmaliger

Münzwurf mit den möglichen Realisationen „Zahl“ und „Kopf“) ist z. B. bernoulli-verteilt mit $p = 0{,}5$, wenn man eine „faire“ Münze voraussetzt. Ein statistisches Experiment, dessen Ausgang durch ein bernoulliverteiltes Merkmal beschrieben wird, heißt **Bernoulli-Experiment**.

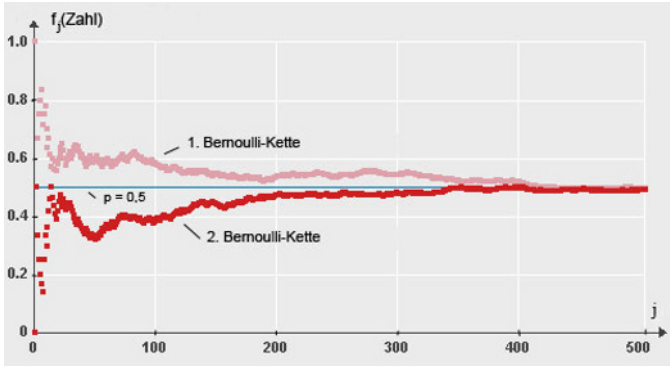


Abb. 11.4: *Relative Häufigkeit für „Zahl“ bei 500-fachem Münzwurf und Vergleich mit dem Parameter $p = 0{,}5$ der Bernoulli-Verteilung*

Interaktives Lernobjekt „Münzwurf“

Abbildung 11.4 vermittelt eine Aussage, die schon in Abbildung 10.2 visualisiert wurde. Sie zeigt den Entwicklungspfad der relativen Häufigkeiten $f_j = f_j(\text{Zahl})$ des Auftretens der Ausprägung „Zahl“ nach j Münzwurfexperimenten, wobei $j = 1, 2, \ldots, n$ mit $n = 500$. Es wird also ein Bernoulli-Experiment wiederholt durchgeführt – man spricht in diesem Zusammenhang unter der hier erfüllten Voraussetzung der Unabhängigkeit der Einzelexperimente auch von einer **Bernoulli-Kette** – und die Zwischenstände $f_j(\text{Zahl})$ bis zum Endstand fortlaufend visualisiert.

Abbildung 11.4 zeigt nicht nur einen, sondern zwei Entwicklungspfade, also zwei Bernoulli-Ketten. Der Endstand $f_{500}(\text{Zahl})$ der beobachteten relativen Häufigkeit liegt in beiden Fällen sehr dicht am Wert $p = 0{,}5$ der Eintrittswahrscheinlichkeit für „Zahl“, gegen den die Bernoulli-Ketten für $n \to \infty$ stochastisch konvergieren. Wenn man die Ausprägungen x_1 und x_2 zu 1 und 0 umcodiert (vgl. (11.3)), wird eine Bernoulli-Verteilung auch **Null-Eins-Verteilung** genannt.

11.2 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 $\overline{x}$ eines Datensatzes $x_1, x_2, \ldots, x_n$, der sich auf ein diskretes Merkmal X mit k Ausprägungen $a_1, a_2, \ldots a_k$ bezieht, lässt sich gemäß (5.4) als Summe der mit den relativen Häufigkeiten gewichteten Merkmalsausprägungen darstellen, also durch $a_1 f_1 + a_2 f_2 + \ldots + a_k f_k$. In ähnlicher Weise lässt sich auch der Schwerpunkt der Verteilung der diskreten Zufallsvariablen (11.1) charakterisieren. Man bildet hier die Summe $x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$ der mit den Eintrittswahrscheinlichkeiten $p_1, p_2, \ldots, p_k$ 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 (11.1) definierten diskreten Zufallsvariablen ist also gegeben durch

Erwartungswert und Varianz einer diskreten Zufallsvariablen

$$\mu := E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \sum_{i=1}^{k} x_i \cdot p_i. \tag{11.6}$$

Aufgabe 11.1

Die Merkmalsausprägungen $a_1, a_2, \ldots a_k$ und die relativen Häufigkeiten $f_1, f_2, \ldots f_k$ aus Kapitel 5 werden also hier, bei der Charakterisierung theoretischer Verteilungsmodelle, durch die Realisationen $x_1, x_2, \ldots x_k$ einer diskreten Zufallsvariablen und deren Eintrittswahrscheinlichkeiten $p_1, p_2, \ldots p_k$ 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 (11.1) mit $\mu = E(X)$ die Darstellung

$$\sigma^2 := V(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i. \tag{11.7}$$

Die Darstellung (11.6) geht in (11.7) über, wenn man in (11.6) anstelle von X den Term $(X - \mu)^2$ einsetzt. Es gilt also

$$\sigma^2 = E\left[(X - \mu)^2\right]. \tag{11.8}$$

Für die Varianz ist manchmal die Darstellung

$$\sigma^2 = E(X^2) - \mu^2 \tag{11.9}$$

nützlich, die sich aus (11.8) ergibt, wenn man dort den Term in der eckigen Klammer ausmultipliziert und dann den Erwartungswert gliedweise anwendet – s. hierzu auch die noch folgenden Formeln (11.11) und (11.13).

Der Varianzdarstellung (11.9) entspricht auf der empirischen Ebene die Zerlegungsformel (5.7).

Die **Standardabweichung** σ (lies: *sigma*) von X ist definiert durch

$$\sigma := \sqrt{V(X)}. \tag{11.10}$$

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.

Lineartransformationen bei Zufallsvariablen

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**).

Unterzieht man eine Zufallsvariable X mit Erwartungswert $\mu = E(X)$ einer Lineartransformation $Y = aX + b$, so gilt

$$E(Y) = E(aX + b) = a \cdot E(X) + b \tag{11.11}$$

$$V(Y) = V(aX + b) = a^2 \cdot V(X). \tag{11.12}$$

Für den Erwartungswert und die Varianz der Summe zweier unabhängiger Zufallsvariablen X und Y seien hier ohne Beweis die Darstellungen

$$E(X + Y) = E(X) + E(Y) \tag{11.13}$$

$$V(X + Y) = V(X) + V(Y). \tag{11.14}$$

wiedergegeben. [3] Die Gleichungen (11.13) und (11.14) gelten entsprechend auch für die Summen von n unabhängigen Zufallsvariablen ($n \geq 2$).

[3] Der Begriff der „Unabhängigkeit“ von zwei oder mehreren Zufallsvariablen wird in Abschnitt 13.1 formalisiert. Im Gegensatz zu (11.13) gilt die Darstellung (11.14) nicht mehr bei Abhängigkeit von X und Y; in diesem Falle ist sie durch (13.14) zu ersetzen. Die Gültigkeit von (11.14) ist schon gegeben, wenn man anstelle der Unabhängigkeit von X und Y lediglich fehlende lineare Abhängigkeit voraussetzt (Unkorreliertheit; vgl. hierzu Abschnitt 13.2).

Kenngrößen der Null-Eins-Verteilung

Erwartungswert und Varianz der Null-Eins-Verteilung ergeben sich unmittelbar aus den allgemeineren Formeln (11.6) und (11.7) 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 (11.9) zurückgreift:

$$\mu = 1 \cdot p + 0 \cdot (1 - p) = p. \tag{11.15}$$

$$\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1 - p). \tag{11.16}$$

Beispiel 11.2: Kenngrößen des Merkmals „Augenzahl“

In Abbildung 11.1 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 ($i = 1, 2, \ldots, 6$), erhält man für den Erwartungswert $\mu = E(X)$ und die Varianz $\sigma^2 = V(X)$ aus (11.6) und (11.7)

$$\mu = \sum_{i=1}^{6} x_i \cdot p_i = \frac{1}{6} \cdot \sum_{i=1}^{6} i = \frac{21}{6} = 3{,}5$$

$$\sigma^2 = \sum_{i=1}^{6} (x_i - \mu)^2 \cdot p_i = \frac{1}{6} \cdot \sum_{i=1}^{6} (i - 3{,}5)^2 = \frac{17{,}5}{6} \approx 2{,}92.$$

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

$$F(x_p) = p \qquad (0 < p < 1) \tag{11.17}$$

definiert, also durch den Wert x_p der Verteilungsfunktion $F(x)$, an dem $F(x)$ den Wert p annimmt. Der **Median** $\widetilde{x} = x_{0,5}$ 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.

Bei diskreten Verteilungen sind die Quantile durch (11.17) noch nicht eindeutig festgelegt. Bei der im zweiten Teil von Abbildung 11.1 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 11.1 erhält man so für den Median $\widetilde{x} = x_{0,5}$ den Wert $\widetilde{x} = 3$.

11.3 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 Münzwurfexperiments wurde schon genannt. Praxisrelevantere Beispiele sind etwa die Geschlechterverteilung bei Geburten, 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 $x_2 = \overline{A}$ und den Eintrittswahrscheinlichkeiten $P(A) = p$ bzw. $P(\overline{A}) = 1 - p$ 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 durchgeführten Bernoulli-Experimente und bezeichnet X die Anzahl der Ausgänge A, so ist die Zählvariable X eine diskrete Zufallsvariable mit den Ausprägungen $0, 1, \ldots, n$. Wenn man den Ausgang jedes der n Bernoulli-Experimente anhand einer Indikatorvariablen

$$X_i = \begin{cases} 1 & \text{bei Eintritt von } x_1 = A \\ 0 & \text{bei Eintritt von } x_2 = \overline{A} \end{cases} \tag{11.18}$$

beschreibt, so lässt sich X als Summe

$$X = \sum_{i=1}^{n} X_i \tag{11.19}$$

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 ($n = 1$).

Aus (11.19) kann man leicht den Erwartungswert $E(X)$ und die Varianz $V(X)$ der binomialverteilten Variablen X ableiten. Die in (11.19) eingehenden n Indikatorvariablen X_i sind voneinander unabhängig und folgen alle einer Null-Eins-Verteilung, besitzen demnach wegen (11.15) und

(11.16) den Erwartungswert $E(X_i) = p$ und die Varianz $V(X_i) = p(1-p)$. Mit den Formeln (11.13) und (11.14), die sich auch für die Summe von n unabhängigen Zufallsvariablen formulieren lassen ($n \geq 2$), folgt hieraus für die Kenngrößen $\mu = E(X)$ und $\sigma^2 = V(X)$ einer Binomialverteilung

Kenngrößen der Binomialverteilung

$$\mu = n \cdot p \tag{11.20}$$

$$\sigma^2 = n \cdot p \cdot (1-p). \tag{11.21}$$

Da eine diskrete Zufallvariable noch nicht durch Erwartungswert und Varianz alleine, sondern erst durch die Wahrscheinlichkeitsfunktion (11.1) oder – alternativ – durch die Verteilungsfunktion (11.2) 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 $\overline{A}$ beobachtet wird, die Bernoulli-Kette also die spezielle Gestalt $A, A, \ldots, A, \overline{A}, \ldots \overline{A}$ 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 10.1 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 $p^x(1-p)^{n-x}$. Für die **Wahrscheinlichkeitsfunktion** $f(x) = P(X = x)$ der Binomialverteilung gilt also

Charakterisierung der Binomialverteilung

Flash-Animation „Galton-Brett und Binomialverteilung“

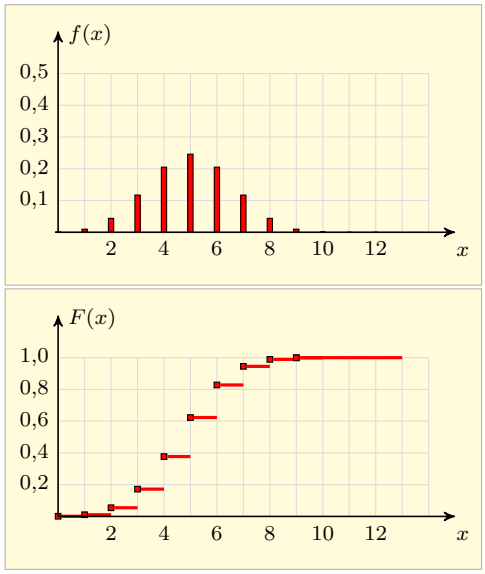
$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für } x = 0, 1, \ldots, n \\ 0 & \text{für alle sonstigen } x. \end{cases} \tag{11.22}$$

Die **Verteilungsfunktion** $F(x) = P(X \leq x)$ ist auf der Trägermenge $\{0,1,\ldots,n\}$ definiert durch

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \qquad x = 0,1,\ldots,n. \tag{11.23}$$

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 $X \sim B(n;p)$ (lies: 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.



x	$f(x)$	$F(x)$
0	0,0010	0,0010
1	0,0098	0,0107
2	0,0439	0,0547
3	0,1172	0,1719
4	0,2051	0,3770
5	0,2461	0,6230
6	0,2051	0,8281
7	0,1172	0,9453
8	0,0439	0,9893
9	0,0098	0,9990
10	0,0010	1,0000
11	0,0000	1,0000
12	0,0000	1,0000

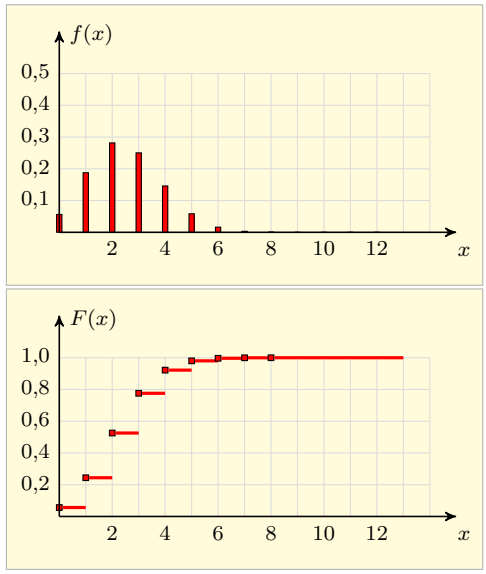
Abb. 11.5: *Binomialverteilung mit* $n = 10$ *und* $p = 0{,}50$

Interaktives Lernobjekt „Binomialverteilung“

Abbildung 11.5 zeigt Wahrscheinlichkeits- und Verteilungsfunktion einer $B(10;0{,}5)$-verteilten Zufallsvariablen. Der Tabelle neben der Grafik entnimmt man z. B., 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 (11.22). Durch Aufsummieren von Werten der Wahrscheinlichkeitsfunktionen ergeben sich also die Werte der Verteilungsfunktion. Umgekehrt kann man aus $F(x)$ durch Differenzenbildung Werte der Wahrscheinlichkeitsfunktion $f(x)$ gewinnen. Der oben 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 also eine der beiden Funktionen $f(x)$ und $F(x)$ zu tabellieren.

Die Wahrscheinlichkeitsfunktion (11.22) ist für $p = 0{,}5$ symmetrisch bezüglich des Erwartungswerts. Für $p \neq 0{,}5$ gilt dies nicht mehr, wie Abbildung 11.6 beispielhaft illustriert. Die Wahrscheinlichkeitsfunktion ist hier links vom Erwartungswert $\mu = 2{,}5$ steiler.

In Tabelle 19.1 des Anhangs sind Verteilungsfunktionen von Binomialverteilungen für $n = 1,2,\ldots,20$ und $p = 0{,}05, 0{,}10, 0{,}20, \ldots, 0{,}50$ 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. Bezüglich der Verwendung von *R* sei



x	$f(x)$	$F(x)$
0	0,0563	0,0563
1	0,1877	0,2440
2	0,2816	0,5256
3	0,2503	0,7759
4	0,1460	0,9219
5	0,0584	0,9803
6	0,0162	0,9965
7	0,0031	0,9996
8	0,0004	1,0000
9	0,0000	1,0000
10	0,0000	1,0000
11	0,0000	1,0000
12	0,0000	1,0000

Abb. 11.6: *Binomialverteilung mit* $n = 10$ *und* $p = 0{,}25$

auf die Einführungen von Wollschläger (2013) und Ligges (2014) verwiesen.[4]

Beispiel 11.3: Anwendung der Binomialverteilung

Interaktives Lernobjekt „Rechnen mit 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 10.2 elementar unter Verwendung des Laplace-Ansatzes (10.5) 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 Gewinne und ein Anteil von $1 - p$ Nieten sind,

[4] Bei Verwendung von SPSS findet man die Wahrscheinlichkeitsfunktion und die Verteilungsfunktion der Binomialverteilung sowie anderer Verteilungen im Menü „Transformieren / Variable berechnen“. Wahrscheinlichkeitsfunktion und Verteilungsfunktion der Binomialverteilung sind dort mit PDF.BINOM(..) resp. mit CDF.BINOM(..) abgekürzt. Bei *R* ist die Wahrscheinlichkeitsfunktion über *dbinom*(x, n, p) und die Verteilungsfunktion über *pbinom*(x, n, p) zugänglich.

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 im Falle $p = 0{,}05$ 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 $1 - F(3) = 0{,}0159$.

Aufgabe 11.2

Exkurs 11.2: Fiasko beim Zentralabitur 2008 in NRW

In Nordrhein-Westfalen gab es beim Zentralabitur 2008 erhebliche Kritik an einer Aufgabe zur Wahrscheinlichkeitsrechnung, die für Schüler der Mathematik-Leistungskurse konzipiert war. Die Kritik wurde im Juni 2008 von *Spiegel online* aufgegriffen. Den Schülern wurde angeboten die Prüfung zu wiederholen.

Die umstrittene Aufgabe bezog sich auf Trefferquoten von D. Nowitzki, Mannschaftsführer der deutschen Basketball-Mannschaft bei der Sommerolympiade 2008 in Peking. Der erste Teil der Aufgabe lautete wie folgt:

Der deutsche Basketball-Profi Dirk Nowitzki spielte in der amerikanischen Profiliga 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)

An der Aufgabe lässt sich verdeutlichen, wie wichtig es ist, zwischen empirischen Befunden (Datenebene) und Modellansätzen zur approximativen Beschreibung solcher Befunde (Modellebene) zu unterscheiden. Die Analogie zum Münzwurfexperiment liegt auf der Hand; auch bei einem Freiwurf gibt es zwei mögliche Ausgänge (Korb wird getroffen / verfehlt). Sei zunächst erneut die Situation beim Münzwurfexperiment betrachtet. Man kann hier davon ausgehen, dass die Wahrscheinlichkeit p für den Eintritt des interessierenden Ereignisses, etwa „Zahl“, sich nicht ändert, wenn man eine Münze n-mal wirft. Bei einer „fairen“ Münze ist $p = 0{,}5$. Abbildung 11.4 zeigt zwei Bernoulli-Ketten, die die Vermutung einer fairen Münze zumindest visuell stützen. Wenn man bei einem Münzwurf Anhaltspunkte dafür hat, dass die Wahrscheinlichkeit p für „Zahl“ nicht der Bedingung $p = 0{,}5$ genügt, kann man p schätzen (vgl. hierzu Kapitel 13) und den Schätzwert $\widehat{p}$ heranziehen. In beiden Fällen – faire oder nicht-faire Münze – ist die Eintrittswahrscheinlichkeit für „Zahl“ von Wurf zu Wurf eine feste Größe und für die Anzahl X der „Treffer“ (Beobachtung von „Zahl“) gilt $X \sim B(n; p)$ bzw. $X \sim B(n; \widehat{p})$.

Wenn die Trefferquote des Basketballers Nowitzki in der Saison 2006/07 als repräsentativ für seine Leistung angesehen werden darf, müsste man dies in der Aufgabe durch die explizite Annahme einer festen Trefferquote von 0,904 zum

Ausdruck bringen. Eine konstante Trefferquote ist keineswegs selbstverständlich; gerade im Sport sind größere Formschwankungen an der Tagesordnung. Ohne die Annahme einer festen Trefferquote p resp. $\widehat{p}$ ist die Information über die Trefferquote in der letzten Saison lediglich ein – möglicherweise einmaliger – empirischer Befund, der nicht ausreicht für die Berechnung der verlangten Trefferwahrscheinlichkeiten. Das Modell der Binomialverteilung setzt ja voraus, dass die n in (11.19) eingehenden Indikatorvariablen X_i (mit dem Wert 1 bei Eintritt des Ereignisses „Treffer“ und 0 beim Ereignis „kein Treffer“) alle derselben Bernoulli-Verteilung (11.4) folgen.

Würde man die Aufgabe um die Annahme einer konstanten Trefferquote 0,904 ergänzen, wäre 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 würde dann z. B. für $f(8)$ nach (11.22) den Wert

$$f(8) = \binom{10}{8} \cdot 0{,}904^8 \cdot 0{,}096^2 = 45 \cdot 0{,}4460129 \cdot 0{,}009216 \approx 0{,}185$$

errechnen. 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. Die Aufgabe (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 heraus zu suchen, 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.

11.4 Die hypergeometrische Verteilung

Die Binomialverteilung beschreibt das Zufallsverhalten der Zählvariablen X aus (11.19) 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 $x_2 = \overline{A}$ und $P(A) = p$ bzw. $P(\overline{A}) = 1 - p$ innerhalb der Bernoulli-Kette auftrat. Als Beispiele wurden Münzwurf- oder auch Würfelexperimente angeführt, wenn man bei letzteren nur zwischen zwei Ausgängen differenziert (etwa „gerade / ungerade Augenzahl“).

Varianten des Urnenmodells

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.

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 (11.19) definierten Zählvariablen X ist bei Annahme einer Stichprobenentnahme 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.[5]

Flash-Animation „Hypergeometrische Verteilung“

Kenngrößen der hypergeometrischen Verteilung

$$\mu = n \cdot \frac{M}{N} \tag{11.24}$$

$$\sigma^2 = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1}. \tag{11.25}$$

Erwartungswert und Varianz einer $H(n, M, N)$-verteilten Zufallsvariablen X stimmen nach (11.20) und (11.21) mit dem Erwartungswert bzw. der

[5] Eine 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).

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 (11.25) auftretende Bruchterm $\frac{N-n}{N-1}$ fehlt. Da dieser Term für $n > 1$ kleiner als 1 ist (für $n = 1$ ist er 1), 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 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 aus der Varianzformel (11.25), denn es gilt dann $\frac{N-n}{N-1} = 0$ und somit $V(X) = 0$.

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 $T = \{x_{min}, \ldots, x_{max}\}$ 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 11.3).

Charakterisierung der hypergeometrischen Verteilung

Die **Wahrscheinlichkeitsfunktion** $f(x) = P(X = x)$ der hypergeometrischen Verteilung ist ebenfalls nicht so einfach ableitbar wie die der Binomialverteilung. Es gilt die Darstellung

$$f(x) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{für } x \in T \\ 0 & \text{für alle sonstigen } x, \end{cases} \tag{11.26}$$

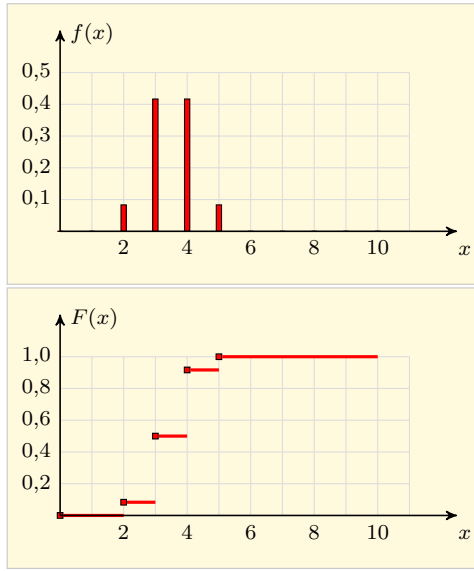
Java-Applet „Hypergeometrische Verteilung“

deren Herleitung in Exkurs 11.3 noch skizziert wird. Für die **Verteilungsfunktion** $F(x) = P(X \leq x)$ gilt dann auf der Trägermenge

$$F(x) = \sum_{k=0}^{x} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \qquad x \in T. \tag{11.27}$$

Da die Wahrscheinlichkeitsfunktion für $x \notin T$ stets 0 ist, bleibt $F(x)$ zwischen zwei benachbarten Elementen der Trägermenge auf dem Niveau des kleineren Werts, um dann in $x_{max} = min(n; M)$ den Endwert 1 anzunehmen (Treppenfunktion).

Abbildung 11.7 veranschaulicht die Wahrscheinlichkeits- und die Verteilungsfunktion einer $H(5;7;10)$-verteilten Zufallsvariablen. Der Erwartungswert $\mu = E(X)$ errechnet sich hier als $\mu = 5 \cdot \frac{7}{10} = 3{,}5$. Neben



i	f_i	F_i
0	0,0000	0,0000
1	0,0000	0,0000
2	0,0833	0,0833
3	0,4167	0,5000
4	0,4167	0,9167
5	0,0833	1,0000
6	0,0000	1,0000
7	0,0000	1,0000
8	0,0000	1,0000
9	0,0000	1,0000
10	0,0000	1,0000

Abb. 11.7: *Hypergeometrische Verteilung mit* $n = 5$, $M = 7$ *und* $N = 10$

den Graphen sind einige Werte der beiden Funktionen tabelliert. Die Trägermenge T der dargestellten hypergeometrischen Verteilung ist durch $T = \{x_{min}, \ldots, x_{max}\}$ gegeben mit $x_{min} = max(0; 5 - 10 + 7) = 2$ und $x_{max} = min(5; 7) = 5$. Der Abbildung entnimmt man, dass die Verteilungsfunktion $F(x)$ an der Stelle $x_{max} = 5$ auf den Endwert 1 springt.

Approximation der hypergeometrischen Verteilung

Die Tabellierung der Verteilungsfunktion ist für die hypergeometrische Verteilung viel aufwändiger als bei der Binomialverteilung, weil die Tabellen hier von drei Parametern abhängen. Aus diesem Grund wird hier auf eine Tabellierung verzichtet. 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 Auswahlsatzes $\frac{n}{N}$ immer weniger ins Gewicht fallen.

Bernoulli-Verteilung als Spezialfall

Die Binomialverteilung und die hypergeometrische Verteilung charakterisieren beide das Zufallsverhalten der Zählvariablen (11.19), allerdings unter verschiedenen Bedingungen. Die Variable (11.19) 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 $x_2 = \overline{A}$ 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 (11.4) beschreibt den Ausgang des einmaligen Bernoulli-Experiments.

Beispiel 11.4: 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 (11.24) durch $\mu = \frac{36}{49} \approx 0{,}735$ gegeben.

Für die Berechnung von Wahrscheinlichkeiten der Art „x Richtige“ oder „mindestens x Richtige“ ist der Rückgriff auf eine Tabelle mit Werten der Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ oder der Verteilungsfunktion $F(x) = P(X \leq x)$ der hypergeometrischen Verteilung am einfachsten.

Abb. 11.8: *„Lottofee“ (ARD-Lottoziehung; Quelle: Hessischer Rundfunk)*

Wenn man nicht über eine solche Tabelle verfügt, kann man die gesuchten Wahrscheinlichkeiten direkt aus (11.26) bzw. aus (11.27) bestimmen. Für das Ereignis „0 Richtige“ erhält man z. B. nach (11.26) 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

$$\binom{49}{6} = \frac{49!}{43! \cdot 6!} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13983816$$

ermittelt (vgl. auch Beispiel 10.3), repräsentiert die Anzahl der möglichen Ausgänge einer Lottoziehung. Für den Zählerterm, der die Anzahl der möglichen Ausgänge mit 0 Richtigen wiedergibt, folgt

$$\binom{43}{6} = \frac{43!}{37! \cdot 6!} = \frac{43 \cdot 42 \cdot 41 \cdot 40 \cdot 39 \cdot 38}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 6096454.$$

Die Wahrscheinlichkeit $f(0) = P(X = 0)$ für das Ereignis „0 Richtige“ ist somit

$$f(0) = \frac{\binom{43}{6}}{\binom{49}{6}} = \frac{6096454}{13983816} \approx 0{,}436,$$

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 13983816 möglichen Ausgängen einer Lottoziehung nur ein einziger Ausgang „6 Richtige“ beinhaltet, kann man $f(6) = P(X = 6)$ einfacher über

$$f(6) = \frac{1}{13983816} \approx 0{,}0000000715 = 7{,}15 \cdot 10^{-8}$$

errechnen. An Lottospieler, die 6 Richtige haben (Gewinnklasse 2), werden 8% der Lottoeinnahmen verteilt. Allerdings werden 50% der Einnahmen als Steuern abgeführt oder für festgelegte Zwecke verwendet und gar nicht ausgeschüttet.

Aufgabe 11.3-4

Die verschwindend kleine Wahrscheinlichkeit für einen Volltreffer verringert sich beim deutschen Lotto noch um den Faktor 10 auf $\frac{1}{139838160} \approx 7{,}15 \cdot 10^{-9}$, wenn man das Spiel „6 aus 49 mit Superzahl“ spielt. Die „Superzahl“ ist eine Zusatzzahl, die aus der Menge $\{0, 1, ..., 8, 9\}$ 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 $f(6) = \frac{1}{622614630} \approx 1{,}61 \cdot 10^{-9}$ der Wahrscheinlichkeitsfunktion einer hypergeometrischen Verteilung mit den Parametern $n = 6$, $M = 6$ und $N = 90$.

Exkurs 11.3: 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 (11.19) 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 – es können nicht mehr rote Kugeln gezählt als gezogen werden. Im Falle $n > M$ ist hingegen M die Obergrenze – es können nicht mehr rote Kugeln gezogen werden als 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 (11.26) kann man auf Tabelle (10.1) zurückgreifen. Der Nenner von (11.26) repräsentiert die Anzahl der Möglichkeiten, aus einer Urne mit N Kugeln insgesamt n Kugeln ohne Zurücklegen zu entnehmen. Nach Tabelle (10.1) 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 (11.26) 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.

12 Stetige Zufallsvariablen

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.

Vorschau auf das Kapitel

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 von 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.

12.1 Dichtefunktion und Verteilungsfunktion

Die in Kapitel 11 behandelten *diskreten* Zufallsvariablen sind dadurch gekennzeichnet, dass man die Anzahl ihrer Ausprägungen abzählen kann. Sie haben also endlich viele Ausprägungen oder zumindest abzählbar unendlich viele Ausprägungen, die die Trägermenge der Variablen definieren. Das Zufallsverhalten einer diskreten Zufallsvariablen X mit k Ausprägungen x_i $(i = 1, \ldots, k)$ und den Eintrittswahrscheinlichkeiten $p_i = P(X = x_i)$ lässt sich vollständig durch die in (11.1) eingeführte Wahrscheinlichkeitsfunktion $f(x)$ beschreiben. Alternativ kann man auch die Verteilungsfunktion $F(x)$ aus (11.2) zur Beschreibung heranziehen, die sich durch Aufsummieren aller Werte ergibt, die die Wahrscheinlichkeitsfunktion bis zur Stelle x annimmt.

Charakterisierung stetiger Zufallsvariablen anhand von Dichte- und Verteilungsfunktion

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*)

$$F(x) = P(X \leq x)$$

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_12

aus (10.17) vollständig charakterisieren. Eine Darstellung der Art (11.2) für die Verteilungsfunktion gibt es aber nicht, wenn die Anzahl der möglichen Werte von X nicht mehr abzählbar ist. Der Ansatz (11.1), der die Wahrscheinlichkeiten bei einer diskreten Zufallsvariablen 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

$$f(x) \geq 0 \qquad \text{für alle reellen } x \tag{12.1}$$

und hat die Eigenschaft, dass sich jeder Wert $F(x)$ der Verteilungsfunktion durch Integration der Dichte bis zur Stelle x ergibt. Es gilt also

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{für alle reellen } x. \tag{12.2}$$

Für alle Werte x, bei denen die Dichtefunktion $f(x)$ stetig ist, stimmt sie mit der Ableitung $F'(x)$ der Verteilungsfunktion überein:

$$F'(x) = f(x). \tag{12.3}$$

Aus (12.2) folgt, dass sich bei einer stetigen Zufallsvariable 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. auch die noch folgende Abbildung 12.3). Setzt man $x = b$ und $x = a$ in (12.2) ein, erhält man Darstellungen der Werte $F(b)$ und $F(a)$ der Verteilungsfunktion und hieraus für die Differenz $F(b) - F(a)$ die Gleichung

$$F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt = \int_{a}^{b} f(t)dt. \tag{12.4}$$

Da die Verteilungsfunktion eine monoton wachsende Funktion ist, die gegen 1 strebt, gilt auch, dass die Gesamtfläche unter der Dichtekurve den Wert 1 besitzt (Normierungseigenschaft):

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{12.5}$$

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

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \le x \le b \\ 0 & \text{für alle sonstigen } x \end{cases} \tag{12.6}$$

besitzt. Die Verteilungsfunktion $F(x)$ einer über $[a, b]$ rechteckverteilten Zufallsvariablen X ergibt sich gemäß (12.2) durch Integration dieser Dichte. Die Integration liefert nur im Bereich von a bis b einen von Null verschiedenen Beitrag, d. h. es ist

Charakterisierung der stetigen Gleichverteilung

$$F(x) = \begin{cases} 0 & \text{für } x < a; \\ \frac{x-a}{b-a} & \text{für } a \le x \le b; \\ 1 & \text{für } x > b. \end{cases} \tag{12.7}$$

Die Funktion (12.6) besitzt alle Eigenschaften, die eine Dichtefunktion auszeichnen. Sie ist zum einen nicht-negativ und erfüllt außerdem die Normierungseigenschaft (12.5). Letzteres ist sofort einsichtig, wenn man sich vergegenwärtigt, dass man die Integration in (12.5) 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 12.1 zeigt die Dichtefunktion (12.6) und die Verteilungsfunktion (12.7) einer Rechteckverteilung über $[a, b]$, wobei hier beispielhaft $a = 2$ und $b = 6$ gewählt wurde. Beide Funktionen sind über die Beziehung (12.3) 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 hat den Inhalt 1.

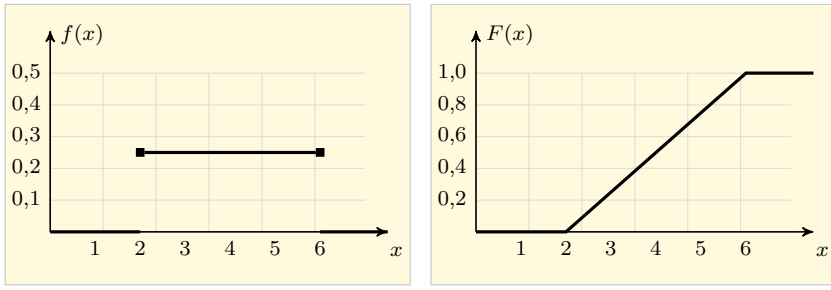


Abb. 12.1: *Dichte- und Verteilungsfunktion der Rechteckverteilung über* [2,6]

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. z. B. Toutenburg / Heumann (2008, Abschnitt 3.4) und auch den folgenden Exkurs 12.1), dass bei einer stetigen Zufallsvariablen X die Wahrscheinlichkeit $P(X = x_0)$ für jeden einzelnen Wert x_0 der Trägermenge Null ist:

$$P(X = x_0) = 0 \text{ für jeden Wert } x = x_0. \tag{12.8}$$

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 gilt es 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 12.1: 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 (12.6) 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 (12.7) 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 12.1: 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 (12.5).

12.2 Kenngrößen stetiger Verteilungen

Erwartungswert und Varianz einer stetigen Zufallsvariablen

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 (11.6) definiert. Bei *stetigen* Zufallsvariablen sind die Ausprägungen nicht mehr abzählbar, d. h. eine Summendarstellung ist nicht mehr möglich. Man kann hier, ausgehend von (11.6), durch Grenzwertbetrachtungen die Integraldarstellung

$$\mu := E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \tag{12.9}$$

gewinnen. Eine analoge, ebenfalls durch Grenzwertbetrachtungen ableitbare Aussage 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 (11.7) ist bei einer stetigen Verteilung zu ersetzen durch

$$\sigma^2 := V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx. \tag{12.10}$$

Die Varianz ist wie im diskreten Fall – vgl. (11.8) und (11.9) – nichts anderes als der Erwartungswert der quadrierten Differenz zwischen X und $\mu = E(X)$, also

$$\sigma^2 = E\left[(X - \mu)^2\right] = E(X^2) - \mu^2,$$

und auch die **Standardabweichung** σ (lies: *sigma*) ist wieder durch

$$\sigma = \sqrt{V(X)}$$

erklärt. Unverändert gültig sind auch die Eigenschaften (11.11) - (11.14), die das Verhalten von Erwartungswert und Varianz bei einfachen Li-

neartransformationen 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:

$$Z = \frac{X - \mu}{\sigma}. \tag{12.11}$$

Der Übergang von X zu Z heißt auch **z-Transformation**. Durch Einsetzen von $a = \frac{1}{\sigma}$ und $b = -\frac{\mu}{\sigma}$ in (11.11) und (11.12) 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 (12.6) oder (12.7) 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 (12.9). Es ist nämlich

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_a^b \frac{x}{b-a}\, dx \\ &= \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{1}{2 \cdot (b-a)}(b+a)(b-a) \end{aligned}$$

und somit

$$\mu = E(X) = \frac{a+b}{2}. \tag{12.12}$$

Aufgabe 12.1

Für die Berechnung der Varianz der Rechteckverteilung kann man die Varianzdarstellung $\sigma^2 = E(X^2) - \mu^2$ nutzen. Man erhält zunächst

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \int_a^b \frac{x^2}{b-a}\, dx \\ &= \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) = \frac{1}{3 \cdot (b-a)}(b^3 - a^3) \end{aligned}$$

und hieraus dann

$$\sigma^2 = E(X^2) - \mu^2 = \frac{b^3 - a^3}{3 \cdot (b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^3}{12 \cdot (b-a)} = \frac{(b-a)^2}{12}. \tag{12.13}$$

Für die über [0,10] rechteckverteilte Zufallsvariable X aus Beispiel 12.1 erhält man z. B. den Erwartungswert $\mu = 5$, die Varianz $\sigma^2 = \frac{25}{3} \approx 8{,}33$ bzw. die Standardabweichung $\sigma = \frac{5}{\sqrt{3}} \approx 2{,}89$. 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 (11.17) 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 Treppenfunktion definiert sind – eindeutig erklärt, da die Verteilungsfunktion streng monoton wächst.[1] Der Median $\widetilde{x} = x_{0,5}$ bezeichnet dann 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.

12.3 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 rührt 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.

Karl Friedrich Gauss

Dichte- und Verteilungsfunktion der Normalverteilung

Eine Zufallsvariable X folgt einer **Normalverteilung**, wenn ihre Dichte die Gestalt

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{für alle reellen } x \tag{12.14}$$

besitzt.[2] Man entnimmt der folgenden Grafik, dass die Dichte der Normalverteilung von μ und σ^2 abhängt und bezüglich μ symmetrisch ist. Anhand der allgemeinen Formeln (12.9) und (12.10) kann man verifizieren, dass μ und σ^2 der Erwartungswert resp. die Varianz der Normalverteilung sind. Für eine Zufallsvariable X mit der Dichte (12.14) sagt man, dass

[1] Das 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 Teilflächen zerlegt, dass die Teilfläche bis zum Punkt x_p den Inhalt p hat, also $p \cdot 100$ % der Gesamtfläche ausmacht. Letztere hat ja stets den Inhalt 1.

[2] Die 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.

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 mü und Varianz sigma-Quadrat*).

Für die Verteilungsfunktion der Normalverteilung gilt mit (12.2)

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{12.15}$$

Die Funktion ist nicht in geschlossener Form darstellbar. Ihre Werte lassen sich aber unter Verwendung von Näherungsverfahren ermitteln. Dichte- und Verteilungsfunktion der Normalverteilung sind nach (12.3) über die Beziehung $F'(x) = f(x)$ verbunden.

Interaktives Lernobjekt „Normalverteilung“

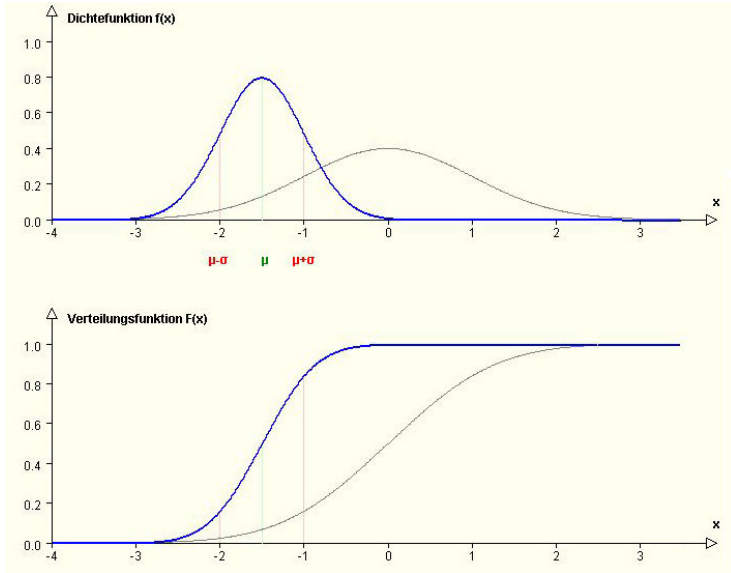


Abb. 12.2: *Darstellung von Standardnormalverteilung $N(0;1)$ und Normalverteilung $N(-1{,}5;0{,}25)$ (kräftigere Kurven)*

Abbildung 12.2 zeigt Dichte- und Verteilungsfunktion zweier Normalverteilungen. Die kräftigeren Kurven beziehen sich auf die Normalverteilung mit $\mu = -1{,}5$ und $\sigma^2 = 0{,}25$, also $\sigma = 0{,}5$. Man kann mit den Mitteln der Differentialrechnung zeigen, dass die Dichtefunktion jeder Normalverteilung in $x = \mu - \sigma$ und $x = \mu + \sigma$ Wendepunkte hat. Die beiden Wendepunkte und der Erwartungswert $\mu = E(X)$ der $N(-1{,}5;0{,}25)$-Verteilung sind im oberen Teil der Abbildung markiert. Zum Vergleich sind auch die Dichte- und Verteilungsfunktion der Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ dargestellt. Man erkennt die Symmetrie der beiden Dichten bezüglich $x = \mu$. Mit Vergrößerung der Varianz σ^2 bzw. der Standardabweichung σ verlaufen Dichte- und Verteilungsfunktion flacher.

Unterzieht man eine normalverteilte Zufallsvariable X mit Erwartungswert μ einer Lineartransformation $Y = aX + b$ mit $a \neq 0$, so gelten (11.11) und (11.12) und die transformierte Variable Y ist ebenfalls normalverteilt:

$$X \sim N(\mu;\sigma^2), Y = aX + b \longrightarrow Y \sim N(a\mu + b; a^2\sigma^2). \tag{12.16}$$

Lineartransformationen bei normalverteilten Zufallsvariablen

Für die Summe zweier unabhängiger normalverteilter Zufallsvariablen X und Y gilt ferner [3]

$$\begin{aligned} &X \sim N(\mu_X;\sigma_X^2); Y \sim N(\mu_Y;\sigma_Y^2),\ X \text{ und } Y \text{ unabh.} \\ &\rightarrow X + Y \sim N(\mu_X + \mu_Y; \sigma_X^2 + \sigma_Y^2). \end{aligned} \tag{12.17}$$

Die Aussage (12.17) lässt sich auch auf Summen von n unabhängigen Zufallsvariablen ($n \geq 2$) übertragen.

Standardisierung der Normalverteilung

Die Gestalt der Dichte- und Verteilungsfunktion der Normalverteilung hängt vom Erwartungswert μ und der Varianz σ^2 ab. Es ist aber möglich, alle Normalverteilungen auf eine einzige Grundform zurückzuführen. Gemeint ist die als **Standardnormalverteilung** bezeichnete Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$, die in Abbildung 12.2 durch die flacher verlaufenden Kurven repräsentiert ist. Hat man nämlich eine normalverteilte Zufallsvariable $X \sim N(\mu;\sigma^2)$, so kann man diese stets der speziellen Lineartransformation $Z := \frac{X-\mu}{\sigma}$ aus (12.11) unterziehen. Für die resultierende 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 (12.14) nach Einsetzen von $\mu = 0$ und $\sigma^2 = 1$ hervor. Da sie häufig verwendet wird, hat sich für sie anstelle von $f(..)$ eine spezielle Notation eingebürgert, nämlich $\phi(..)$ (lies: *Klein-Phi*):

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \tag{12.18}$$

Für die Verteilungsfunktion der Standardnormalverteilung hat sich die Bezeichnung $\Phi(..)$ (lies: *Groß-Phi*) etabliert. Sie ist durch

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt \tag{12.19}$$

[3] Der Begriff der „Unabhängigkeit“ von Zufallsvariablen wird in Abschnitt 13.1 noch formalisiert. Eine Herleitung von (12.16) und eine Verallgemeinerung von (12.17) findet man bei Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 6.3.1).

erklärt und wie (12.15) nicht in geschlossener Form darstellbar. Ihre Werte lassen sich anhand numerischer Verfahren bestimmen. Aus der zweiten Teilgrafik der Abbildung 12.2 erkennt man, dass für $\Phi(z)$ die nachstehende Symmetriebeziehung gilt:

$$\Phi(-z) = 1 - \Phi(z). \tag{12.20}$$

Interaktives Lernobjekt „Standardnormalverteilung“

Wegen $\Phi'(z) = \phi(z)$ ist der Wert $\Phi(z)$ an der Stelle $z = a$ auch interpretierbar als Inhalt der Fläche unter der Dichtekurve bis zum Punkt $z = a$.Abbildung 12.3 illustriert dies für $z = 1{,}38$. Die im oberen Teil von Abbildung 12.3 nicht markierte Restfläche unter der Dichte entspricht dem Wert $P(Z > 1{,}38) = 1 - \Phi(1{,}38)$. In Tabelle 19.2 des Anhangs sind

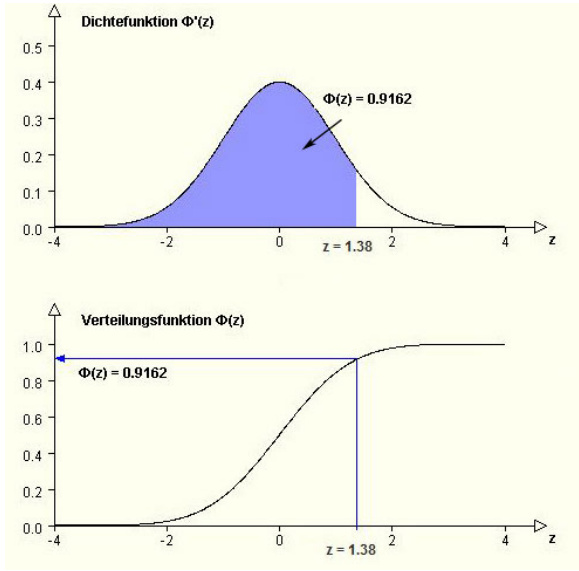


Abb. 12.3: *Interpretation von Werten der Verteilungsfunktion* $\Phi(z)$ *am Beispiel* $z = 1{,}38$

Interaktives Lernobjekt „Rechnen mit der Standardnormalverteilung“

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 (12.20) bestimmen. Mit den Werten $\Phi(z)$ aus Tabelle 19.2 kann man Werte $F(x)$ der Verteilungsfunktion *jeder* beliebigen Normalverteilung bestimmen. Gilt nämlich $X \sim N(\mu; \sigma^2)$ und damit $Z \sim N(0; 1)$, 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:

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{12.21}$$

$$P(X > a) = 1 - P(X \leq a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{12.22}$$

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{12.23}$$

Das p-Quantil der Normalverteilung ist nach (11.17) der Wert x_p, an dem die Verteilungsfunktion $F(x)$ den Wert p erreicht. Die **p-Quantile der Standardnormalverteilung** sind also durch

$$\Phi(z_p) = p \tag{12.24}$$

definiert. Der in Abbildung 12.3 auf der z-Achse betonte Punkt $z = 1{,}38$ ist demnach das 0,9162-Quantil der Standardnormalverteilung. Da die Dichte der Standardnormalverteilung symmetrisch zum Nullpunkt ist, gilt dies auch für z_p und z_{1-p}, d. h. es gilt

$$z_p = -z_{1-p}. \tag{12.25}$$

Abbildung 12.4 veranschaulicht diese Symmetrieeigenschaft der Quantile der Standardnormalverteilung für $p = \frac{\alpha}{2}$ resp. $p = 1 - \frac{\alpha}{2}$ mit $\alpha = 0{,}05$. Wegen (12.25) darf in der Grafik $z_{\alpha/2}$ durch $-z_{1-\alpha/2}$ ersetzt werden.

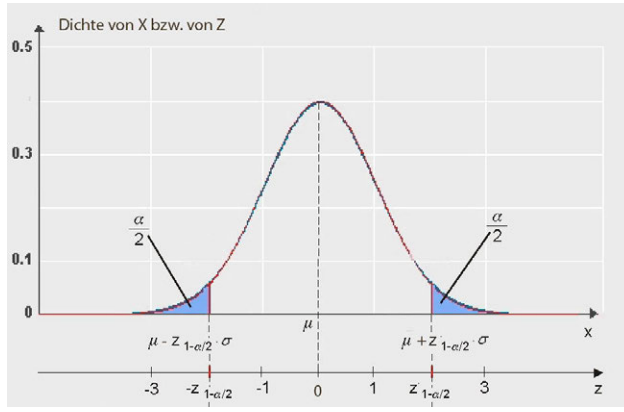


Abb. 12.4: *Symmetrie der Quantile von Normal- und Standardnormalverteilung bezüglich des Erwartungswerts (hier: $\alpha = 0{,}05$)*

Lernobjekt „Quantile der Standardnormalverteilung“

Die Beziehung (12.11) gilt analog auch für die Quantile z_p und x_p von Standardnormalverteilung resp. Normalverteilung. Ist also $X \sim N(\mu; \sigma^2)$, so sind die Quantile x_p dieser Verteilung mit denen der Standardnormal-

verteilung über die Gleichung

$$z_p = \frac{x_p - \mu}{\sigma} \tag{12.26}$$

verbunden, d. h. es ist $x_p = \mu + z_p \cdot \sigma$. Auch dies ist in Abbildung 12.4 veranschaulicht. Man erkennt auch hier, dass die Standardisierung einer normalverteilten Zufallsvariablen X nichts anderes beinhaltet als eine Reskalierung der x-Achse.

Beispiel 12.2: Berechnung von Wahrscheinlichkeiten und Quantilen

Es sei eine normalverteilte, aber nicht-standardnormalverteilte Zufallsvariable X betrachtet, etwa eine $N(\mu;\sigma^2)$-verteilte Zufallsvariable mit $\mu = 1$ und $\sigma^2 = 2{,}25$. Für diese leitet man mit (12.20) - (12.21) und Tabelle 19.2 folgende Aussagen her:

$$P(X \leq -0{,}5) = \Phi\left(\frac{-0{,}5-1}{1{,}5}\right) = \Phi(-1) = 1 - \Phi(1) = 0{,}1587;$$

$$P(X \leq 4) = \Phi\left(\frac{4-1}{1{,}5}\right) = \Phi(2) = 0{,}9772;$$

$$P(-0{,}5 \leq X \leq 4) = \Phi(2) - \Phi(-1) = 0{,}8185.$$

Aufgabe 12.2-3

Der Wert 0,1587 entspricht der Fläche unter der Dichte von X bis zum Punkt $x = -0{,}5$ oder, äquivalent, der Fläche unter der Dichte der Standardnormalverteilung bis zum Punkt -1. Analog ist der Wert 0,9772 zu interpretieren. 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 (12.25). Man erhält $z_{0{,}975} = 1{,}96$ und $z_{0{,}025} = -1{,}96$. Daraus folgt dann mit (12.26) resp. mit $x_p = \mu + \sigma \cdot z_p$ für die gesuchten Quantile

$$x_{0{,}975} = 1 + 1{,}96 \cdot 1{,}5 = 3{,}94, \qquad x_{0{,}025} = 1 - 1{,}96 \cdot 1{,}5 = -1{,}94.$$

Die Wahrscheinlichkeit dafür, dass X im Intervall $[x_{0{,}025}; x_{0{,}975}] = [-1{,}94; 3{,}94]$ liegt, ist 0,95. Dieser Wert ist mit der Wahrscheinlichkeit identisch, dass Z Werte innerhalb des Intervalls $[z_{0{,}025}; z_{0{,}975}] = [-1{,}96; 1{,}96]$ annimmt (vgl. erneut Abbildung 12.4).

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 (12.25) beschränkt sich Tabelle 19.3 auf p-Quantile mit $p > 0{,}5$.

Beispiel 12.3: 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 Punktrohwert 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 man sich anhand von Abbildung 12.4 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 (12.11) 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

$$P(85 \le Y \le 115) = P(-1 \le Z \le 1) = \Phi(1) - \Phi(-1) = 2 \cdot \Phi(1) - 1 \approx 0{,}683,$$

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 12.2: 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. Aber selbst ein realitätsnahes Modell kann nur innerhalb unspektakulärer Börsenperioden hilfreich sein. Bei Eintritt von Sonderereignissen (Anschlag auf das World Trade Center 2001, Konkurs der Investmentbank Lehman Brothers 2008) zeigen sich die Grenzen von Modellen zur Risikoabschätzung von Anlagen.

12.4 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.

Charakterisierung der χ^2-Verteilung

Geht man von n unabhängigen standardnormalverteilten Variablen $Z_1, Z_2, \ldots, Z_n$ aus und bildet die Summe

$$X := Z_1^2 + Z_2^2 + ... + Z_n^2 = \sum_{i=1}^{n} Z_i^2 \tag{12.27}$$

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 (12.14), über die Dichtefunktion definieren. So verfahren etwa Zucchini / Schlegel / Nenadic / Sperlich (2009, Abschnitt 6.4.1). Dieser Weg ist naheliegend – die Aussage, dass die Zufallsvariable X aus (12.27) 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 Dichtefunktion beider Verteilungen relativ sperrig ist.

Die Anzahl n der in (12.27) 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. Es sei ohne Beweis erwähnt, dass sich aus (12.9) und (12.10) sowie der Dichtefunktion der Verteilung für den Erwartungswert und die Varianz einer χ_n^2-verteilten Variablen X die Gleichungen

$$E(X) = n$$
$$V(X) = 2n$$

ableiten lassen. Die Dichtefunktion der χ^2-Verteilung ist in Abbildung 12.5 für zwei ausgewählte Freiheitsgrade n grafisch dargestellt.

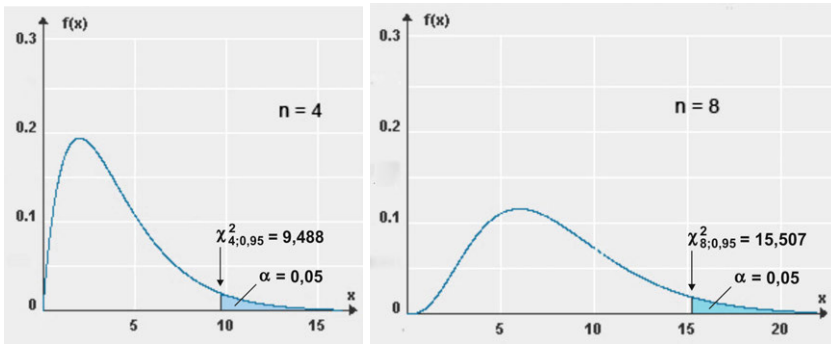


Abb. 12.5: *Dichtefunktion der χ^2-Verteilungen mit $n = 4$ und $n = 8$ Freiheitsgraden (jeweils mit 0,95-Quantilen)*

Die wiedergegebenen Dichtekurven fallen – ähnlich wie die empirischen Verteilungen aus Abbildung 4.7 – 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 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 (11.17) erklärten Quantile, die mit $\chi^2_{n;p}$ abgekürzt werden (lies: *p-Quantil der χ^2-Verteilung mit n Freiheitsgraden*). Bei den Dichten der beiden Verteilungen in Abbildung 12.5 sind auch die 0,95-Quantile $\chi^2_{4;0,95}$ und $\chi^2_{8;0,95}$ angedeutet. Sie bezeichnen jeweils den Punkt auf der Abszissenachse, an dem die farbig betonten Flächen beginnen.

Interaktives Lernobjekt „Quantile der χ^2-Verteilung“

In der Praxis werden die Funktionsdarstellungen für Dichte- und Verteilungsfunktion χ^2-Verteilung i. Allg. nur zur Berechnung von Quantilen der Verteilung 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. In Tabelle 19.4 sind Quantile $\chi^2_{n;p}$ für $n = 1$ bis $n = 40$ und ausgewählte Werte p zusammengestellt. Man entnimmt der Tabelle z. B., dass die 0,95-Quantile der χ^2-Verteilung mit $n = 4$ resp. 8 Freiheitsgraden die Werte $\chi^2_{4;0,95} = 9{,}488$ bzw. $\chi^2_{8;0,95} = 15{,}507$ besitzen.

William S. Gosset

Aus der Standardnormalverteilung und der χ^2-Verteilung leitet sich die **t-Verteilung** ab, die gelegentlich auch **Student-Verteilung** genannt wird. Die t-Verteilung wurde erstmals von William S. Gosset (1876 - 1937) beschrieben. Dabei verwendete er anstelle seines Namens das Pseudonym Student. Hieraus erklärt sich die Bezeichnung „Student-Verteilung“. Sind X und Z unabhängige Zufallsvariablen mit $X \sim \chi^2_n$ und $Z \sim N(0;1)$,[4] dann folgt die Zufallsvariable

$$T := \frac{Z}{\sqrt{\frac{X}{n}}} \tag{12.28}$$

Charakterisierung der t-Verteilung

einer t-Verteilung mit n Freiheitsgraden und man schreibt $T \sim t_n$ (lies: *T ist t-verteilt mit n Freiheitsgraden*). Auch hier wird auf die direkte Einführung der t-Verteilung über ihre Dichtefunktion aufgrund der Komplexität der Dichteformel verzichtet und auf Zucchini / Schlegel / Nenadic / Sperlich (2009, Abschnitt 6.4.3) verwiesen. Die Anzahl der Freiheitsgrade bezeichnet wieder einen Formparameter. Für den Erwartungswert und die Varianz einer t_n-verteilten Variablen T lässt sich

[4] Eine Formalisierung des Begriffs „Unabhängigkeit von Zufallsvariablen“ erfolgt in Abschnitt 13.1.

zeigen, dass

$$E(T) = 0 \qquad (n > 1)$$
$$V(T) = \frac{n}{n-2} \qquad (n > 2).$$

Die Dichte der t-Verteilung ist wie die der Standardnormalverteilung symmetrisch zum Nullpunkt. Für die Quantile der t-Verteilung gilt analog zu (12.25)

$$t_{n;p} = -t_{n;1-p}. \tag{12.29}$$

In Abbildung 12.6 ist die Dichtefunktion der t-Verteilung für zwei ausgewählte Freiheitsgrade n visualisiert. Eingezeichnet sind für $\alpha = 0{,}05$ auch das $\frac{\alpha}{2}$-Quantil und das $(1 - \frac{\alpha}{2})$-Quantil. Beide unterscheiden sich wegen (12.29) nur bezüglich des Vorzeichens. Die Grafik weist auch die Dichte der Standardnormalverteilung mit den entsprechenden Quantilen aus.

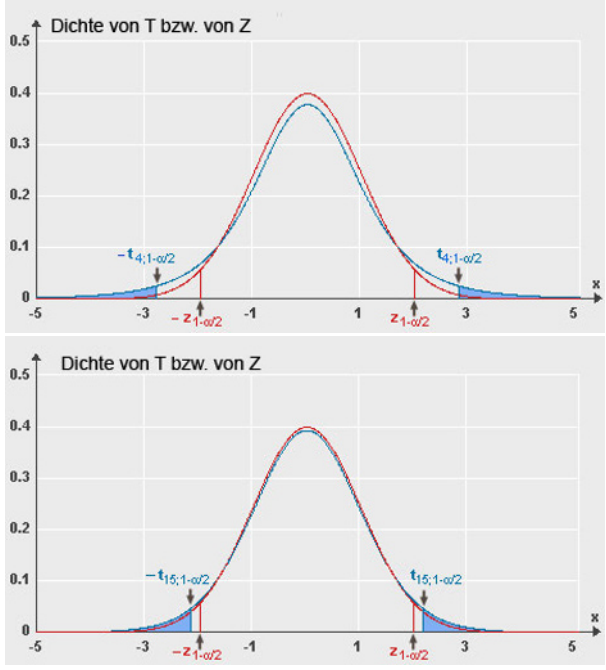


Abb. 12.6: *Dichte zweier t-Verteilungen ($n = 4$ und $n = 15$ Freiheitsgrade) und der Standardnormalverteilung mit Quantilen ($\frac{\alpha}{2} = 0{,}025$)*

Interaktives Lernobjekt „Quantile der t-Verteilung“

Man erkennt, dass die Gestalt der Dichte der t-Verteilung der der Standardnormalverteilung ähnelt. Die Dichtekurve der t-Verteilung verläuft im Bereich des Erwartungswerts $\mu = 0$ flacher und ist an den Flanken etwas breiter. Die p-Quantile der t-Verteilung liegen daher bei kleinem oder großem p etwas weiter vom Nullpunkt entfernt. Mit zunehmender

Anzahl n der Freiheitsgrade nähert sich aber die Dichte der t-Verteilung der der Standardnormalverteilung an. Für große n kann man die mit $t_{n;p}$ (lies: *p-Quantil der t-Verteilung mit n Freiheitsgraden*) abgekürzten und durch (11.17) erklärten Quantile der t-Verteilung durch die Quantile z_p der Standardnormalverteilung approximieren.

Tabelle 12.1 illustriert die Größenordnung der Quantile $t_{n;p}$ und z_p für einige Werte n und p. 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.

Aufgabe 12.4

p	$t_{4;p}$	$t_{15;p}$	$t_{30;p}$	$t_{40;p}$	z_p
0,95	2,132	1,753	1,697	1,684	1,6449
0,975	2,776	2,131	2,042	2,021	1,9600
0,99	3,747	2,602	2,457	2,4233	2,3263

Tab. 12.1: *Quantile der t-Verteilung und der Standardnormalverteilung*

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 $X_1 \sim \chi^2_m$ und $X_2 \sim \chi^2_n$, so folgt die Zufallsvariable

$$Y := \frac{X_1/m}{X_2/n} \tag{12.30}$$

einer F-Verteilung mit m und n Freiheitsgraden und man schreibt $Y \sim F_{m;n}$ (lies: *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).

Charakterisierung der F-Verteilung

In Abbildung 12.7 sind beispielhaft die Dichtekurven zweier $F_{m;n}$-verteilter Zufallsvariablen und für ausgewählte Werte p auch die mit $F_{m;n;p}$ bezeichneten p-Quantile dieser F-Verteilungen visualisiert. Teil a veranschaulicht die Dichtefunktion der F-Verteilung mit $m = 10$ und $n = 15$ Freiheitsgraden sowie das zugehörige 0,95-Quantil $F_{10;15;0,95}$. Teil b der Grafik bezieht sich auf die Dichtefunktion der F-Verteilung mit $m = 5$ und $n = 30$ Freiheitsgraden und das 0,99-Quantil $F_{5;30;0,99}$ dieser Verteilung. Die Quantile bezeichnen jeweils die Positionen auf der Abszissenachse, bei der die markierten Flächen unter den Dichten beginnen. Die Inhalte der markierten Flächen betragen 0,05 im Falle des 0,95-Quantils resp. 0,01 beim 0,99-Quantil.

Interaktives Lernobjekt „Quantile der F-Verteilung“

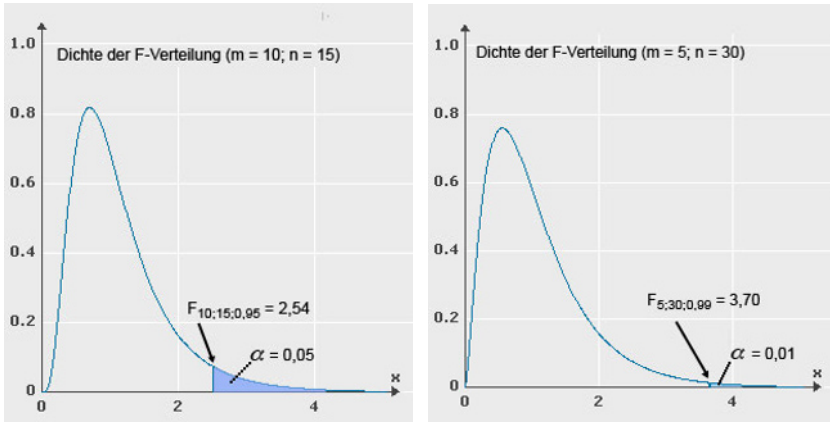


Abb. 12.7: *Dichtefunktion zweier F-Verteilungen mit Quantilen*

Die in Abbildung 12.7 beispielhaft dargestellten Quantile sind 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 (2012, Abschnitt 11.2.3) die Gleichungen

$$E(Y) = \frac{n}{n-2} \qquad (n > 2)$$
$$V(Y) = \frac{2n^2 \cdot (m+n-2)}{m \cdot (n-2)^2 \cdot (n-4)} \qquad (n > 4).$$

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 leitet sich hieraus die Beziehung

$$F_{m;n;p} = \frac{1}{F_{n;m;1-p}} \tag{12.31}$$

ab.[5] Bei der Tabellierung von Quantilen der F-Verteilung kann man sich daher auf Quantile $F_{m;n;p}$ mit $m \leq n$ beschränken.

[5] Folgt 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 $Y \sim \chi^2_n$ resp. $Y \sim t_n$ und $Y \sim F_{m;n}$ verwendet und für die p-Quantile $\chi^2_{n;p}$, $t_{n;p}$ 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 $\chi^2_p(n)$, $t_p(n)$ und $F_p(m;n)$.

Exkurs 12.3: Formparameter von Verteilungen

Für die in den Kapiteln 11 - 12 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 (12.11) 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 gilt $\gamma_1 = 0$, 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 12.7 wiedergegebenen F-Verteilungen und auch die in Abbildung 11.6 dargestellte Binomialverteilung mit $p = 0{,}25$ sind linkssteile Verteilungen. Binomialverteilungen mit $p > 0{,}5$ sind Beispiele für rechtssteile Verteilungen.

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 Wahrscheinlichkeitskeitsmasse 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, ... x_n für ein Merkmal X, so ist die mit $\widehat{\gamma}_1$ bezeichnete *empirische Schiefe* der Verteilung des Datensatzes durch

$$\widehat{\gamma}_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s}\right)^3$$

erklärt. Analog ist die *empirische Wölbung* definiert. Das Symbol $\widehat{\ }$ über γ_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.

13 Bivariate Verteilungen

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 unabhängig 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.

Vorschau auf das Kapitel

Zur Messung des Zusammenhangs zwischen Zufallsvariablen werden die theoretische Kovarianz als nicht-normiertes und der Korrelationskoeffizient ρ als normiertes Maß vorgestellt.

13.1 Unabhängigkeit von Zufallsvariablen

In Abschnitt 10.3 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 (10.16) 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.

Gemeinsame Verteilung zweier Zufallsvariablen

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**

$$F(x; y) := P(X \leq x; Y \leq y) \tag{13.1}$$

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_13

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 (10.16) für alle Elemente der Trägermengen von X und Y als Produkt

$$F(x;y) = F_X(X \leq x) \cdot F_Y(Y \leq y) \tag{13.2}$$

der Verteilungsfunktionen $F_X(x)$ und $F_Y(y)$ der Einzelvariablen darstellen lässt. Die Unabhängigkeitsbedingung gilt entsprechend auch für mehr als zwei Zufallsvariablen.

Beispiel 13.1: Unabhängige und abhängige Zufallsvariablen

Wenn man einen Würfel n-mal wirft, so kann man jeden Wurf durch eine Zufallsvariable X_i modellieren ($i = 1, 2, ..., n$), 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 z. B. 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 aus $F_X(2) = P(X \leq 2) = \frac{1}{3}$ und $F_Y(3) = P(Y \leq 3) = \frac{1}{2}$ gegeben, also 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 man die einzelnen Ziehungen wieder anhand von Zufallsvariablen X_i, so sind diese Zufallsvariablen nicht mehr stochastisch unabhängig. Die Ziehung der Lottozahlen wurde schon in Abschnitt 11.4 als Beispiel für ein solches Experiment genannt.

Exkurs 13.1: 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 $x_1, \ldots, x_k$ resp. $y_1, \ldots, y_l$ und bezeichnet $p_{ij} := P(X = x_i; Y = y_j)$ die Eintrittswahr-

scheinlichkeit für die Realisation $(x_i; y_j)$, so lautet das bivariate Analogon zur Wahrscheinlichkeitsfunktion (11.1)

$$f(x;y) = \begin{cases} p_{ij} & \text{für } (x;y) = (x_i;y_j); \quad i = 1,2,\ldots,k; j = 1,2,\ldots,l; \\ 0 & \text{für alle sonstigen } (x;y). \end{cases}$$

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 (12.2) 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:

$$F(x;y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s;t)dsdt \qquad \text{für alle reellwertigen Paare } (x;y).$$

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 Einzelvariablen und bedingte Dichten resultieren – analog zu (8.7) oder (8.8) 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 / Künstler / Pigeot / Tutz (2010, Kapitel 8) oder Toutenburg / Heumann (2008, Abschnitt 3.7).

Wichtige Stichprobenfunktionen

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 $X_1, X_2, ..., X_n$ modelliert, für die man Realisationen $x_1, x_2, ..., x_n$ beobachtet und verwertet. Die Zufallsvariablen $X_1, X_2, ..., X_n$ werden meist nicht direkt herangezogen, sondern anhand einer **Stichprobenfunktion** aggregiert:

$$X_1, X_2, \ldots, X_n \xrightarrow{\text{Verdichtung der Stichprobeninformation}} g(X_1, X_2, \ldots, X_n)$$

Wenn eine Stichprobenfunktion im Kontext der Schätzung verwendet wird, spricht man sie auch als **Schätzfunktion** an, beim Testen als **Test-** oder **Prüfstatistik**.

Eine wichtige Stichprobenfunktion ist der **Stichprobenmittelwert**

$$\overline{X} := \frac{1}{n} \cdot (X_1 + X_2 + \ldots + X_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i, \tag{13.3}$$

der auf der Datenebene seine Entsprechung in (5.2) findet. Eine weitere Stichprobenfunktion, die beim Schätzen und Testen oft gebraucht wird, ist die **Stichprobenvarianz**

$$S^2 := \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{13.4}$$

bzw. die **korrigierte Stichprobenvarianz**

$$S^{*2} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} \cdot S^2, \tag{13.5}$$

die in (5.6) und (5.9) ihre empirische Entsprechung 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 (12.27), (12.28) resp. (12.30) vorgestellt.

Verteilung des Stichprobenmittelwerts

Wenn die Stichprobenvariablen $X_1, X_2, ..., X_n$ alle unabhängig $N(\mu; \sigma^2)$-verteilt sind, so kann man auch für die Stichprobenfunktionen (13.3) und (13.5) Verteilungsaussagen ableiten, die u. a. beim Testen von Hypothesen eine wichtige Rolle spielen. Überträgt man (12.17) 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 $\overline{X}$ verifiziert man dann mit (12.16), wenn man dort speziell $a = \frac{1}{n}$ und $b = 0$ einsetzt, dass er normalverteilt ist mit Erwartungswert $E(\overline{X}) = \mu$ und Varianz $V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, also [2]

$$\overline{X} \sim N(\mu; \sigma^2_{\overline{X}}) \quad \text{mit} \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}. \tag{13.6}$$

Standardisiert man den Stichprobenmittelwert gemäß (12.11), folgt

$$\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} \sim N(0; 1). \tag{13.7}$$

[1] Beim Schätzen und Testen wird vor allem die korrigierte Stichprobenvarianz (13.5) gebraucht, die günstigere Schätzeigenschaften hat. Die Bezeichnungen sind in der Literatur nicht einheitlich; in vielen Lehrbüchern wird (13.5) Stichprobenvarianz genannt und (13.4) gar nicht verwendet.

[2] Die Formeln für den Erwartungswert und die Varianz von $\overline{X}$ sind nicht an die Normalverteilungsannahme gebunden, wie in Abschnitt 14.2 noch gezeigt wird.

Für die aus n unabhängigen $N(\mu;\sigma^2)$-verteilten Stichprobenvariablen X_i gebildete Stichprobenvarianz lässt sich eine Beziehung zur χ^2-Verteilung ableiten. Auch die Variablen X_i kann man zunächst gemäß (12.11) standardisieren. Für die Summe der Quadrate der resultierenden standardnormalverteilten Variablen Z_i gilt (vgl. (12.27)), dass sie χ^2-verteilt ist mit n Freiheitsgraden:

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2. \tag{13.8}$$

Verteilung der Stichprobenvarianz

Hieraus kann man mit einigen Überlegungen ableiten, dass die mit dem Faktor $\frac{n}{\sigma^2}$ multiplizierte Stichprobenvarianz S^2 bzw. – äquivalent – die mit $\frac{n-1}{\sigma^2}$ multiplizierte korrigierte Stichprobenvarianz S^{*2} einer χ^2-Verteilung mit $n-1$ Freiheitsgraden folgt:

$$\frac{n \cdot S^2}{\sigma^2} = \frac{(n-1) \cdot S^{*2}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2. \tag{13.9}$$

Ferner lässt sich mit (13.9) zeigen, dass eine Ersetzung von σ in (13.7) durch die als Schätzung für σ verwendete **korrigierte Stichprobenstandardabweichung** $S^* := \sqrt{S^{*2}}$ zu einer t-Verteilung mit $n-1$ Freiheitsgraden führt:

$$\frac{\overline{X} - \mu}{S} \cdot \sqrt{n-1} = \frac{\overline{X} - \mu}{S^*} \cdot \sqrt{n} \sim t_{n-1}. \tag{13.10}$$

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 (13.10) z. B. bei Mosler / Schmid (2011, Abschnitt 4.3.2).

Exkurs 13.2: Der Zentrale Grenzwertsatz

Aussage (13.6) bezieht sich auf die Verteilung eines Stichprobenmittelwerts $\overline{X}$, der aus n unabhängigen, mit gleichem Erwartungswert μ und gleicher Varianz σ^2 normalverteilten Zufallsvariablen $X_1, X_2, ..., X_n$ gebildet ist, während (13.7) eine Verteilungsaussage für den aus $\overline{X}$ 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 $X_1, X_2, ..., X_n$ zwar unabhängig sind und bezüglich Erwartungswert und Varianz übereinstimmen, aber nicht mehr normalverteilt sind.

Seien also $X_1, X_2, ..., X_n$ 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 $\overline{X}$ hier noch einen Index anbringt, also die Bezeichnung $\overline{X}_n$ verwendet, so sagt der Zentrale Grenzwertsatz, dass die Verteilungsfunktion der standardisierten Summe

$$Z_n := \frac{Y_n - E(Y_n)}{\sqrt{V(Y_n)}} = \sum_{i=1}^{n} \frac{X_i - n \cdot \mu}{\sqrt{n \cdot \sigma^2}} = \frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}}$$

unter diesen Voraussetzungen für $n \to \infty$ gegen die Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung konvergiert. Hieraus lässt sich folgern, dass bei großen Werten n für $\overline{X}_n$ näherungsweise (13.6) gilt und dass die Summe Y_n der Variablen $X_1, X_2, ..., X_n$ approximativ $N(n \cdot \mu; n \cdot \sigma^2)$-verteilt ist.

Java-Applet „Approximation der Binomialverteilung“

Wählt man speziell die n Variablen $X_1, X_2, ..., X_n$ wie in (11.19), also als identisch bernoulli-verteilt, so folgt aus den vorausgegangenen Ausführungen und bei Beachtung von (11.15) und (11.16), dass die binomialverteilte Zählvariable $Y_n := X$ aus (11.19) 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 also 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.

13.2 Kovarianz und Korrelation

In den Abschnitten 11.2 und 12.2 wurden univariate Wahrscheinlichkeitsverteilungen von Zufallsvariablen anhand von Kenngrößen charakterisiert. Als Lageparameter für die Verteilung von X wurde hier der durch (11.6) resp. (12.9) definierte Erwartungswert $\mu = E(X)$ aufgeführt und als Streuungsparameter die Varianz $V(X) = \sigma^2 = E\left[(X-\mu)^2\right]$ aus (11.8) oder die Standardabweichung $\sigma = \sqrt{V(X)}$ aus (11.10).

Ein nicht-normiertes Zusammenhangsmaß

Hat man *zwei* Zufallsvariablen X und Y mit den Erwartungswerten $\mu_X = E(X)$ und $\mu_Y = E(Y)$ und 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 nichtnormiertes Maß für einen linearen Zusammenhang ist die mit $Cov(X;Y)$ abgekürzte **Kovarianz** von X und Y, die zwecks Unterscheidung von der empirischen Kovarianz (9.9) auch **theoretische Kovarianz** genannt

wird. Sie ist definiert als Erwartungswert von $(X - \mu_X)(Y - \mu_Y)$:

$$Cov(X;Y) := E\left[(X - E(X))(Y - E(Y))\right]. \tag{13.11}$$

Durch Ausmultiplizieren der beiden Differenzterme $X - \mu_X$ und $Y - \mu_Y$ und anschließende gliedweise Anwendung des Erwartungswertoperators gewinnt man aus (13.11) noch die äquivalente Darstellung

$$Cov(X;Y) = E(XY) - E(X) \cdot E(Y). \tag{13.12}$$

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 $Cov(X;Y) = 0$ 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

$$X \text{ und } Y \text{ sind unabhängig } \rightarrow Cov(X;Y) = 0. \tag{13.13}$$

Sind X und Y zwei Zufallsvariablen mit der Kovarianz $Cov(X;Y)$, so gilt für die Varianz ihrer Summe

$$V(X + Y) = V(X) + V(Y) + 2 \cdot Cov(X;Y). \tag{13.14}$$

Ein normiertes Zusammenhangsmaß

Wie die empirische Kovarianz ist auch die theoretische Kovarianz maßstabsabhängig. Sie hat daher keine untere oder obere Schranke. Eine zur Definition (9.10) 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]

$$\rho = \frac{Cov(X;Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}. \tag{13.15}$$

Der **Korrelationskoeffizient** ρ liegt wie sein empirisches Analogon r stets zwischen -1 und $+1$, d. h. es gilt

$$-1 \leq \rho \leq 1. \tag{13.16}$$

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

[3] Man 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 (13.13) folgt, dass Unabhängigkeit von X und Y stets Unkorreliertheit impliziert:

$$X \text{ und } Y \text{ sind unabhängig } \rightarrow \rho = 0. \tag{13.17}$$

Der Umkehrschluss gilt nicht, d.h. unkorrelierte Zufallsvariablen sind nicht zwingend auch stochastisch unabhängig.

Beispiel 13.2: 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 (13.15) 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, nicht – wie bei der Berechnung des empirischen Korrelationskoeffizienten r nach Bravais-Pearson – aus Daten.

Aufgabe 13.1

14 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).

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, wie in Abschnitt 13.1 erläutert, zu einer Stichprobenfunktion aggregiert. Deren Ausprägung wird dann für die Schätzung des unbekannten Parameters herangezogen. Da die Informationsbasis bei Verwendung von Stichproben schmaler 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

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_14

die aus ihnen abgeleiteten Schlüsse tendenziell zuverlässiger sind als bei kleinen Stichproben.

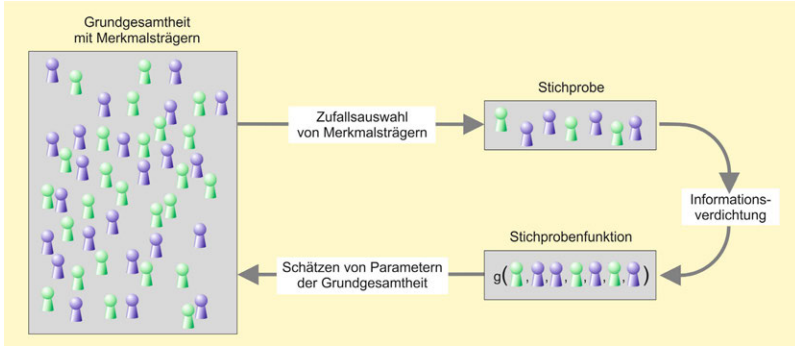


Abb. 14.1: *Vorgehensweise bei der Schätzung*

Punkt- und Intervallschätzung

Wenn man für ein stochastisches Merkmal X ein geeignetes Verteilungsmodell spezifiziert hat, also eine bestimmte diskrete oder stetige Verteilung, kommen für die Schätzung des interessierenden unbekannten 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 unbekannte Parameter mit einer Wahrscheinlichkeit von mindestens $1-\alpha$ liegt, wobei α eine vorgegebene kleine Irrtumswahrscheinlichkeit ist.

Beispiel 14.1: 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 (Gewinnung empirisch fundierter Prognosen).
- In der *Marktforschung* verwertet man u. a. Stichprobendaten zum Fernsehverhalten von Zuschauern. Aus diesen will man z. B. den mittleren Fernsehkonsum oder 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.

14.1 Punktschätzungen und ihre Eigenschaften

Flash-Animation „Punktschätzung“

Will man für einen Parameter θ (lies: *theta*) der Verteilung eines Merkmals in einer Grundgesamtheit eine Punktschätzung anhand von Stichprobendaten $x_1, x_2, ..., x_n$ gewinnen, verwendet man die Realisation einer **Stichprobenfunktion** $g(x_1, x_2, ..., x_n)$ als Schätzwert. Da die Stichprobendaten als Ausprägungen von Zufallsvariablen $X_1, X_2, ..., X_n$ interpretiert werden, ist auch der aus ihnen errechnete Schätzwert eine Realisation einer Zufallsvariablen $g(X_1, X_2, ..., X_n)$, 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 $\widehat{\theta}$ angesprochen (lies: *theta-Dach*). Die Verwendung von $\widehat{\ }$ über einer Kenngröße ist in der Statistik für die Kennzeichnung von Schätzungen üblich.

Gütekriterien für Schätzfunktionen:

Bevor Schätzer $\widehat{\theta}$ 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]

$$E(\widehat{\theta}) = \theta. \tag{14.1}$$

Wenn ein Schätzer $\widehat{\theta}$ nicht erwartungstreu ist, heißt die Differenz

$$B(\widehat{\theta}) := E(\widehat{\theta}) - \theta = E(\widehat{\theta} - \theta) \tag{14.2}$$

- keine oder geringe Verzerrung

Verzerrung oder **Bias** (engl.: *bias*). Ein Schätzer für θ ist also genau dann erwartungstreu, wenn seine Verzerrung Null ist. Manchmal ist ein Schätzer $\widehat{\theta}$ zwar verzerrt, besitzt aber eine Verzerrung, die gegen Null strebt, wenn der Umfang n des zur Berechnung von $\widehat{\theta}$ verwendeten Datensatzes gegen ∞ (lies: *unendlich*) strebt:

$$\lim_{n \to \infty} E(\widehat{\theta}) = \theta. \tag{14.3}$$

Ein Schätzer $\widehat{\theta}$ mit dieser Eigenschaft heißt **asymptotisch erwartungstreu** oder **asymptotisch unverzerrt**.

- kleine Varianz

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

[1] Der Erwartungswert $E(\widehat{\theta})$ wird unter der Annahme bestimmt, dass der zu schätzende unbekannte Parameter den Wert θ hat. Gelegentlich wird dies durch die Notation $E_\theta(\widehat{\theta})$ betont (analoge Indizierung für andere Schätzercharakteristika). Erwartungstreue beinhaltet, dass (14.1) für alle möglichen Werte von θ gilt.

Statistiksoftwarepaketen bei der Anwendung von Schätzprozeduren routinemäßig neben den Schätzwerten ausgewiesen (vgl. Abbildung 16.3).

Flash-Animation „Beurteilung von Schätzern“

Abbildung 14.2 zeigt die – hier als symmetrisch angenommenen – Dichtefunktionen für drei Schätzfunktionen, wobei die ersten beiden Schätzer, etwa $\widehat{\theta}_1$ und $\widehat{\theta}_2$, den Erwartungswert $E(\widehat{\theta}_1) = E(\widehat{\theta}_2) = a$ und der dritte Schätzer $\widehat{\theta}_3$ den Erwartungswert $E(\widehat{\theta}_3) = b$ habe. Geht man davon aus, dass der zu schätzende unbekannte Parameter θ den Wert $\theta = b$ hat, so ist $\widehat{\theta}_3$ erwartungstreu, während $\widehat{\theta}_1$ und $\widehat{\theta}_2$ verzerrte Schätzer sind. Allerdings hat $\widehat{\theta}_2$ eine kleinere Varianz als $\widehat{\theta}_3$. Da sich die Schätzer $\widehat{\theta}_1$ und $\widehat{\theta}_2$ nicht bezüglich des Erwartungswerts unterscheiden, ist von beiden der Schätzer mit der geringeren Streuung (steilere Dichtekurve) vorzuziehen, also $\widehat{\theta}_2$.

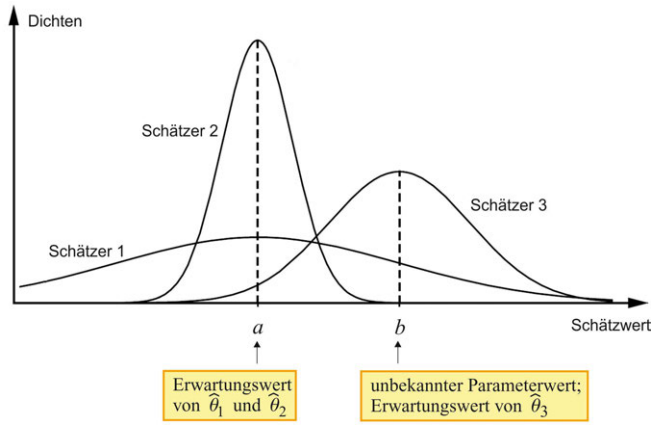


Abb. 14.2: *Vergleich dreier Schätzfunktionen für* $\theta = b$

- kleiner MSE

In der Praxis hat man häufig verzerrte Schätzer. Wie aber soll man sich zwischen zwei Schätzern entscheiden, wenn – wie in Abbildung 14.2 anhand der Schätzer $\widehat{\theta}_3$ und $\widehat{\theta}_2$ 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.: *m*ean *s*quared *e*rror)

$$MSE(\widehat{\theta}) := E\left[\left(\widehat{\theta} - \theta\right)^2\right]. \tag{14.4}$$

MSE-Zerlegungsformel

Nach elementaren Umformungen erhält man mit (14.2) und der Varianzdefinition (11.8) die äquivalente Darstellung

$$MSE(\widehat{\theta}) = E\left[\left(\widehat{\theta} - E(\widehat{\theta})\right)^2\right] + \left[E(\widehat{\theta}) - \theta\right]^2 = V(\widehat{\theta}) + B(\widehat{\theta})^2. \tag{14.5}$$

Der MSE repräsentiert eine additive Verknüpfung von Varianz und quadrierter Verzerrung. Man wird von den beiden Schätzern $\widehat{\theta}_2$ und $\widehat{\theta}_3$ in Abbildung 14.2 denjenigen als „besser“ ansehen, dessen MSE kleiner ausfällt. Bei erwartungstreuen Schätzern sind MSE und Varianz identisch.

14.2 Schätzung von Erwartungswerten, Varianzen und Anteilen

Schätzung des Erwartungswerts

Will man den Erwartungswert μ einer Zufallsvariablen anhand der Ausprägungen unabhängiger Stichprobenvariablen $X_1, X_2, ..., X_n$ schätzen, bietet sich als Stichprobenfunktion der in (13.3) eingeführte **Stichprobenmittelwert** $\overline{X}$ an. Da man die Erwartungswertbildung nach (11.13) auf die Stichprobenvariablen einzeln anwenden kann, gilt

$$E(\overline{X}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \ldots + E(X_n)] = \frac{1}{n} \cdot n \cdot \mu = \mu. \quad (14.6)$$

Der Stichprobenmittelwert liefert also eine *unverzerrte* Schätzung für den Erwartungswert. Wenn die Stichprobenvariablen $X_1, X_2, ..., X_n$ unabhängig sind und die feste Varianz σ^2 haben, kann man für die Varianz $V(\overline{X}) = \sigma^2_{\overline{X}}$ von $\overline{X}$ mit (11.12) und (11.14) die Darstellung

$$V(\overline{X}) = \frac{\sigma^2}{n} \quad (14.7)$$

ableiten. Wegen der Unverzerrtheit von $\overline{X}$ stimmt $V(\overline{X})$ mit dem mittleren quadratischen Fehler $MSE(\overline{X})$ von $\overline{X}$ überein. Die Qualität des Schätzers $\overline{X}$ verbessert sich also bei Erhöhung des Stichprobenumfangs.

Schätzung der Varianz

Zur Schätzung der Varianz σ^2 einer Zufallsvariablen kommt zunächst die **Stichprobenvarianz** S^2 aus (13.4) in Betracht. Mit elementaren Umformungen und Anwendung der Zerlegungsformel (11.9) auf die Varianz von $\overline{X}$ kann man zeigen, dass[2]

$$E(S^2) = \frac{n-1}{n} \cdot \sigma^2. \quad (14.8)$$

Die Stichprobenvarianz 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 allerdings 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 14.1

[2] Die Zerlegung der Varianz von $\overline{X}$ hat die Gestalt $\sigma^2_{\overline{X}} = E(\overline{X}^2) - \mu^2$. Eine Herleitung von (14.8) findet man bei Mosler / Schmid (2011, Abschnitt 5.1.4).

unverzerrt ist, verwendet man anstelle von S^2 zur Varianzschätzung die **korrigierte Stichprobenvarianz** S^{*2} aus (13.5). Für sie gilt

$$E(S^{*2}) = \frac{n}{n-1} \cdot E(S^2) = \sigma^2. \tag{14.9}$$

Schätzung 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 $\overline{A}$, die mit Wahrscheinlichkeit $p = P(A)$ resp. $1 - p = P(\overline{A})$ auftreten. In Abschnitt 11.1 wurde als Beispiel ein Münzwurfexperiment genannt (Ausgänge „Zahl“ und „Kopf“). Wenn man ein Bernoulli-Experiment n-mal durchführt, kann man den Ausgang jedes Einzelexperiments anhand der Indikatorvariablen (11.3) modellieren, die gesamte Bernoulli-Kette also durch eine Folge unabhängiger Stichprobenvariablen $X_1, X_2, ..., X_n$. Der hieraus gebildete Stichprobenmittelwert $\overline{X}$ lässt sich zur Schätzung des Erwartungswerts p heranziehen. Der Erwartungswert p repräsentiert hier den zu erwartenden *Anteil* der Ausgänge mit A. Für die Schätzfunktion $\widehat{p} := \overline{X}$ gilt analog zu (14.6)

$$E(\widehat{p}) = \frac{1}{n} \cdot [E(X_1) + E(X_2) + \ldots + E(X_n)] = \frac{1}{n} \cdot n \cdot p = p. \tag{14.10}$$

Da die bernoulli-verteilten Variablen X_i nach (11.16) die Varianz $\sigma^2 = p(1-p)$ haben, erhält man für die Varianz $V(\widehat{p})$ von $\widehat{p}$ mit (14.7)

$$V(\widehat{p}) = V(\overline{X}) = \frac{p \cdot (1-p)}{n}. \tag{14.11}$$

Beispiel 14.2: Schätzung bei mangelhafter Datenqualität

Bei der Schätzung des „durchschnittlichen“ Einkommens in einer größeren Population von Personen wird man anstelle des Stichprobenmittelwerts $\overline{X}$ den Stichprobenmedian $\widetilde{X}$ heranziehen, der robuster gegenüber Ausreißern reagiert. Noch entscheidender aber als die Wahl der Stichprobenfunktion ist die Qualität der Daten, auf die eine Stichprobenfunktion angewendet wird. Selbst die beste Schätzmethode kann zu unbrauchbaren Ergebnissen führen, wenn die Daten von zweifelhafter Qualität sind.

Diese Aussage lässt sich anhand eines Fallbeispiels illustrieren, das unter dem Schlagwort „Maserati-Affäre“ Anfang 2010 durch die Presse ging (Bericht im *Spiegel* vom 25. 2. 2010). Der Geschäftsführer einer als gemeinnützig geltenden und in der Obdachlosenhilfe tätigen Berliner Organisation war aufgrund der Verwendung eines für den Tätigkeitsbereich „Wohlfahrtspflege“ ungewöhnlichen Dienstwagens, einem Maserati, in die Schlagzeilen geraten. Dabei kam auch ans Licht, dass der Geschäftsführer ein Jahresbruttogehalt von deutlich über 400.000 Euro bezog. Der Landesverband Berlin des Paritätischen Wohlfahrtsverbands, der den besagten Träger als Mitglied führte, kam rasch unter öffentlichen Druck. Um dem Imageschaden entgegenzuwirken, beauftragte der Landesverband eine

Wirtschaftsprüfungsgesellschaft damit, bei den insgesamt 650 Mitgliedsorganisationen anhand eines Fragebogens die Geschäftsführergehälter zu ermitteln und hieraus Durchschnittsgehälter zu schätzen.

In einer Ende Juli 2010 veröffentlichten Pressenotiz zu den Ergebnissen der Erhebung sah der Auftraggeber mit der Studie den Beweis als erbracht, dass die Affäre nur einen Einzelfall betraf und die Gehälter von Führungskräften bei den beteiligten Mitgliederorganisationen i. Allg. durchaus angemessen sind. Aus den eingegangenen Fragebögen hatte sich für das Merkmal „Gesamt-Bruttojahresgehalt“ (Festgehalt einschließlich sonstiger Einkommenskomponenten) der nicht-ehrenamtlich tätigen Geschäftsführer ein Mittelwert von ca. 56.100 Euro und ein Median von ca. 53.300 Euro ergeben. Die Werte ergaben sich aus 246 Fragebögen, die von 650 angeschriebenen Trägern zurück kamen (Rücklaufquote von nur 38 %). Dass der Median unterhalb des Mittelwerts liegt, ist plausibel und mit den Befunden aus Abbildung 4.7 kompatibel.

Wenn man den Fragebogen am Ende des Originalberichts ansieht, stellt man fest, dass nach dem Arbeit*geber*bruttogehalt gefragt wurde, nicht nach dem Arbeit*nehmer*brutto, und zwar – unverständlicherweise – nach dem *ausgezahlten* Arbeitgeberbrutto. Die Variable, um die es in der Erhebung maßgeblich geht, ist also aus dem Fragebogen heraus nicht zu verstehen. Allein die mangelhafte Operationalisierung der Variablen „Geschäftsführergehalt“ spricht gegen eine ausreichende Reliabilität und Validität der Ergebnisse.

Bedenklich ist zudem, dass alle im Fragebogen abgefragten statistischen Informationen anonym, freiwillig und damit ohne jede Kontrolle oder logische Konsistenzprüfung zustande kamen. Ob die Non-Response-Rate bei den umsatzstärkeren Organisationen mit höheren Gehältern für Führungskräfte höher war als bei kleineren Trägern, kann aus der Studie nicht erschlossen werden. Der Fragebogen differenzierte auch nicht zwischen kostensatzfinanzierten Organisationen und solchen, die im wesentlichen mit Zuwendungen aus öffentlichen Mitteln operieren. Nur bei den Zuwendungsempfängern unterliegen die Gehälter stärkeren Beschränkungen (Deckelung durch das sog. „Besserstellungsverbot“ gegenüber öffentlich Bediensteten mit vergleichbarer Tätigkeit). Es hätte sich angeboten für verschiedene Klassen von Trägern – analog zu Abbildung 5.4 – Boxplots zu generieren, um auch die Streuung innerhalb unterschiedlicher Trägerklassen sichtbar zu machen.

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 Angemessenheit der zur statistischen Analyse eingesetzten Instrumente liefern. Die von der Wirtschaftsprüfungsgesellschaft ermittelten Schätzergebnisse für Geschäftsführerbezüge haben zwar eventuell zur Beruhigung der Öffentlichkeit beigetragen, die Art der Durchführung und Auswertung der Erhebung begründen aber erhebliche Zweifel an der Datenqualität und damit auch an den Schätzwerten. Die Ergebnisse der Auftragsstudie dürften daher kaum als Basis für Politikentscheidungen getaugt haben.

14.3 Konfidenzintervalle für Erwartungswerte

Eine **Punktschätzung** $\widehat{\theta}$ für einen Parameter θ liefert einen einzigen Schätzwert, der meist mit θ nicht exakt übereinstimmt. Zur Beurteilung der Güte einer Punktschätzung spielt die Verzerrung (14.2) 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 (14.4) 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.

Illustration für normalverteiltes Merkmal:

- Varianz bekannt

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 $x_1, x_2, \ldots, x_n$ werden als Ausprägungen unabhängiger $N(\mu;\sigma^2)$-verteilter Zufallsvariablen $X_1, X_2, \ldots, X_n$ interpretiert. Die Zufallsvariable $Z := \frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ ist dann gemäß (13.7) 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 $\left[-z_{1-\alpha/2}; z_{1-\alpha/2}\right]$, das in Abbildung 12.4 veranschaulicht ist. Es gilt also für den standardisierten Stichprobenmittelwert Z die Wahrscheinlichkeitsaussage

$$P\left(-z_{1-\alpha/2} \le \frac{\overline{X}-\mu}{\sigma} \cdot \sqrt{n} \le z_{1-\alpha/2}\right) = 1-\alpha. \tag{14.12}$$

Wenn man die drei Terme der Ungleichungskette in der Klammer mit $\frac{\sigma}{\sqrt{n}}$ erweitert, dann jeweils $\overline{X}$ subtrahiert und mit -1 multipliziert, folgt

$$P\left(\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \tag{14.13}$$

Das mit KI bezeichnete Intervall

$$KI = \left[\overline{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{14.14}$$

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 (14.14) setzt voraus, dass die Varianz σ^2 bzw.

[3]Bei 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.

die Standardabweichung σ der $N(\mu;\sigma^2)$-verteilten Variablen X bekannt ist, also nicht erst über eine Schätzung zu ermitteln ist.

In (14.13) bzw. (14.14) geht die Ausprägung $\widehat{\mu} = \overline{x}$ des Schätzers $\overline{X}$ ein, die von Stichprobe zu Stichprobe variiert. Die Intervallgrenzen sind also *zufallsabhängig*. Die Länge des Konfidenzintervalls ist fest und durch

$$\text{Länge}(KI) = 2 \cdot z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{14.15}$$

gegeben, hängt also von der Irrtumswahrscheinlichkeit α und vom Stichprobenumfang n ab. Mit abnehmender Irrtumswahrscheinlichkeit α (wachsendem Konfidenzniveau $1-\alpha$) nimmt die Länge (14.15) zu, weil das Quantil $z_{1-\alpha/2}$ dann größere Werte annimmt (vgl. erneut Abbildung 12.4). Mit zunehmendem n wird das Konfidenzintervall schmaler.

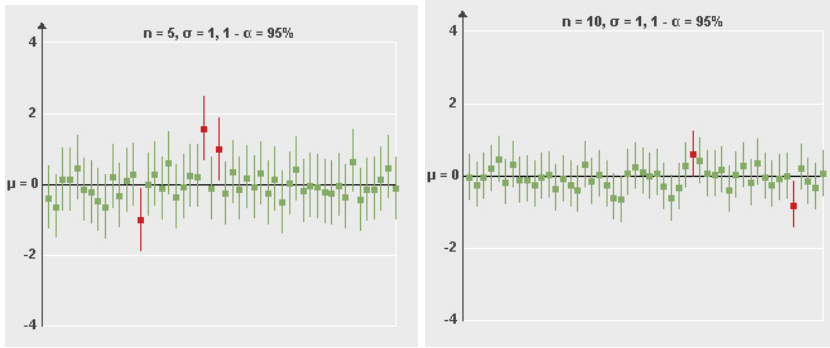


Abb. 14.3: *Konfidenzintervalle für μ (Normalverteilung; Varianz bekannt; $n = 5$ und $n = 10$)*

Java-Applet „Konfidenzintervalle für μ (Varianz bekannt)“

Abbildung (14.3) zeigt Konfidenzintervalle zum Konfidenzniveau 0,95 ($\alpha = 0{,}05$), die nach (14.14) berechnet wurden und jeweils auf n per Simulation generierten Stichprobendaten basieren. Für die Stichprobenvarianz wurde $\sigma^2 = 1$ gewählt und für den zu schätzenden Erwartungswert $\mu = 0$, d. h. es wurden n standardnormalverteilte Daten generiert. Der Vorgang wurde k-mal ausgeführt mit $k = 50$. Insgesamt wurden somit jeweils $k = 50$ Konfidenzintervalle per Simulation erzeugt. Eine Simulation bietet den Vorteil, dass der üblicherweise unbekannte Parameter μ, für den man Intervallschätzungen berechnen will, ausnahmsweise bekannt ist (kontrollierte Laborsituation).

Für die Länge der Konfidenzintervalle erhält man mit (14.15) und Tabelle 19.3 im Falle $n = 5$ den Wert $\frac{2 \cdot 1{,}96}{\sqrt{5}} \approx 1{,}75$. Im Falle $n = 10$ sind die Konfidenzintervalle erwartungsgemäß schmaler (breitere Datenbasis). Ihre Länge errechnet sich zu $\frac{2 \cdot 1{,}96}{\sqrt{10}} \approx 1{,}24$. Die Verdopplung von n führt nach (14.15) 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.

Aufgabe 14.2

In Teil a von Abbildung 14.3 überdecken drei (6%), in Teil b zwei (4%) der $k = 50$ Intervalle den Parameter $\mu = 0$ nicht. Der theoretische Wert, den man approximativ bei Wahl eines sehr großen Wertes für k erreicht, ist $\alpha = 0{,}05$ (5%). Man müsste k deutlich erhöhen, um den theoretischen Wert $\alpha = 0{,}05$ besser zu treffen. Es ist jedenfalls festzuhalten, dass ein konkretes Konfidenzintervall den unbekannten Parameter – auch bei klein gewählter Irrtumswahrscheinlichkeit α – nicht zwingend überdeckt.

- Varianz unbekannt

Die vorstehenden Ableitungen sind leicht zu modifizieren, wenn man die Varianz σ^2 nur in Form einer Schätzung $\widehat{\sigma}^2$ kennt. Ausgangspunkt ist hier nicht mehr der standardisierte Stichprobenmittelwert aus (13.7), sondern die mit $n-1$ Freiheitsgraden t-verteilte Zufallsvariable aus (13.10). Man erhält anstelle von (14.14)

$$KI = \left[\overline{X} - t_{n-1;1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}}; \overline{X} + t_{n-1;1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}}\right]. \tag{14.16}$$

Java-Applet „Konfidenzintervalle für μ (Varianz unbekannt)“

Diese Formel unterscheidet sich von (14.14) darin, dass statt der festen Größe σ nun die Zufallsvariable S^* erscheint und statt zweier Quantile der Standardnormalverteilung die entsprechenden Quantile der t-Verteilung mit $\nu = n-1$ Freiheitsgraden. Die erste Ersetzung hat wegen (14.9) 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 12.1). Das Konfidenzintervall (14.16) ist also im Mittel länger. Ein wesentlicher Unterschied gegenüber (14.14) besteht auch darin, dass die Länge

$$\text{Länge}(KI) = 2 \cdot t_{n-1;1-\alpha/2} \cdot \frac{S^*}{\sqrt{n}} \tag{14.17}$$

des Konfidenzintervalls (14.16) nun nicht mehr nur von der Irrtumswahrscheinlichkeit α und dem Stichprobenumfang n, sondern auch von der jeweiligen Ausprägung von S^* abhängt, also zufallsabhängig ist.

Exkurs 14.1: Stichprobenfehler bei 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 Annahme grundsätzlich nur näherungsweise, d. h. man begeht bei einer Schätzung anhand einer Stichprobe stets einen Fehler, den sog. *Stichprobenfehler*. Wie gravierend der Stichprobenfehler ist hängt natürlich vom Umfang n der Stichprobe ab. Ist n im Verhältnis zum Umfang N der Grundgesamtheit sehr klein, kann der Stichprobenprobenfehler erheblich und die Qualität der Schätzung fragwürdig sein.

Das Auftreten des Stichprobenfehlers sei an einer Schätzung von Armutsgefährdungsquoten in den Kreisen des Bundeslandes Rheinland-Pfalz illustriert. Die Schätzung basiert auf Daten des Jahres 2010 aus der europaweiten Erhebung *EU-SILC* (*E*uropean *U*nion *S*tatistics on *I*ncome and *L*iving *C*onditions; s. auch Exkurs 5.3). Diese Erhebung umfasst für ganz Deutschland nur knapp 15000 Haushalte – zu wenig, um für das Bundesland Rheinland-Pfalz auf Kreisebene Schätzungen von Armutsgefährdungsquoten in ausreichender Qualität zu gewinnen. Um dies zu illustrieren, wurden an der Universität Trier im Rahmen einer kleinen Simulationsstudie wiederholt einfache Zufallsstichproben aus synthetischen, 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 dann eine Schätzung der Armutsgefährdungsquoten. Die Ergebnisse der Schätzung sind in Abbildung 14.4 für drei Stichproben anhand eingefärbter Landkarten dargestellt.

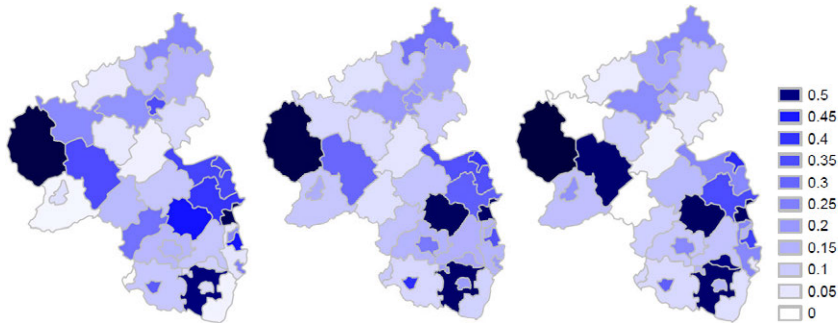


Abb. 14.4: *Stichprobenfehler bei der Schätzung von Armutsgefährdungsquoten in den Kreisen von Rheinland-Pfalz (Karte: Bundesamt für Kartographie und Geodäsie / Geodatenzentrum)*

Die unterschiedlichen Färbungen repräsentieren – analog zu Abbildung 4.2 – Klassen für die Armutsgefährdungsquote, wobei die Definition der Klassengrenzen in der Legende wiedergegeben ist – der Wert 0,2 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, wie in Exkurs 5.3 näher ausgeführt, 60 % des Medians der Einkommensverteilung für Deutschland.

Man kann aus den drei Karten auch ohne Angabe der genauen Schätzergebnisse erkennen, dass der Stichprobenfehler bei kleinen Auswahlsätzen $\frac{n}{N}$ beachtlich sein kann. 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.[4]

[4] Details zur Schätzung von Armutsgefährdung in der EU auf der Basis von EU-SILC für das Jahr 2010 sind einer von Eurostat in der Reihe *Statistics in Focus* herausgegebenen Schrift zu entnehmen; s. Antuofermo / Di Meglio (2012). Danach lag die Armutsgefährdungsquote in Deutschland bei 19,7 %.

15 Statistische Testverfahren

Das 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. nichtparametrische Tests) oder nach der Verteilung der Prüfgröße (z. B. t-Test, χ^2-Test, Gauß-Test bei Normalverteilung). 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 Gütefunktion 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 (χ^2-Test). Abschließend wird noch ein Test mit χ^2-verteilter Testgröße vorgestellt, der auf die Prüfung der Unabhängigkeit zweier Merkmale abzielt.

Vorschau auf das Kapitel

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 der Stichprobe inhärente Information zu einer Stichprobenfunktion und nutzt deren Ausprägung dazu, eine Entscheidung bezüglich der zu testenden Hypothesen zu fällen:

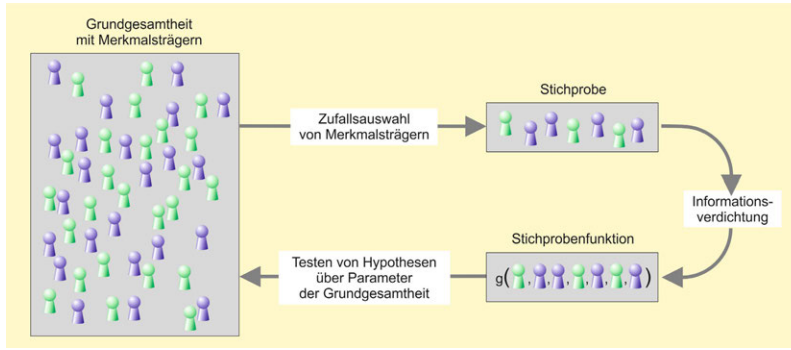


Abb. 15.1: *Vorgehensweise beim Testen von Hypothesen*

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_15

Die Grundidee besteht also wie bei der Schätzung darin, Stichprobenbefunde zu verwenden und aus diesen Aussagen für eine umfassendere Grundgesamtheit abzuleiten.

15.1 Arten statistischer Tests

Klassifikationen für statistische Tests:

- nach der Anzahl der Stichproben

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.

- in Abhängigkeit von Verteilungsannahmen

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**.

- nach dem Inhalt der Hypothesen

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.

- nach der Verteilung der Prüfstatistik

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 normalverteilter 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.

- nach der Hypothesenformulierung

Bei der Prüfung von Hypothesen über Parameter kann es darauf ankommen, 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 15.1: 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 in der europäischen Sozialpolitik der Ausgangspunkt für Strategien zur Verringerung eines geschlechtsspezifischen Verdienstgefälles. Mit den Daten der Erhebung lässt sich auch klären, ob das Ausbildungsniveau, operationalisiert über die ordinalskalierte Variable „Höchster erreichter Bildungsabschluss“, einen wesentlichen Effekt auf das Einkommen hat.
- 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 mit einem neuen Medikament tatsächlich eine von einem Pharmakonzern behauptete Wirkung erzielt wird. 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 werden.
- Bei der industriellen Serienproduktion ist man daran interessiert, die mittlere Lage $\mu = E(X)$ eines häufig als normalverteilt spezifizierten Qualitätsmerkmals X zu überwachen. Anhand von Stichproben, die der laufenden Produktion (Grundgesamtheit) in regelmäßigen Abständen entnommen werden, will man eine Aussage darüber ableiten, ob der Verteilungsparameter μ noch auf einem Sollniveau μ_0 liegt oder sich verändert hat. Bei Eintritt eines Shifts soll möglichst rasch korrigierend in den Fertigungsprozess eingegriffen werden.

Unterscheidung von statistischen und psychologischen Tests

In der Psychologie wird der Begriff „Test“ häufig in anderem Sinne verwendet, nämlich in der **Diagnostik** als routinemäßig einsetzbares Messinstrument zur Erfassung latenter Variablen bzw. hypothetischer Konstrukte anhand von Fragebögen – vgl. etwa Sedlmeier / Renkewitz (2013). Es geht dabei um die Bestimmung der relativen Position von Individuen oder Gruppen bezüglich bestimmter Persönlichkeitsmerkmale, etwa „Leistungsmotivation“, „Intelligenz“ oder „Teamfähigkeit“. Oft werden in der Psychologie anstelle solcher Einzelmerkmale auch ganze Bündel von Persönlichkeitsmerkmalen anhand eines einzigen Tests gemessen, etwa die sog. *Big Five* der Persönlichkeitspsychologie. [1] Um solche psychologische Tests geht es aber in diesem Kapitel *nicht*. Es geht im Folgenden vielmehr um die Konfrontation von Forschungshypothesen mit Daten mit dem Ziel, Aufschluss darüber zu gewinnen, ob eine Hypothese mit vorhandenen Beoachtungen verträglich ist und daher bis auf weiteres beizubehalten ist oder ob sie aufgrund des empirischen Befunds zu verwerfen ist. Die letztgenannte Testentscheidung wird getroffen, 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 15.1).

15.2 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 einer nicht heilbaren Krankheit, bei der der Erwartungswert μ für die verbleibende Restlebenszeit bisher den Wert $\mu = \mu_0$ hatte, anhand von Patientendaten belegen, dass eine neue Operationsmethode oder ein neues Medikament zu einem größeren Erwartungswert für die Restlebenszeit

[1] Die „Big Five“ sind „Offenheit gegenüber neuen Erfahrungen“ (engl.: ***o**penness to new experience*), Gewissenhaftigkeit (***c**onscientiousness*), Extraversion“ (***e**xtraversion*), Verträglichkeit (***a**greeableness*) und „Neurotizismus (***n**eurotizism*). Sie werden nach den englischsprachigen Faktorenbezeichnungen auch mit OCEAN abgekürzt – zu Details und Modifikationen vgl. z. B. Asendorpf / Neyer (2012).

führt, wird man $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ testen (einseitiger Test). Bei der Überwachung des mittleren Füllvolumens für Tinte bei Tintendruckerpatronen, wird sich das Eichamt oder der Kunde vor allem für Unterschreitungen des angegebenen Füllvolumens interessieren und $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$ testen (einseitiger Test). Der Hersteller wird hingegen das mittlere Niveau μ möglichst genau auf dem Zielwert μ_0 halten wollen, um den gesetzlichen Vorschriften zu genügen (keine Unterschreitung des etikettierten Füllvolumens) und gleichzeitig nichts zu verschenken (keine Überschreitung), d. h. er wird $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ testen (zweiseitiger Test).

Teststatistiken sind Zufallsvariablen

Ein Test basiert auf einer **Prüfvariablen**, auch **Prüf-** oder **Teststatistik** genannt, deren Ausprägung sich im Ein-Stichprobenfall aus einer Stichprobe $x_1, x_2, .., x_n$ ergibt. Letztere wird als Realisation von Stichprobenvariablen $X_1, X_2, .., X_n$ interpretiert. Die Stichprobenvariablen werden nicht direkt verwendet; man aggregiert sie vielmehr anhand einer Stichprobenfunktion $g(X_1, X_2, .., X_n)$, z. B. anhand des Stichprobenmittelwerts $\overline{X}$ 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 $g(x_1, x_2, .., x_n)$ der herangezogenen Stichprobenfunktion ab.

Zweiseitiger Test für den Erwartungswert

Die Vorgehensweise bei einem Hypothesentest sei anhand der Überwachung eines Produktmerkmals X bei der industriellen Serienfertigung illustriert. Man weiß aufgrund von Voruntersuchungen des Produktionsprozesses, dass das Merkmal X (z. B. Durchmesser einer Produktkomponente) exakt oder approximativ normalverteilt ist mit Erwartungswert $\mu = E(X)$ und Varianz $\sigma^2 = V(X)$. Für die Qualität des Endprodukts ist es wichtig, dass die Ausprägungen von X innerhalb eines bestimmten Toleranzintervalls liegen, weil sonst die Funktionsfähigkeit des Produkts nicht mehr gewährleistet ist und Ausschuss produziert wird. Zielwert für X ist die Mitte des Toleranzintervalls, die mit μ_0 bezeichnet sei. Man will sich während der Fertigung vergewissern, dass sich das Fertigungsniveau $\mu = E(X)$ für das Merkmal nicht zu weit nach oben oder unten vom Zielwert entfernt hat und testet in regelmäßigen Abständen

$$H_0 : \ \mu = \mu_0 \quad \text{gegen} \quad H_1 : \ \mu \neq \mu_0. \tag{15.1}$$

Um den *zweiseitigen* Test durchführen zu können, benötigt man Daten aus einer Stichprobe $x_1, \ldots, x_n$, die bei dem hier gewählten Beispiel der laufenden Produktion entnommen wird. Die Stichprobeninformation ermöglicht es, für den unbekannten Lageparameter μ eine Schätzung $\widehat{\mu}$ zu gewinnen. Als Schätzfunktion bietet sich der Stichprobenmittelwert $\widehat{\mu} = \overline{X}$ an, der als Prüf- oder Testgröße für den Test (15.1) fungiert. Wenn H_0 zutrifft, 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 $\alpha = 0{,}01$. Liegt die aus den Stichprobendaten errechnete Ausprägung der Prüfstatistik außerhalb des Intervalls, wird die Nullhypothese verworfen. Die Testentscheidung basiert also auf der Verteilung der Prüfstatistik unter H_0.

Prüfgröße bei bekannter Varianz

Wenn die Varianz σ^2 von X bekannt ist, gilt unter H_0, also für $\mu = \mu_0$, nach (13.6) die Aussage $\overline{X} \sim N(\mu_0; \sigma^2_{\overline{X}})$ mit $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. Man kann die Prüfgröße $\overline{X}$ direkt verwenden oder aber zweckmäßigerweise erst einmal nach (12.11) standardisieren. Für die standardisierte Testvariable

$$Z := \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n} \tag{15.2}$$

gilt, dass eine Ausprägung 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 liegt (vgl. erneut Abbildung 12.4). Nur wenn die standardisierte Prüfgröße (15 2) innerhalb dieses Intervalls liegt, wird weiter von der Gültigkeit der Nullhypothese H_0 ausgegangen. Das Intervall heißt **Annahmebereich** für H_0. Der Bereich außerhalb des genannten Intervalls definiert den **Ablehnungsbereich** für die Nullhypothese und die Grenzen des Intervalls werden als **kritische Werte** bezeichnet. Im Falle der Verwerfung von H_0 ist die Alternativhypothese H_1 statistisch „bewiesen“ in dem Sinne, dass ihre Gültigkeit mit einer Irrtumswahrscheinlichkeit α als gesichert angenommen werden kann. Die fälschliche Zurückweisung der Nullhypothese wird **Fehler 1. Art** oder **α-Fehler** genannt. Die Wahrscheinlichkeit α für den Eintritt eines Fehlers 1. Art definiert das **Signifikanzniveau** des Tests.

Abbildung 15.2 zeigt den Annahme- und Ablehnungsbereich für den mit der Prüfgröße $\overline{X}$ bzw. der standardisierten Variablen Z operierenden zweiseitigen Hypothesentest (15.1). Der Test wird auch als **Gauß-Test** bezeichnet, weil er mit einer normalverteilten Prüfvariablen arbeitet. Die Grafik zeigt die Dichte des Stichprobenmittelwerts $\overline{X}$ unter H_0, die nach Transformation der Abszissenachse auch die Dichte der standardnormalverteilten Variablen Z darstellt. Die obere Grenze des Annahmebereichs, also das $(1 - \frac{\alpha}{2})$-Quantil $z_{1-\alpha/2}$, hat z. B. bei Wahl von $\alpha = 0{,}05$ nach Tabelle 19.3 den Wert $z_{0{,}975} = 1{,}96$. Für die untere Grenze $z_{0{,}025}$ gilt wegen (12.25) dann $z_{0{,}025} = -z_{0{,}975} = -1{,}96$. Für $\alpha = 0{,}05$ ist der Annahmebereich also durch das Intervall $[-1{,}96; 1{,}96]$ gegeben. Die Wahrscheinlichkeit α für den Eintritt eines Fehlers 1. Art ist in Abbildung 15.2 durch den Inhalt der beiden markierten Flächen repräsentiert.

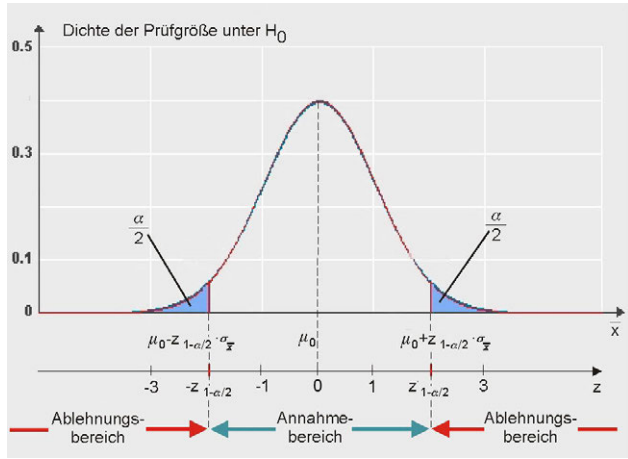


Abb. 15.2: *Annahme- und Ablehnungsbereich für H_0 (zweiseitiger Test für den Erwartungswert (normalverteiltes Merkmal; Varianz bekannt)*

Falls die Nullhypothese $H_0 : \mu = \mu_0$ zutrifft, wird sie demnach mit Irrtumswahrscheinlichkeit α verworfen. Die Ablehnung von H_0 erfolgt, wenn sich für die aus der Stichprobenfunktion $\widehat{\mu} = \overline{X}$ durch Standardisierung hervorgegangene Variable Z eine Realisation ergibt, die außerhalb des Intervalls $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$ liegt. Letzteres beinhaltet, dass für den Betrag $|z|$ der Teststatistik (15.2) die Bedingung

$$|z| > z_{1-\alpha/2} \tag{15.3}$$

erfüllt ist. Abbildung 15.2 verdeutlicht auch, dass man den Test ebenso anhand der nicht-standardisierten Prüfgröße $\overline{X}$ durchführen kann. Obwohl beide Ansätze äquivalent sind, bietet sich die Standardisierung natürlich an, weil man hier zu einer Ablehnungsbedingung kommt, die nicht mehr vom Wert μ_0 abhängt.

Beispiel 15.2: Anwendung von Qualitätsregelkarten

Eine effiziente und weitverbreitete Methode zur Vermeidung von Fehlern in der industriellen Massenfertigung ist die statistische Prozessregelung (engl.: *s*tatistical *p*rocess *c*ontrol, kurz *SPC*) mit sog. *Qualitätsregelkarten*. Deren Anwendung entspricht der wiederholten Durchführung eines Tests. Man geht 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 oder 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. Bei Qualitätsregelkarten zur Überwachung des mittleren Merk-

malsniveaus wird die Streuung, repräsentiert durch die Standardabweichung σ der Normalverteilung, aufgrund der Auswertung von Prozessvorläufen i. Allg. als bekannt angenommen.

Abb. 15.3: *Anwendung einer Qualitätsregelkarte (Serienproduktion von Platinen; Quelle: Fa. Böhme und Weihs Systemtechnik)*

Abbildung 15.3 zeigt eine solche Qualitätsregelkarte, die hier zur Überwachung der Länge von Platinen eingesetzt wird. Eine möglichst gute Längenkonstanz ist qualitätsrelevant, weil die Platinen maßgenau in ein Steuergerät einer Maschine (Optomouse) eingebaut werden. Die obere Grafik zeigt den beobachteten Stichprobenmittelwert $\overline{x}$ im zeitlichen Verlauf. Sobald die Zeitreihe eine untere Linie unter- bzw. eine obere Linie überschreitet, erfolgt ein prozesskorrigierender Eingriff. Diese beiden – im e-Buch rot dargestellten – Linien werden in der Qualitätssicherung *Eingriffsgrenzen* genannt. Sie definieren die in Abbildung 15.2 veranschaulichten Grenzen zwischen Annahme- und Ablehnbereich des Gauß-Tests. Die auf dem Foto sichtbare Betonung eines Zeitfensters dient der Hervorhebung und Analyse von Prozesstrends.

Einseitiger Test für den Erwartungswert bei Normalverteilung

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

$$H_0 : \ \mu \leq \mu_0 \quad \text{gegen} \quad H_1 : \ \mu > \mu_0 \tag{15.4}$$

und im *linksseitigen* Fall

$$H_0 : \ \mu \geq \mu_0 \quad \text{gegen} \quad H_1 : \ \mu < \mu_0. \tag{15.5}$$

Während der Annahmebereich beim zweiseitigen Test, wie in Abbildung 15.2 veranschaulicht, durch das $\frac{\alpha}{2}$-Quantil und das $(1-\frac{\alpha}{2})$-Quantil der Standardnormalverteilung begrenzt wird, ist der Annahmebereich

beim einseitigen Test nur durch ein einziges Quantil vom Ablehnungsbereich getrennt. Beim rechtsseitigen Test ist es das $(1-\alpha)$-Quantil, beim linksseitigen Test das α-Quantil. Die Bedingung für die Ablehnung der Nullhypothese lautet also beim *rechtsseitigen* Test

$$z > z_{1-\alpha}, \tag{15.6}$$

und beim *linksseitigen* Test

$$z < z_\alpha = -z_{1-\alpha}. \tag{15.7}$$

Der Annahmebereich beim einseitigen Hypothesentest wird allein durch die Verteilung der Prüfgröße an der Stelle $\mu = \mu_0$ bestimmt, hängt also von der Verteilung der Prüfgröße am Rande des Gültigkeitsbereichs der Nullhypothese ab.

Fehlerarten beim Testen

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 insgesamt vier denkbare Fälle, von denen zwei falsche Entscheidungen darstellen. Neben dem schon genannten **Fehler 1. Art** oder **α-Fehler**, der fälschlichen Verwerfung der Nullhypothese, kann auch eine Nicht-Verwerfung einer nicht zutreffenden Nullhypothese eintreten. Beim Test (15.1) kann ja der Fall eintreten, dass der unbekannte Parameter μ nicht mit μ_0 übereinstimmt, die Realisation der Prüfgröße $\overline{X}$ aber dennoch in den Annahmebereich fällt. Diese Fehlentscheidung bei einem Hypothesentest heißt **Fehler 2. Art** oder auch **β-Fehler**. Bei dem herangezogenen Beispiel der Überwachung des mittleren Niveaus eines Qualitätsmerkmals bei der Serienfertigung ist der Fehler 1. Art als blinder Alarm zu interpretieren, der Fehler 2. Art als unterlassener Alarm.

Tabelle 15.1 zeigt, welche Ausgänge bei einem Hypothesentest möglich sind und wie man die Testentscheidungen zu bewerten hat.

Testentscheidung	**tatsächlicher Zustand**	
	Nullhypothese richtig	Nullhypothese falsch
Nullhypothese nicht verworfen	richtige Entscheidung	**Fehler 2. Art** (β-Fehler)
Nullhypothese verworfen	**Fehler 1. Art** (α-Fehler)	richtige Entscheidung

Tab. 15.1: *Ausgänge bei einem Hypothesentest*

Die Wahrscheinlichkeiten für die in Tabelle 15.1 aufgeführten Testfehler sind offenbar bedingte Wahrscheinlichkeiten:

$$P(\text{Fehler 1. Art}) = P(\text{Ablehnung von } H_0 | H_0 \text{ ist wahr}) \tag{15.8}$$

$$P(\text{Fehler 2. Art}) = P(\text{Nicht-Verwerfung von } H_0 | H_1 \text{ ist wahr}). \tag{15.9}$$

Abbildung 15.4 veranschaulicht die Fehlerarten am Beispiel des *rechtsseitigen* Tests (15.4). Sie zeigt in beiden Abbildungsteilen zwei Dichten, wobei die erste Kurve jeweils die Dichte des Stichprobenmittelwerts für $\mu = \mu_0$ darstellt. Der Inhalt der rechts vom kritischen Wert $z_{1-\alpha}$ markierten Fläche hat den Wert α. Wenn man die auf den Fall $\mu = \mu_0$ bezogene Dichtekurve nach links verschiebt, also den Wert μ auf einen Wert $\mu < \mu_0$ absenkt, bleibt die Nullhypothese gültig; die rechts vom kritischen Wert betonte Fläche (rot im e-Buch) wird kleiner. Das **Signifikanzniveau** α ist bei einem einseitigen Test also als *obere Schranke* für den Eintritt eines Fehlers 1. Art zu interpretieren.

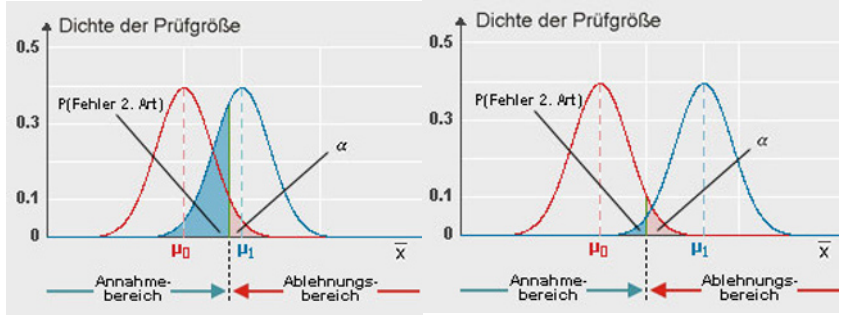


Abb. 15.4: *Wahrscheinlichkeiten für den Eintritt eines Fehlers 2. Art (rechtsseitiger Test)*

Java-Applet „Fehler 1. und 2. Art“

Ein Fehler 2. Art kann nur eintreten, wenn H_1 zutrifft. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art hängt vom Wert des Parameters μ ab. Sie wird – anders als die Wahrscheinlichkeit für den Eintritt eines Fehlers 1. Art – nicht durch das Testdesign unter Kontrolle gehalten. Dies geht aus Abbildung 15.4 hervor. Die beiden Teile der Abbildung unterscheiden sich durch die Lage der zweiten Dichtekurve. Letztere zeigt die Verteilung von $\overline{X}$ unter H_1 für zwei unterschiedliche Werte μ_1, die zum Gültigkeitsbereich von H_1 gehören. Die Wahrscheinlichkeit für den Eintritt eines Fehlers 2. Art wird durch den Inhalt der markierten Fläche links vom kritischen Wert repräsentiert (blau im e-Buch). Je weiter μ_1 nach rechts von μ_0 wegrückt, desto kleiner wird der β-Fehler. Die beiden Fehlerwahrscheinlichkeiten stehen nicht in einer Komplementärbeziehung zueinander, ergänzen sich z. B. nicht zum Wert 1.

Abbildung 15.4 zeigt Dichtekurven der noch nicht standardisierten Prüfgröße $\overline{X}$, deren Streuung vom Stichprobenumfang n abhängt. Wenn n

vergrößert wird, bleibt das Zentrum der Verteilung von $\overline{X}$ nach (14.6) unverändert. Die Streuung von $\overline{X}$ nimmt hingegen gemäß (14.7) ab; die Dichtekurven werden steiler. Dies wiederum impliziert, dass alle markierten Flächeninhalte in Abbildung 15.4 kleiner werden.[2]

Bewertung der Leistungsfähigkeit eines Tests

Zur allgemeinen Beurteilung der Leistungsfähigkeit eines Tests, im zweiseitigen Fall also eines Tests der Hypothesen (15.1) und im einseitigen Fall der Hypothesen (15.4) bzw. (15.5), zieht man die sog. **Gütefunktion**

$$G(\mu) = P(\text{Ablehnung von } H_0|\mu) \tag{15.10}$$

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 Ablehnungswahrscheinlichkeit 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 Gütefunktion die Fehlerwahrscheinlichkeiten für jeden Wert μ ablesen. Von zwei mit dem Signifikanzniveau α arbeitenden Tests wird man den Test bevorzugen, dessen Gütefunktion 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.

Gütefunktion beim rechtsseitigen Test

Wie die Gütefunktion beim *rechtsseitigen* Gauß-Test verläuft, kann man qualitativ schon anhand von Abbildung 15.4 erschließen. Für $\mu = \mu_0$ nimmt $G(\mu)$ den Wert α an, der in der Abbildung als Flächeninhalt rechts vom kritischen Wert betont ist (rot im e-Buch). Bei Werten $\mu < \mu_0$, für die ja H_0 auch gilt, gilt $G(\mu) < \alpha$. Der Wert $G(\mu)$ rückt um so näher an 0 heran, je weiter μ den Wert μ_0 unterschreitet. Dies lässt sich nachvollziehen, wenn man gedanklich in Abbildung 15.4 die auf den Fall $\mu = \mu_0$ bezogene Dichtekurve nach links verschiebt – die rechts vom kritischen Wert betonte Fläche unterhalb der Dichte wird kleiner.

Wenn μ oberhalb von μ_0 liegt, gilt hingegen $G(\mu) > \alpha$. Je weiter μ den Wert μ_0 überschreitet, desto größer wird $G(\mu)$ und um so kleiner wird die Wahrscheinlichkeit $1-G(\mu)$ für den Eintritt eines Fehlers 2. Art. Die Werte $G(\mu)$ streben schließlich gegen den Wert 1, die Fehlerwahrscheinlichkeit $1 - G(\mu)$ gegen 0. Die Gütefunktion des rechtsseitigen Tests ist somit eine *monoton wachsende* Funktion mit Werten zwischen 0 und 1 und mit

[2] Eine Erhöhung des Stichprobenumfangs reduziert, wie Abbildung 15.5 noch zeigt, bei dem hier betrachteten rechtsseitigen Test (15.4) für jedes μ mit $\mu > \mu_0$ die Eintrittswahrscheinlichkeit für einen Fehler 2. Art. Die Eintrittswahrscheinlichkeit für einen Fehler 1. Art hat für $\mu = \mu_0$ den vom Testdesign vorgegebenen Wert α, vermindert sich aber für $\mu < \mu_0$.

$G(\mu_0) = \alpha$. Sie ist, wie in Exkurs 15.1 gezeigt wird, durch

$$G(\mu) = 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) \tag{15.11}$$

gegeben. Dabei bezeichnet $\Phi(..)$ wieder die Verteilungsfunktion der Standardnormalverteilung. Mit (15.11) 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 Gütefunktionen 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 Gütefunktion als Funktion von d formuliert, hat dies den Vorzug, dass man von den jeweiligen Werten μ_0 und σ abstrahieren kann.

Einfluss des Stichprobenumfangs

Den Verlauf der Funktion (15.11) für $\alpha = 0{,}05$ und für $n = 5$ sowie für $n = 10$ zeigt Abbildung 15.5. 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 Gütefunktion für den Test mit dem größeren Stichprobenumfang unter H_1 einen steileren Verlauf aufweist, also trennschärfer ist.

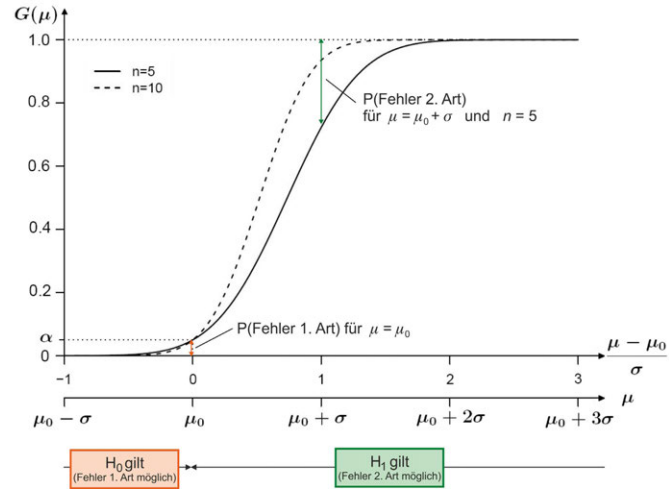


Abb. 15.5: *Gütefunktion für den rechtsseitigen Gauß-Test* ($\alpha = 0{,}05$)

Abbildung 15.5 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.

Gütefunktion beim linksseitigen Test

Für den *linksseitigen* Test gilt analog zum rechtsseitigen Fall, dass die Gütefunktion eine von 1 nach 0 streng *monoton fallende* Funktion ist und in μ_0 ebenfalls den Wert $G(\mu_0) = \alpha$ annimmt (vgl. Aufgabe 15.2). Ihre Formeldarstellung ist gegeben durch

$$G(\mu) = \Phi\left(-z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right). \tag{15.12}$$

Die Herleitung gleicht der Herleitung der Gütefunktion (15.11).

Gütefunktion beim zweiseitigen Test

Beim *zweiseitigen* Gauß-Test (15.1) 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 Gütefunktion 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

$$G(\mu) = \Phi\left(-z_{1-\alpha/2} + \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right). \tag{15.13}$$

Java-Applet „Gütefunktion (zweiseitiger Gauß-Test)“

Abbildung 15.6 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. In Abbildung 15.6 ist wieder eine zweite Abszissenachse eingezeichnet (Darstellung der Gütefunktion als Funktion der relativen Abweichung d).

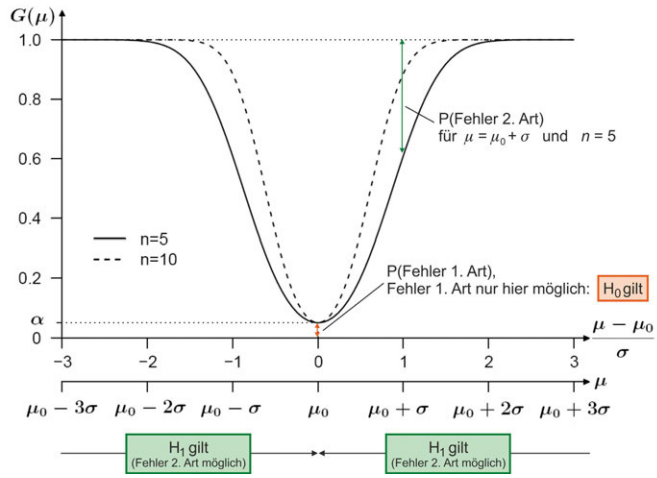


Abb. 15.6: *Gütefunktion für den zweiseitigen Gauß-Test* ($\alpha = 0{,}05$)

Wozu braucht man eine Gütefunktion?

In der Praxis liefern Gütefunktionen 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 Gütefunktion 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. Stichprobendaten kommen erst bei der Durchführung eines Tests bei der Berechnung eines Wertes für die Prüfgröße und der danach erfolgenden Testentscheidung ins Spiel.

Was sagt der p-Wert aus?

Es gibt eine Alternative für die Durchführung von Hypothesentests. Der einzige 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:

$$\text{Ablehnung von } H_0 \Leftrightarrow \text{p-Wert } \alpha' \text{ erfüllt die Bedingung } \alpha' \leq \alpha \quad (15.14)$$

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 $F_{10;15;0,95} = 2{,}54$ waren schon in Teil a von Abbildung 12.7 dargestellt. In Abbildung 15.7 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.

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 hier 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.

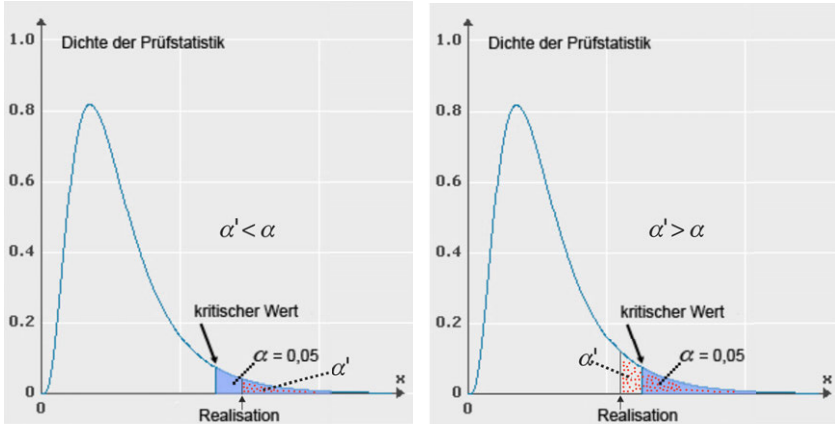


Abb. 15.7: *p-Wert bei einem F-Test*

Beispiel 15.3: 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 $\overline{x} = 1{,}996$ kg errechnet. Es soll anhand eines statistischen Tests (15.5) 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 (15.7) genau dann, wenn die Ausprägung

$$z = \frac{\overline{x} - 2}{0{,}01} \cdot \sqrt{10} = 100 \cdot (1{,}996 - 2) \cdot \sqrt{10} \approx -1{,}2649$$

der Prüfstatistik Z aus (15.2) den Wert $z_{0,05} = -z_{0,95} = -1{,}6449$ unterschreitet. Da dies hier nicht zutrifft, kann H_0 nicht verworfen werden. Die Differenz zwischen $\overline{x} = 1{,}996$ kg und dem Soll-Füllgewicht von 2 kg ist also nur auf zufällige Abweichungen zurückzuführen und statistisch nicht signifikant.

Aufgabe 15.1-3

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.

Exkurs 15.1: Gütefunktion beim rechtsseitigen Gauß-Test

Beim rechtsseitigen Gauß-Test (15.4) erfolgt die Ablehnung der Nullhypothese H_0, wenn für die Realisation z der Prüfgröße Z aus (15.2) die Bedingung $z > z_{1-\alpha}$ erfüllt ist. Für die Gütefunktion (15.11) beinhaltet dies, dass

$$G(\mu) = P(Z > z_{1-\alpha}|\mu) = P\left(\frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} > z_{1-\alpha}|\mu\right)$$
$$= P\left(\frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n} > z_{1-\alpha}|\mu\right).$$

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:

$$G(\mu) = P\left(\frac{\overline{X} - \mu_0 + \mu - \mu}{\sigma} \cdot \sqrt{n} > z_{1-\alpha}|\mu\right)$$
$$= P\left(\frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} > z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}|\mu\right).$$

Da der in der Klammer vor dem Ungleichheitszeichen stehende Term standardnormalverteilt ist, folgt bei Beachtung von (12.20)

$$G(\mu) = 1 - P\left(\frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} < z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}|\mu\right)$$
$$= 1 - \Phi\left(z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right).$$

Dies ist die herzuleitende Darstellung (15.11).

15.3 t-Test für Erwartungswerte

Die Hypothesen (15.1), (15.4) und (15.5) beziehen sich auf den Erwartungswert eines normalverteilten Merkmals X. Die Verwendung der Prüfgröße (15.2) 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 (15.2) ist dann σ durch eine Schätzung $\widehat{\sigma}$ zu ersetzen, wobei man wegen (14.9) anstelle der Stichprobenstandardabweichung S die korrigierte Stichprobenstandardabweichung $S^* = \widehat{\sigma}$ wählt. Nach (13.10) ist die resultierende Prüfstatistik

$$T := \frac{\overline{X} - \mu_0}{S^*} \cdot \sqrt{n} \tag{15.15}$$

t-verteilt mit $n-1$ Freiheitsgraden. Man kann den Annahme- und Ablehnungsbereich des mit (15.15) operierenden zweiseitigen Tests analog zu Abbildung 15.1 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.

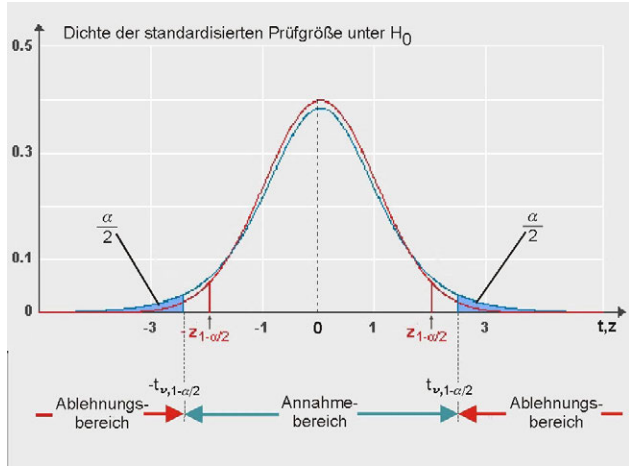


Abb. 15.8: *Annahme- und Ablehnungsbereich für H_0 (zweiseitiger Test für den Erwartungswert (normalverteiltes Merkmal, Varianz geschätzt, $\nu := n-1=6$)*

Die an Abbildung 12.6 anknüpfende Abbildung 15.8 zeigt die Dichte der t-verteilten Variablen (15.15) mit 6 Freiheitsgraden unter H_0 und den Annahme- und Ablehnungsbereich des *zweiseitigen* t-Tests. Die Dichte der Standardnormalverteilung aus Abbildung 15.1 ist zu Vergleichszwecken ebenfalls eingezeichnet. Der Annahmebereich $[-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$ des t-Tests ist stets breiter als der Annahmebereich $[-z_{1-\alpha/2}; z_{1-\alpha/2}]$ des zweiseitigen 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 (15.15) operierenden zweiseitigen Test verworfen, wenn die Prüfgröße außerhalb des in Abbildung 15.8 veranschaulichten Intervalls $[-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$ liegt, d. h. wenn

$$|t| > t_{n-1;1-\alpha/2} \tag{15.16}$$

Java-Applet „Ablehnungsbereich (t-Test)“

gilt. Beim *rechtsseitigen* t-Test erfolgt die Ablehnung von H_0 im Falle

$$t > t_{n-1;1-\alpha} \tag{15.17}$$

und im *linksseitigen* Fall für

$$t < t_{n-1;\alpha} = -t_{n-1;1-\alpha}. \tag{15.18}$$

Abbildung 15.9 zeigt wie Abbildung 15.8 im rechtsseitigen Fall zu modifizieren ist. Anstelle des $\frac{\alpha}{2}$-Quantils und des $(1-\frac{\alpha}{2})$-Quantils, die in Abbildung 15.7 die Grenzen des Annahmebereichs markieren, wird der Annahmebereich nun durch ein einziges Quantil vom Ablehnungsbereich getrennt. Beim *rechtsseitigen* Test ist es das $(1-\alpha)$-Quantil, beim *linksseitigen* Test das α-Quantil.

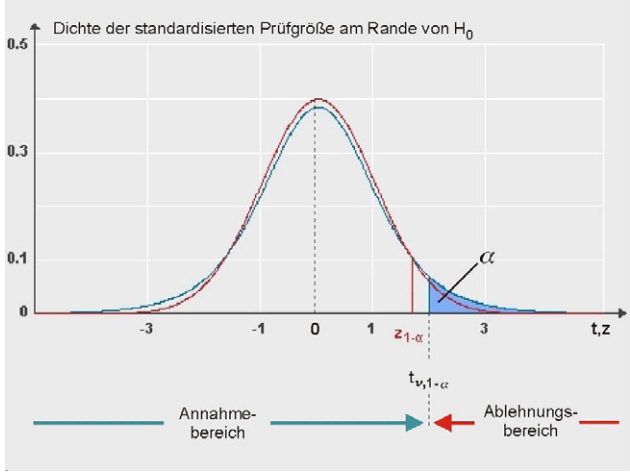


Abb. 15.9: *Annahme- und Ablehnungsbereich für H_0 (rechtsseitiger Test, Varianz geschätzt, $\nu := n-1 = 6$)*

15.4 χ^2-Test für Varianzen

Hypothesen, Prüfgröße und Ablehnbedingungen

Die Ausführungen aus Abschnitt 15.3 über das Testen zwei- und einseitiger Hypothesen für Erwartungswerte bei normalverteiltem Merkmal lassen sich leicht auf Hypothesen für Varianzen übertragen. Die Vorgehensweise sei hier nur angerissen.

Die zu (15.1) analogen Hypothesen für den zweiseitigen Fall lauten nun

$$H_0: \ \sigma^2 = \sigma_0^2 \qquad \text{gegen} \qquad H_1: \ \sigma^2 \neq \sigma_0^2. \tag{15.19}$$

Der Test wird durchgeführt mit der Prüfstatistik

$$T := \frac{n \cdot S^2}{\sigma_0^2} = \frac{(n-1) \cdot S^{*2}}{\sigma_0^2}. \tag{15.20}$$

Diese folgt nach (13.9) einer χ^2-Verteilung mit $n-1$ Freiheitsgraden: $T \sim \chi^2_{n-1}$. Die Nullhypothese aus (15.19) wird bei diesem zweiseitigem **χ^2-Test** mit Irrtumswahrscheinlichkeit α verworfen, wenn die Realisation t der Prüfgröße entweder kleiner als $\chi^2_{n-1;\alpha/2}$ oder größer als $\chi^2_{n-1;1-\alpha/2}$

ist, wenn also die berechnete Testgröße die Bedingung

$$t \notin [\chi^2_{n-1;\alpha/2}; \chi^2_{n-1;1-\alpha/2}] \tag{15.21}$$

(lies: t ist *nicht Element* von ..) 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 15.8 – nicht symmetrisch zueinander liegen, weil die χ^2-Verteilung asymmetrisch ist.

Für den einseitigen Fall hat man anstelle von (15.4) und (15.5)

$$H_0: \ \sigma^2 \leq \sigma_0^2 \quad \text{gegen} \quad H_1: \ \sigma^2 > \sigma_0^2 \qquad \textit{(rechtsseitiger Test)} \tag{15.22}$$

resp.

$$H_0: \ \sigma^2 \geq \sigma_0^2 \quad \text{gegen} \quad H_1: \ \sigma^2 < \sigma_0^2 \qquad \textit{(linksseitiger Test)}. \tag{15.23}$$

Beim rechtsseitigen Test wird H_0 verworfen, wenn für die Realisation t der Testgröße T aus (15.20)

$$t > \chi^2_{n-1;1-\alpha} \tag{15.24}$$

gilt. Die Ablehnbedingung für H_0 beim linksseitigen Test lautet

$$t < \chi^2_{n-1;\alpha}. \tag{15.25}$$

Die Ablehnbereiche lassen sich analog zu Abbildung 15.8 und Abbildung 15.9 veranschaulichen. Man muss nur die Dichte der χ^2-Verteilung mit $n-1$ Freiheitsgraden visualisieren (vgl. Abbildung 12.5) und dann in der Grafik beim zweiseitigen Test die Quantile $\chi^2_{n-1;\alpha/2}$ und $\chi^2_{n-1;1-\alpha/2}$, beim rechtsseitigen Test das Quantil $\chi^2_{n-1;1-\alpha}$ und beim linksseitigen Test das Quantil $\chi^2_{n-1;\alpha}$ auf der Abszissenachse markieren. Die Quantile sind jeweils Tabelle 19.4 des Anhangs zu entnehmen.

15.5 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 (15.2) als Prüfvariable herangezogen (**Gauß-Test**), bei geschätzter Varianz die in (15.15) 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 Teilpopulationen 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_1 und X_2 interpretieren, für die zwei separate Stichproben des Umfangs n_1 bzw. n_2 vorliegen, und anhand eines **Zweistichproben-Tests** untersuchen, ob sich die Erwartungswerte $\mu_1 := E(X_1)$ und $\mu_2 := E(X_2)$ beider Zufallsvariablen signifikant unterscheiden. Getestet wird also im hier ausschließlich betrachteten *zweiseitigen* Fall anstelle von (15.1)

$$H_0 : \ \mu_1 = \mu_2 \qquad \text{gegen} \qquad H_1 : \ \mu_1 \neq \mu_2. \tag{15.26}$$

Die Zufallsvariablen X_1 und X_2 seien unabhängig, d. h. es wird unterstellt, dass **unabhängige Stichproben** vorliegen. Man spricht auch von **unverbundenen Stichproben**. [3]

Wie bei den Einstichproben-Tests wird auch bei Zweistichproben-Tests für Erwartungswerte meist Normalverteilung unterstellt, also $X_1 \sim N(\mu_1; \sigma_1^2)$ und $X_2 \sim N(\mu_2; \sigma_2^2)$. 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

Prüfvariablenkonstruktion

$$D := \overline{X}_1 - \overline{X}_2 \tag{15.27}$$

der Stichprobenmittelwerte aus. Nach (13.6) gilt $\overline{X}_1 \sim N(\mu_1; \sigma^2_{\overline{X}_1)}$ und $\overline{X}_2 \sim N(\mu_2; \sigma^2_{\overline{X}_2})$. Für die Differenz D ergibt sich daraus mit (12.17) und der vorausgesetzten Unabhängigkeit der Stichproben

$$D \sim N(\mu_D; \sigma_D^2) \quad \text{mit} \quad \mu_D = \mu_1 - \mu_2; \quad \sigma_D^2 = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}. \tag{15.28}$$

Für die Varianz σ_D^2 kann man wegen (14.7) auch

$$\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{15.29}$$

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

[3] Bei Zwei-Stichproben-Untersuchungen in der Psychologie ist die Unabhängigkeit von X_1 und X_2 zum Beispiel verletzt, wenn man für dieselben Personen zu zwei verschiedenen Zeitpunkten Daten erhebt, also eine *Messwertwiederholung* durchführt, etwa um Effekte intervenierender Maßnahmen zu quantifizieren. Man hat dann **abhängige Stichproben**, auch **verbundene Stichproben** genannt.

Hypothesen (15.26) anhand der standardnormalverteilten Prüfgröße

$$Z = \frac{D}{\sigma_D} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{15.30}$$

durchführen kann. Haben die beiden Varianzen denselben Wert, etwa $\sigma^2 := \sigma_1^2 = \sigma_2^2$, vereinfacht sich (15.30) zu

Test bei Gleichheit von σ_1^2 und σ_2^2

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}.$$

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.

Test bei Ungleichheit von σ_1^2 und σ_2^2

Bei unbekannten Varianzen σ_1^2 und σ_2^2 ist σ_D^2 zu schätzen. Die Vorgehensweise sei nur angerissen. Bezeichnet an die analog zu (13.5) definierten korrigierten Stichprobenvarianzen mit S_1^{*2} resp. S_2^{*2}, so liefert

$$\widehat{\sigma}_D^2 := \frac{(n_1 - 1) \cdot S_1^{*2} + (n_2 - 1) \cdot S_2^{*2}}{(n_1 - 1) + (n_2 - 1)} \tag{15.31}$$

eine erwartungstreue Schätzung für σ_D^2, die die beiden Stichprobenvarianzen mit dem Umfang der Stichprobenumfänge gewichtet. Einsetzen in (15.30) führt zur Prüfstatistik

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1-1)\cdot S_1^{*2} + (n_2-1)\cdot S_2^{*2}}{n_1+n_2-2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{15.32}$$

des **Zweistichproben-t-Tests**. Für die Prüfvariable (15.32) 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.

15.6 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 8 wurde die gemeinsame empirische Verteilung der beiden Merkmale in einer Kontingenztabelle zusammengefasst (vgl. Tabelle 8.1). Diese weist die Häufigkeiten h_{ij} für die möglichen Kombinationen der Ausprägungen $a_1, \ldots, a_k$ von X resp. $b_1, \ldots, b_m$ von Y aus.

Der durch (9.1) und (8.9) definierte χ^2**-Koeffizient** wurde in Kapitel 9 als Zusammenhangsmaß eingeführt. In der schließenden Statistik bietet sich der χ^2-Koeffizient als Testgröße für einen Test

$$\begin{aligned} H_0 &: \; X \text{ und } Y \text{ sind unabhängig} \qquad \text{gegen} \\ H_1 &: \; X \text{ und } Y \text{ sind abhängig} \end{aligned} \tag{15.33}$$

auf Unabhängigkeit zweier diskreter Merkmale X und Y an. Dieser nichtparametrische 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 (9.1) 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 (9.1) errechnete Testgröße

$$\chi^2 > \chi^2_{(k-1)\cdot(m-1);1-\alpha} \tag{15.34}$$

gilt. Das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $(k-1)\cdot(m-1)$ Freiheitsgraden entnimmt man Tabelle 19.4 des Anhangs.

Beispiel 15.4: Unabhängigkeitstest mit Politbarometer-Daten

In Beispiel 9.1 wurde anhand der Daten des ZDF-Politbarometers vom 16. Oktober 2009 für das Zusammenhangsmaß (9.1) der Wert $\chi^2 = 9{,}79$ errechnet. Da die Werte in der zugrunde liegenden Kontingenztabelle 8.9 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 $(6-1)(2-1) = 5$ Freiheitsgraden. Die Nullhypothese H_0 aus (15.33) kann somit bei Vorgabe der Irrtumswahrscheinlichkeit $\alpha = 0{,}05$ nicht verworfen werden, weil

$$\chi^2 = 9{,}79 < \chi^2_{5;0,95} = 11{,}070$$

gilt, die Ablehnbedingung (15.34) also nicht erfüllt ist. Testet man hingegen mit $\alpha = 0{,}10$, kann H_0 abgelehnt werden. Es gilt dann ja

$$\chi^2 = 9{,}79 > \chi^2_{5;0,90} = 9{,}236.$$

16 Das lineare Regressionsmodell

Regressionsmodelle zielen darauf ab, die Werte eines Merkmals oder mehrerer Merkmale (unabhängige Variablen) zur Erklärung der Werte eines anderen Merkmals (abhängige Variable) 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.

Vorschau auf das Kapitel

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.

Sir Francis Galton

Sir Francis Galton (1822 - 1911), Sohn einer Quäkerfamilie und Halbcousin von Charles Darwin (1809 - 1882), war ein wissbegieriger Weltreisender und vor allem ein überaus vielseitiger Naturforscher, der u. a. Wetterdaten auswertete und Klimakarten publizierte, erstmals Verfahren zur Personenidentifikation anhand von Fingerabdrücken entwickelte und sich – mit polarisierender Wirkung – auch zu Fragen der Vererbungslehre äußerte. 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.

Im Bereich der Statistik hat Galton nicht nur das heute als Galtonbrett benannte Demonstrationsmodell für das Zustandekommen bestimmter Wahrscheinlichkeitsverteilungen hervorgebracht, sondern 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.[1]

[1] Details zu dieser Studie sind einer Monografie über Regressionsmodelle von Fahrmeir / Kneip / Lang (2009, Kapitel 1) zu entnehmen. Das Buch behandelt auch verallgemeinerte lineare Modelle, etwa Modelle mit diskreten erklärenden Variablen, und auch nicht-parametrische Regressionsansätze. Eine umfassende Einführung in die Regressionsanalyse unter Verwendung der inzwischen weit verbreiteten Programmierumgebung *R* findet man bei Schlittgen (2013).

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_16

Galton stellte fest, dass die beobachteten Datenpaare $(x_1; y_1), \ldots, (x_n; y_n)$ 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 aber durch nicht-systematische Zufallseinflüsse überdeckt ist. Diese Studie kann als erste Regressionsanalyse der Statistik angesehen werden. 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 $X_1, \ldots, X_k$ zu erklären, wobei der Zusammenhang über eine Funktion f modelliert wird. Letztere wird **Regressionsfunktion** genannt. Im Falle nur *eines* erklärenden Merkmals spricht man vom **einfachen Regressionsmodell**, bei Verwendung *mehrerer* erklärender Merkmale vom **multiplen Regressionsmodell**.

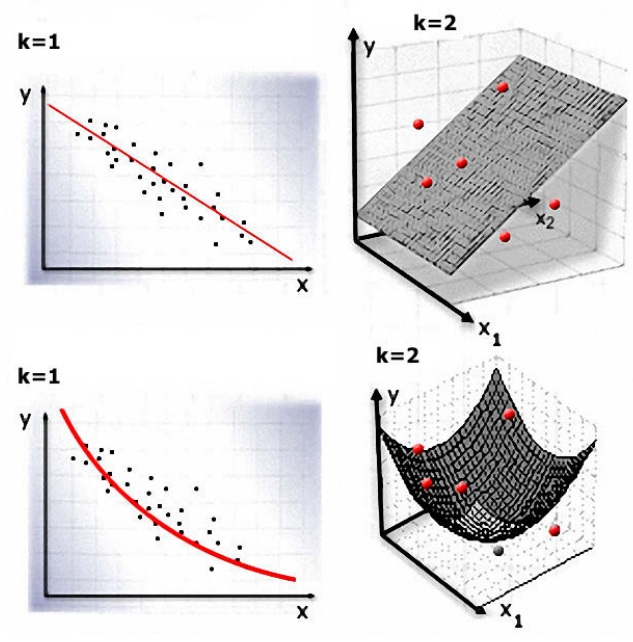


Abb. 16.1: *Lineares und nicht-lineares Regressionsmodell mit 1 - 2 erklärenden Variablen*

Wenn die Funktion f als linear spezifiziert ist, liegt ein **lineares Regressionsmodell** vor, bei nicht-linearem Funktionstyp f ein **nicht-lineares Regressionsmodell**. In allen Fällen wird angenommen, dass der funktio-

nale Zusammenhang nicht exakt gilt, sondern durch nicht-systematische zufällige Störeinflüsse überlagert ist. Nicht-systematisch meint, dass sich die Störungen „im Mittel“ aufheben.

Abbildung 16.1 zeigt im oberen Teil Datenpunkte in der Ebene bzw. im Raum und eine lineare Funktion, mit der ein funktionaler Zusammenhang zwischen einem Merkmal Y und k erklärenden Variablen modelliert wird. Der Fall $k = 1$ führt zum einfachen linearen Regressionsmodell, der Fall $k = 2$ zum multiplen linearen Regressionsmodell mit zwei unabhängigen Variablen. Der untere Teil der Abbildung bezieht sich ebenfalls auf die Fälle $k = 1$ und $k = 2$, aber auf einen nicht-linearen Regressionsansatz.

16.1 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 $(x_1, y_1), \ldots, (x_n, y_n)$ für die beiden Merkmale X und Y festlegen. Wenn – wie im linken oberen Teil von Abbildung 16.1 beispielhaft illustriert – 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 Beziehung[2]

$$y_i = \alpha + \beta x_i + u_i \qquad i = 1, \ldots, n. \tag{16.1}$$

Flash-Animation „Das einfache Regressionsmodell“

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:

[2] Die 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 findet man auch ϵ_i oder e_i statt u_i.

Modellvariable X	**Modellvariable** Y
erklärende Variable	erklärte Variable
unabhängige Variable	abhängige Variable
exogene Variable	endogene Variable
Regressor	Regressand

Tab. 16.1: *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 oder - etwa in der Psychologie - als Stimulus, während das Merkmal Y als Ziel- oder Responsevariable angesprochen wird. Im Kontext des Einsatzes von Regressionsmodellen zu Prognosezwecken 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**.

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

$$Y_i = \alpha + \beta x_i + U_i \qquad i = 1, \ldots, n \tag{16.2}$$

schreiben. Die zu (16.2) gehörenden Modellannahmen sind nachstehend aufgelistet:

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 (Fehlen von *Autokorrelation*).[3]

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 $i = 1, \ldots, n$ (Ausschluss eines trivialen Falls).

16.2 KQ-Schätzung im einfachen Regressionsmodell

Ohne den Störterm u_i wäre die lineare Regression (16.1) 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 $\widehat{R}$ verwendet und für die Geradengleichung

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta} x. \tag{16.3}$$

Flash-Animation „Bestimmung der Regressionsgeraden“

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

$$\widehat{u_i} = y_i - \widehat{y_i} = y_i - \widehat{\alpha} - \widehat{\beta} x_i \qquad i = 1, \ldots, n \tag{16.4}$$

zwischen dem Beobachtungswert y_i und dem Wert $\widehat{y_i}$ der Regressionsgeraden in der Beobachtungsperiode i zurück. Die Differenzen (16.4) werden **Residuen** genannt. Da diese sowohl positiv als auch negativ

[3] Oft wird anstelle von Unkorreliertheit von Störvariablen aus verschiedenen Beobachtungsperioden die etwas stärkere Forderung stochastischer Unabhängigkeit gefordert, die nach (13.17) Unkorreliertheit impliziert.

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 $\widehat{R}$ aus, bei der die Summe der *quadrierten* Residuen $\widehat{u_i}^2$ bezüglich der beiden Geradenparameter minimal ist:

$$\sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2 \to Min. \tag{16.5}$$

Prinzip der KQ-Schätzung

Abbildung 16.2 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 $\widehat{u_i} = y_i - \widehat{y_i}$ visualisiert. Die KQ-Regressionsgerade ist dann dadurch charakterisiert, dass für sie die in (16.5) wiedergegebene Summe ein Minimum erreicht.

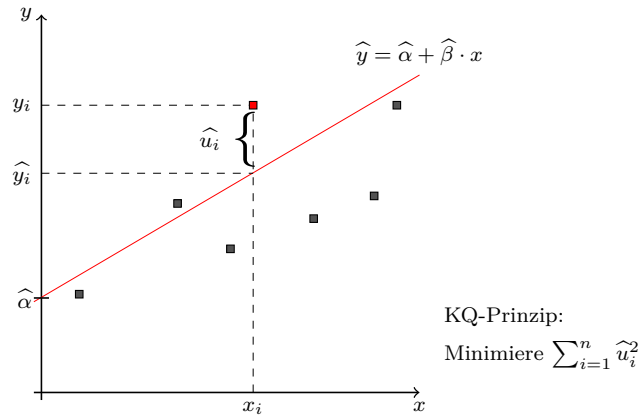


Abb. 16.2: *Veranschaulichung von „Residuum“ und KQ-Methode*

Für die KQ-Gerade und deren Koeffizienten könnte man zur Kennzeichnung des Schätzverfahrens einen Index „KQ“ anbringen, also z. B. $\widehat{R}_{KQ}$, $\widehat{\alpha}_{KQ}$ und $\widehat{\beta}_{KQ}$ schreiben.[4] Da in diesem Manuskript nur die KQ-Methode zur Schätzung von Regressionskoeffizienten verwendet wird, wird der Index unterdrückt.

KQ-Schätzung der Regressionskoeffizienten

Um eine Formel für die KQ-Schätzungen α und β zu erhalten, muss man die Summe (16.5), 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 Varianzzer-

[4] Wenn 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“.

legungsformel (5.7) für die Regressionskoeffizienten β und α[5]

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \tag{16.6}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta} \cdot \overline{x}. \tag{16.7}$$

Wenn man in die Gleichung (16.3) der Regressionsgeraden für x den Wert $x = \overline{x}$ einsetzt, resultiert für die abhängige Variable der Wert $\widehat{y} = \widehat{\alpha} + \widehat{\beta} \cdot \overline{x}$, nach (16.7) also $\widehat{y} = \overline{y}$. Dies bedeutet, dass die nach der KQ-Methode ermittelte Regressionsgerade stets durch den Schwerpunkt $(\overline{x}, \overline{y})$ des für die Schätzung herangezogenen Datensatzes $(x_1; y_1), \ldots, (x_n; y_n)$ geht. Mit (16.7) kann man außerdem ableiten, dass die Summe der KQ-Residuen stets Null ist. Setzt man nämlich in (16.4) für $\widehat{\alpha}$ gemäß (16.7) den Term $\overline{y} - \widehat{\beta} \cdot \overline{x}$ ein, erhält man zunächst

Eigenschaften der KQ-Regression

$$\widehat{u}_i = y_i - (\overline{y} - \widehat{\beta} \cdot \overline{x}) - \widehat{\beta} \cdot x_i = y_i - \overline{y} + \widehat{\beta} \cdot \overline{x} - \widehat{\beta} \cdot x_i \qquad i = 1, \ldots, n$$

und hieraus durch Aufsummieren der n Terme

$$\sum_{i=1}^{n} \widehat{u}_i = \sum_{i=1}^{n} y_i - n\overline{y} + \widehat{\beta} \cdot n\overline{x} - \widehat{\beta} \cdot \sum_{i=1}^{n} x_i = n \cdot (\overline{y} - \overline{y} + \overline{x} - \overline{x}) = 0.$$

Dies bedeutet, dass die KQ-Schätzung fehlerausgleichend wirkt, in dem Sinne, dass sich die Abweichungen $\widehat{u}_i$ zwischen den Ordinatenwerten der Datenpunkte und denen der Regressionsgeraden herausmitteln.

Nicht nur die Regressionskoeffizienten β und α, sondern auch die in A3a eingehende Varianz $\sigma^2 := \sigma_u^2$ der Störvariablen lässt sich anhand der Beobachtungsdaten schätzen. Man verwendet hierfür die Summe der quadrierten Residuen $\widehat{u}_i^2$, die man noch durch $n - 2$ dividiert, weil diese Korrektur zu einer erwartungstreuen Schätzung führt. Man erhält mit (16.4) sowie mit $\widehat{\beta}$ und $\widehat{\alpha}$ aus (16.6) und (16.7)

KQ-Schätzung der Varianz der Störvariablen

$$\widehat{\sigma^2} = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \widehat{u}_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2. \tag{16.8}$$

Beispiel 16.1: 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

[5] Bezüglich der Herleitung der beiden KQ-Schätzformeln sei auf Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 3.6.2) verwiesen.

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 $\overline{x} = 30$ und $\overline{y} = 4$. Es wird angenommen, dass der Modellansatz (16.1) 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	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
1	-20	400	-2	40
2	0	0	-1	0
3	20	400	3	60
	Summe:	800		100

Tab. 16.2: *Arbeitstabelle für die manuelle KQ-Berechnung*

Für die KQ-Schätzung $\widehat{\beta}$ von β folgt dann wegen $s_{xy} = \frac{100}{3}$ und $s_x^2 = \frac{800}{3}$ gemäß (16.6) zunächst $\widehat{\beta} = 0{,}125$ und hieraus mit $\overline{x} = 30$ und $\overline{y} = 4$ der nach (16.7) berechnete KQ-Schätzwert $\widehat{\alpha} = 0{,}25$. Man verifiziert durch Einsetzen von $\overline{x} = 30$, dass die nach der KQ-Methode geschätzte Regressionsgerade $\widehat{y} = 0{,}25 + 0{,}125x$ durch den Schwerpunkt $(\overline{x};\overline{y}) = (30;4)$ des Datensatzes verläuft.

Interpretation der Ergebnisse

Der Schätzwert $\widehat{\beta} = 0{,}125$ 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. Das Modell kann somit auch für Vorhersagen eingesetzt werden. Bei einem Rechnungsbetrag in Höhe von z. B. $x = 16$ wäre der prognostizierte Wert für das Trinkgeld durch $\widehat{y} = 0{,}25 + 0{,}125x = 2{,}25$ gegeben. Der Schätzwert $\widehat{\alpha} = 0{,}25$ ist formal der Wert, den das Modell für $x = 0$ liefert. Man erkennt, dass das Modell, dessen Parameter auf der Basis von x-Werten zwischen $x = 10$ und $x = 50$ geschätzt wurden, nicht mehr zwangsläufig weit außerhalb des Stützbereichs anwendbar sein muss. Für einen Rechnungsbetrag in Höhe von $x = 0$ würde sich hier ein Trinkgeld in Höhe von 0,25 Euro errechnen. Die Erfahrung lehrt jedoch, dass bei Nicht-Konsum in einem Restaurant i. d. R. auch kein Trinkgeld anfällt.

Aufgabe 16.1

Bei größeren Datensätzen wird man einen Taschenrechner oder eine geeignete Statistik-Software heranziehen - z. B. SPSS, STATISTICA, JMP von SAS, STATA, EViews oder *R*, und sollte dann natürlich dieselben Ergebnisse erhalten. In Abbildung 16.3 sind ein SPSS- und darunter ein EViews-Screenshot für dieses Beispiel wiedergegeben. Die oben berechneten KQ-Schätzwerte $\widehat{\beta}$ und $\widehat{\alpha}$ 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		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	,250	1,479		,169	,893
	Rechnungsbetrag	,125	,043	,945	2,887	,212

Equation: UNTITLED Workfile: KQ-SCHÄTZUNG (TRINK...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: TRINKGELD
Method: Least Squares
Date: 04/25/07 Time: 15:29
Sample: 1 3
Included observations: 3

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.250000	1.479020	0.169031	0.8934
RECHNUNGSBETRAG	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. 16.3: *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 (16.6) und (16.7) 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.

Deskriptive vs. induktive Regressionsanalyse

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 $\widehat{\beta}$, $\widehat{\alpha}$ und $\widehat{\sigma^2}$, aus denen sich nach (16.6) - (16.8) Schätzwerte aus den Daten errechnen, lässt sich mit den hier getroffenen Annahmen ableiten, dass sie erwartungstreu sind, d. h. es gilt

Eigenschaften der KQ-Schätzer

$$E(\widehat{\beta}) = \beta; \qquad E(\widehat{\alpha}) = \alpha; \qquad E(\widehat{\sigma^2}) = \sigma^2. \tag{16.9}$$

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.

Eigenschaften bei Normalverteilung

Setzt man für die Störvariablen nach A3 Normalverteilung voraus und bezeichnet die Varianzen $V(\widehat{\beta})$ und $V(\widehat{\alpha})$ der Schätzer $\widehat{\beta}$ resp. $\widehat{\alpha}$ mit $\sigma^2_{\widehat{\beta}}$ und $\sigma^2_{\widehat{\alpha}}$, so gelten für die beiden Schätzer die Normalverteilungsaussagen

$$\widehat{\beta} \sim N(\beta; \sigma^2_{\widehat{\beta}}) \tag{16.10}$$

$$\widehat{\alpha} \sim N(\alpha; \sigma^2_{\widehat{\alpha}}). \tag{16.11}$$

Die Formeldarstellungen für die Varianzen seien noch der Vollständigkeit halber und ebenfalls ohne Beweis angeführt (s. hierzu z. B. Fahrmeir / Künstler / Pigeot / Tutz (2010, Abschnitt 12.1)). Es gilt mit der nun mit s_x^2 bezeichneten unkorrigierten empirischen Varianz aus (5.6)

$$\sigma^2_{\widehat{\beta}} = \frac{1}{n \cdot s_x^2} \cdot \sigma^2 \tag{16.12}$$

$$\sigma^2_{\widehat{\alpha}} = \frac{\overline{x^2}}{n \cdot s_x^2} \cdot \sigma^2, \tag{16.13}$$

wobei wieder σ^2 die Varianz der Störvariablen U_i aus (16.2) bezeichnet.

16.3 Das Bestimmtheitsmaß

Hat man eine Regressionsgerade anhand eines Datensatzes $(x_1; y_1)$, $\ldots, (x_n; y_n)$ 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 vor und zerlegt die Gesamtvarianz s_y^2 der abhängigen Variablen in zwei Komponenten, nämlich in die durch den Regressionsansatz erklärte Varianz $s^2_{\widehat{y}}$ und eine durch den Ansatz nicht erklärte Restvarianz $s^2_{\widehat{u}}$. Alle drei empirischen Varianzen sind gemäß (5.6) definiert, also

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2}_{s^2_{\widehat{y}}} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{u}_i - \overline{\widehat{u}})^2}_{s^2_{\widehat{u}}}.$$

Da die Summe der n Residuen Null ist, kann man im letzten Summenterm $\overline{\widehat{u}} = 0$ setzen und im mittleren Summenterm $\overline{\widehat{y}} = \overline{y}$. Setzt man noch anstelle von $\widehat{u}_i$ den äquivalenten Term $y_i - \widehat{y}_i$ ein, folgt

$$\underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}_{s_y^2} = \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{s^2_{\widehat{y}}} + \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{s^2_{\widehat{u}}}. \tag{16.14}$$

Messung der Anpassungsgüte

Als Maß für die Anpassungsgüte eines bivariaten Datensatzes an eine Regressionsgerade wird dann das **Bestimmtheitsmaß** R^2 verwendet. Dieses auch gelegentlich als **Determinationskoeffizient** bezeichnete Maß vergleicht den durch die lineare Regression erklärten Varianzanteil $s^2_{\widehat{y}}$ mit der Gesamtvariation s^2_y der endogenen Variablen. Das Bestimmtheitsmaß ist also gegeben durch

$$R^2 = \frac{s^2_{\widehat{y}}}{s^2_y} = 1 - \frac{s^2_{\widehat{u}}}{s^2_y}. \tag{16.15}$$

Diese Gleichung lässt sich noch in eine weniger technisch aussehende, kürzere Form bringen. Wenn man Gleichung (16.14) mit n erweitert, erhält man eine Zerlegung der Streuung in drei Summen, die mit SQ abgekürzt (Summe von Abweichungs$q$uadraten) und mit einem aussagekräftigen Index versehen werden: [6]

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SQ_{Total}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{SQ_{Regression}} + \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{SQ_{Residual}}. \tag{16.16}$$

Anstelle von (16.15) hat man dann die äquivalente Darstellung

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}. \tag{16.17}$$

Aus der Nicht-Negativität aller Komponenten der Zerlegungen (16.14) und (16.16) folgt, dass R^2 zwischen Null und Eins liegt.

Abbildung 16.4 zeigt zwei Datensätze des Umfangs $n = 20$, 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%).

[6] In der Literatur findet man häufig auch die Abkürzungen SQT (engl: *s*um of s*q*uares *t*otal), SQE (*s*um of s*q*uares *e*xplained) und SQR (*s*um of s*q*uares *r*esiduals).

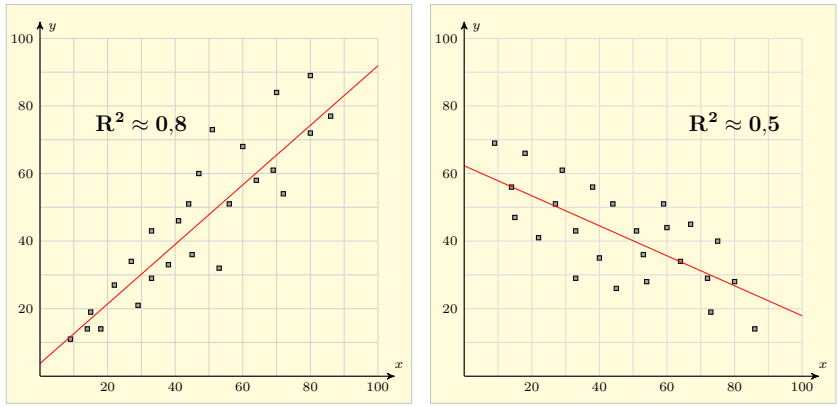


Abb. 16.4: *Datensatz mit KQ-Schätzgeraden und Bestimmtheitsmaß* R^2

Interaktives Lernobjekt „KQ-Schätzung und Bestimmtheitsmaß“

Wenn $R^2 = 1$ gilt, ist das Modell perfekt ($s^2_{\widehat{u}} = 0$). Im Falle $R^2 = 0$ liefert das lineare Modell keinen Erklärungsbeitrag ($s^2_{\widehat{y}} = 0$), was aber keinesfalls ausschließt, dass zwischen den Variablen X und Y ein nichtlinearer Zusammenhang besteht.

Für die praktische Berechnung von R^2 bietet sich anstelle von (16.15) bzw. (16.17) eine Formel an, die direkt von den Daten ausgeht. Setzt man in (16.14) beim mittleren Summenterm für $\widehat{y}_i$ den Term $\widehat{\alpha} + \widehat{\beta} \cdot x_i$ und für $\overline{y}$ den Term $\widehat{\alpha} + \widehat{\beta} \cdot \overline{x}$ ein, erhält man für die Varianzkomponente $s^2_{\widehat{y}}$ die Darstellung

$$s^2_{\widehat{y}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \Big[(\widehat{\alpha} + \widehat{\beta} x_i) - (\widehat{\alpha} + \widehat{\beta} \cdot \overline{x}) \Big]^2 = \widehat{\beta}^2 s_x^2 .$$

Hieraus folgt

$$R^2 = \frac{\widehat{\beta} s_{xy}}{s_y^2} = \frac{(s_{xy})^2}{s_x^2 s_y^2} = r^2 . \tag{16.18}$$

Im einfachen Regressionsmodell stimmt demnach das Bestimmtheitsmaß R^2 mit dem Quadrat des in (9.10) eingeführten empirischen Korrelationskoeffizienten r nach Bravais-Pearson überein.

Beispiel 16.2: Berechnung des Bestimmtheitsmaßes

Für den Datensatz aus Beispiel 16.1 (Trinkgeldbeträge von drei Restaurantbesuchern) wurden in Tabelle 16.2 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 16.2 verifiziert man leicht, dass $s_y^2 = \frac{14}{3}$ ist. Mit (16.18) erhält man hieraus für das Bestimmtheitsmaß

den Wert

$$R^2 = \frac{(\frac{100}{3})^2}{\frac{800}{3} \cdot \frac{14}{3}} = \frac{25}{28} \approx 0{,}893.$$

Dieser Wert ist in Abbildung 16.3 mit etwas größerer Genauigkeit ausgewiesen.

Aufgabe 16.2

16.4 Das multiple lineare Regressionsmodell

Eine Verallgemeinerung des Modellansatzes (16.1) mit nur *einer* erklärenden Variablen ist das **multiple Regressionsmodell**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i \qquad i = 1, \ldots, n \qquad (16.19)$$

mit k erklärenden Variablen. Man erkennt, dass (16.19) für $k = 1$ in das einfache lineare Regressionsmodell (16.1) übergeht, wenn man dort $\alpha =: \beta_0$ und $\beta =: \beta_1$ setzt.

Mit (16.19) 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.[7] Dazu fasst man, analog zu (18.1), die n Werte der abhängigen Variablen und auch die n Werte der Störvariablen zu Spaltenvektoren $\mathbf{y}$ resp. $\mathbf{u}$ zusammen:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = (y_1, y_2, \ldots, y_n)' \qquad (16.20)$$

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = (u_1, u_2, \ldots, u_n)'. \qquad (16.21)$$

Auch die in (16.19) auftretenden $k + 1$ Koeffizienten β_0, $\beta_1, \beta_2, \ldots, \beta_k$ lassen sich zu einem Vektor zusammenfassen:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'. \qquad (16.22)$$

[7] Grundzüge der Vektor- und Matrixrechnung sind in Kapitel 18 wiedergegeben.

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 (16.19) 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:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}. \tag{16.23}$$

Mit den Vektoren (16.20) - (16.22) und der Matrix (16.23) kann man die n Gleichungen (16.19) des multiplen linearen Regressionsmodells ausführlich in der Form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \tag{16.24}$$

schreiben oder kürzer als

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{16.25}$$

Modellannahmen

Fasst man die Störterme u_i aus (16.18) wieder als Realisationen von Zufallsvariablen U_i auf, so sind auch hier die Werte y_i der abhängigen Variablen Realisationen stochastischer Größen Y_i. Spezifiziert man, wie in der Praxis üblich, noch die Werte x_{i1}, x_{i2}, ... , x_{ik} der unabhängigen Variablen als nicht-stochastisch, lässt sich das multiple lineare Regressionsmodell (16.19) wie folgt schreiben:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + U_i \qquad i = 1, \ldots, n. \tag{16.26}$$

Auch diese Gleichung lässt sich durch (16.25) 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. [8] Das Modell ist durch die folgenden Annahmen charakterisiert, die allerdings – wie auch beim

[8] Es werden dann z. B. sowohl der Vektor $(u_1, u_2, ..., u_n)'$ aus (16.21) wie auch der Vektor $(U_1, U_2, ..., U_n)'$ der in (16.26) eingehenden Zufallsvariablen mit $\mathbf{u}$ abgekürzt. Würde man Zufallsvektoren mit Großbuchstaben kennzeichnen, hätte dies den Nachteil, dass sie fälschlich als Matrizen interpretiert werden könnten.

einfachen Regressionsmodell – 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 k Regressoren $X_1, X_2, \ldots, X_k$ und der abhängigen Variablen Y vermittelnde lineare Funktion ist fest, d. h. die $k+1$ Parameter β_0, $\beta_1, \beta_2, \ldots, \beta_k$, die die lineare Funktion festlegen, 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 $U_1, \ldots, U_n$, sind unabhängig identisch $N(0; \sigma^2)$-verteilt.

Annahmen bezüglich der unabhängigen Modellvariablen:

MA4: Die Werte der k unabhängigen Variablen $X_1, X_2, \ldots, X_k$ 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*).

Wenn die Elemente der Matrix $\mathbf{X}$ in (16.25) 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. Aus Annahme (MA3a) folgt z. B. dass für den Erwartungswert $E(\mathbf{u})$ von $\mathbf{u}$ die Gleichung

$$E(\mathbf{u}) = \mathbf{0} \tag{16.27}$$

gilt. Dabei bezeichnet **0** den Nullvektor, dessen Elemente nur Nullen sind. Für die im Matrizenanhang durch (18.12) und (18.13) definierte Kovarianzmatrix $V(\mathbf{u})$ des Zufallsvektors $\mathbf{u}$ erhält man mit (MA3a) und (M3b) die Darstellung

$$V(\mathbf{u}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \cdot \mathbf{I}_n. \tag{16.28}$$

Für $k = 1$ und mit $\beta_1 =: \beta$ sowie $x_1 =: x$ geht nicht nur (16.19) in (16.1) über, sondern natürlich auch (16.25). Dies gilt, weil (16.22) im Falle $k = 1$ nur aus den ersten beiden Elementen β_0 und β_1 und die Matrix (16.23) nur aus den ersten beiden Spalten besteht:

Spezialfall $k = 1$

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} &= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \\ &= \begin{pmatrix} \alpha + \beta x_1 \\ \alpha + \beta x_2 \\ \vdots \\ \alpha + \beta x_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} \alpha + \beta x_1 + u_1 \\ \alpha + \beta x_2 + u_2 \\ \vdots \\ \alpha + \beta x_n + u_n \end{pmatrix}. \end{aligned} \tag{16.29}$$

Dies ist genau das Gleichungssystem (16.1). Für $k = 2$ ist (16.29) in naheliegender Weise zu modifizieren – der Vektor $\boldsymbol{\beta}$ hat dann drei Elemente β_0, β_1, β_2 und die Matrix $\mathbf{X}$ drei Spalten.

16.5 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), ... , (x_n, y_n) in der Ebene repräsentiert sind, sind sie nun durch Punkte $(x_{11}, \dots, x_{1k}; y_1)$, ... , $(x_{n1}, \dots, x_{nk}; y_n)$ 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* ($k > 2$). 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 $\widehat{u}_i^{\,2}$ bezüglich der Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_k$ minimal ist. Die **Residuen** für eine beliebige Regressionshyperebene sind analog zu (16.4) durch

Prinzip der KQ-Schätzung

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik} \qquad i = 1, \ldots, n. \tag{16.30}$$

definiert.Abbildung 16.5 visualisiert die Residuen für einen Datensatz, nun aber für Datenpunkte im dreidimensionalen Raum ($k = 2$).

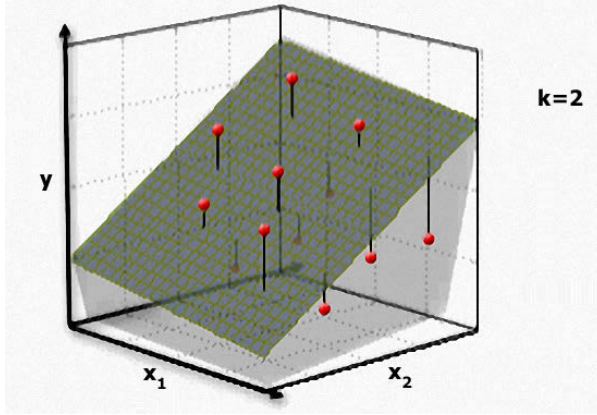


Abb. 16.5: *Veranschaulichung der Residuen und der KQ-Schätzung im Modell mit zwei erklärenden Variablen*

Die (16.5) entsprechende Minimierungsaufgabe lautet also:

$$\sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik})^2 \to Min. \tag{16.31}$$

Die nach der KQ-Methode optimale **Regressionshyperebene** ist durch den Vektor

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)' \tag{16.32}$$

definiert. Fasst man die n Residuen aus (16.30) zum **Residuenvektor**

$$\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{16.33}$$

zusammen, kann man (16.31) äquivalent als

$$\sum_{i=1}^{n} \widehat{u}_i^{\,2} = \widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \to Min \tag{16.34}$$

schreiben, wobei die Minimierung bezüglich aller denkbaren Vektoren $\widehat{\boldsymbol{\beta}}$ von Regressionskoeffizienten erfolgt. Zur Lösung dieses Minimierungsproblems wird nach $\widehat{\boldsymbol{\beta}}$ differenziert, Null gesetzt und nach $\widehat{\boldsymbol{\beta}}$ aufgelöst. [9] Dies führt zur Darstellung

Aufgabe 16.3

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{16.35}$$

für die gesuchte KQ-Schätzung von $\boldsymbol{\beta}$. Der Vektor $\widehat{\boldsymbol{\beta}}$ minimiert (16.34). Die Invertierbarkeit der Matrix $\mathbf{X}'\mathbf{X}$ ist durch die Annahme (MA5) des multiplen Regressionsmodells gesichert.

Spezialfall $k = 1$

Im Spezialfall $k = 1$ und mit den Setzungen $x_1 =: x$ sowie $\widehat{\beta}_0 =: \alpha$ und $\widehat{\beta}_1 =: \beta$ hat $\mathbf{X}$ die in (16.29) schon aufgeführte spezielle Gestalt einer $(n \times 2)$-Matrix und der Vektor $\widehat{\boldsymbol{\beta}}$ geht über in den zweielementigen Spaltenvektor $\widehat{\boldsymbol{\beta}} = (\widehat{\alpha}, \widehat{\beta})'$. Wenn man diese spezielle Ausprägungen für $\mathbf{X}$ und $\widehat{\boldsymbol{\beta}}$ in (16.35) einsetzt, resultieren nach elementaren Umformungen für $\widehat{\beta}$ und $\widehat{\alpha}$ die KQ-Schätzformeln (16.6) und (16.7).

KQ-Schätzung der Varianz der Störvariablen

Die KQ-Residuen (16.31) werden wie im einfachen Regressionsmodell auch für die Schätzung der Varianz der Störvariablen U_i herangezogen. Man verwendet wieder die Summe der quadrierten Residuen $\widehat{u}_i^2$, die man nun noch durch $n - (k + 1)$ dividiert, um eine unverzerrte Schätzung zu erhalten. Man erhält in Verallgemeinerung von (16.8)

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n-k-1} \cdot \sum_{i=1}^{n} \widehat{u}_i^2 \\ &= \frac{1}{n-k-1} \cdot \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \ldots - \widehat{\beta}_k x_{ik})^2. \end{aligned} \tag{16.36}$$

[9] Eine ausführliche Herleitung findet man z. B. bei Toutenburg / Heumann (2008, Abschnitt 9.3.1)

Dabei sind $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ die Elemente des KQ-Schätzvektors aus (16.35).

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 (16.15) bzw. (16.17) 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.

Exkurs 16.1: Binäre Regressionsmodelle

Bei dem in diesem Kapitel vorgestellten Regressionsmodell, das einen linearen Zusammenhang zwischen einer oder mehreren erklärenden Variablen und einer erklärten Variable Y herstellt, ist die Responsevariable Y ein *quantitatives* stetiges Merkmal, kann also alle Werte auf der Achse der reellen Zahlen annehmen. In der Praxis ist man aber oft auch daran interessiert, einen Zusammenhang zwischen mehreren erklärenden Variablen und einem *qualitativen* diskreten Merkmal zu modellieren, also einem Merkmal Y, dessen Ausprägungen Kategorien sind. Der einfachste Fall ist der, dass der Regressand Y nur *zwei* Ausprägungen hat, also den Charakter einer Binärvariablen aufweist. Als Beispiel genannt seien die Merkmale „Beschäftigtenstatus“ mit den beiden Kategorien „arbeitslos“ und „nicht-arbeitslos“ angeführt oder „Erfolg“ mit den Ausprägungen „ja / tritt ein“ und „nein / tritt nicht ein“. Codiert man die Ausprägungen zu „1“ und „0“ um und bezeichnet man 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.

Wenn man – analog zu (16.26) – die Responsevariable Y für jeden Beobachtungsindex i durch eine binäre Zufallsvariable Y_i und den beobachteten Wert mit y_i bezeichnet ($i = 1, 2, \ldots, n$), erkennt man, dass man den Modellansatz (16.19) nicht mehr unmittelbar verwenden kann. Letzteres folgt u. a. daraus, dass bei diesem Ansatz nun auf der linken Seite der Gleichung eine Variable stünde, die nur zwei Werte annehmen kann, auf der rechten Seite aber eine stetige Störvariable erscheint, deren Ausprägungen u_i sowohl positiv als auch negativ sein können. Der Erwartungswert $p_i = E(Y_i)$ der bernoulli-verteilten Variablen Y_i wäre nicht mehr durch $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$, sondern gemäß (11.15) durch $E(Y_i) = p_i$ gegeben.

Man muss daher bei binärer Responsevariablen den Modellansatz (16.19) modifizieren. Hierzu unterzieht man $p_i = P(Y_i = 1)$ einer Transformation

$$p_i = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})$$

mit einer streng monoton wachsenden Funktion $F(..)$, deren Werte im Intervall $[0;1]$ liegen. Geeignet ist z. B. die Funktion

$$F(x) = \frac{\exp x}{1 + \exp x}.$$

Der Zusammenhang zwischen dem Erwartungswert der Responsevariablen und den erklärenden Variablen wird also hier über eine Funktion $F(..)$ vermittelt.

Aus den beiden letzten Gleichungen lässt sich auch die Modelldarstellung

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

ableiten. Das beschriebene Regressionsmodell, bei dem die abhängige Variable Y nur zwei Ausprägungen hat, nennt man **binäres Responsemodell**. Bei Verwendung der speziellen transformierenden Funktion $F(x) = \frac{\exp x}{1+\exp x}$ nennt man das binäre Responsemodell auch **Logit-Modell**.

Man kann allgemeiner auch Regressionsmodelle betrachten, bei der die abhängige Variable ein qualitatives Merkmal ist, das *mehr als zwei* Ausprägungen aufweist. Ein Beispiel ist das Merkmal „Parteipräferenz“ mit den Ausprägungen „Parteien des aktuellen Politbarometers“. Ein Beispiel aus der Medizin für eine Responsevariable mit drei Ausprägungen ist das Merkmal „Gesundheitsstatus“ mit den Realisationen „Erkrankung vom Typ 1“, „Erkrankung vom Typ 2“ und „gesund“. Man spricht dann von einem **kategorialen Regressionsmodell**. Eine ausführliche Behandlung kategorialer Regressionsmodelle einschließlich des Sonderfalls „Binäres Responsemodell“ findet man bei Fahrmeir / Kneip / Lang (2009, Kapitel 4-5).

17 Grundzüge der Varianzanalyse

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.

Vorschau auf das Kapitel

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 15.5 wurde der **Zweistichproben-t-Test** vorgestellt. Mit diesem lassen sich für zwei normalverteilte Stichproben die in (15.26) 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.

Sir Ronald A. Fisher

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,

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_17

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 Datengewinung eingesetzt werden.

Grundbegriffe der Varianzanalyse

Die Varianzanalyse geht wie das lineare Regressionsmodell (16.2) oder (16.26) von einem linearen Zusammenhang zwischen einer Einflussgröße X oder mehreren Einflussgrößen $X_1, X_2, ..., X_k$ 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 **mehrfaktorieller 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: *an*alysis *o*f *va*riance) 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: *m*ultivariate *an*alysis *o*f *va*riance) belegt.

Beispiel 17.1: 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.

17.1 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 $(i = 1, 2, \ldots, s)$ 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 (Unabhängigkeitsannahme). Die Unabhängigkeitsannahme 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 17.1 verdeutlicht das Design und die Basisannahmen:

Grundgesamtheit $Y \sim N(\mu;\sigma^2)$		**Ziehung von Zufallsstichproben** (Gesamtumfang aller Stichproben: n)
Teilgesamtheit 1	$\rightarrow$	Stichprobe 1; Umfang n_1: $Y \sim N(\mu_1;\sigma^2)$ mit $\mu_1 = \mu + \alpha_1$
Teilgesamtheit 2	$\rightarrow$	Stichprobe 2; Umfang n_2: $Y \sim N(\mu_2;\sigma^2)$ mit $\mu_2 = \mu + \alpha_2$
$\vdots$		$\vdots$
Teilgesamtheit s	$\rightarrow$	Stichprobe s; Umfang n_s: $Y \sim N(\mu_s;\sigma^2)$ mit $\mu_s = \mu + \alpha_s$

Tab. 17.1: *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

$$Y_{ik} = \mu_i + U_{ik} \qquad i = 1, \ldots, s; \quad k = 1, \ldots, n_i \tag{17.1}$$

schreiben mit $E(U_{ik}) = 0$ und $n_1 + n_2 + \ldots n_s = n$. [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 (17.1) über in

$$Y_{ik} = \mu + \alpha_i + U_{ik} \qquad i = 1, \ldots, s; \quad k = 1, \ldots, n_i. \tag{17.2}$$

[1] In 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. Modelle der Varianzanalyse mit Messwiederholung behandeln u. a. Rasch / Kubinger (2006, Abschnitt 12.4).

[2] Sind die Stichprobenumfänge n_i alle gleich groß, spricht man von einem **varianzanalytischen Modell mit balanciertem Design**.

Dabei ist $n_1 \cdot \alpha_1 + n_2 \cdot \alpha_2 + \ldots n_s \cdot \alpha_s = 0$, weil sich die Effekte der Faktorstufen im Mittel ausgleichen. Die Modellvariante (17.2) 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.

Was leistet die Varianzanalyse?

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 $\mu_1, \mu_2, \ldots, \mu_s$ 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 (17.1)

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \ldots = \mu_s \quad \text{gegen} \\ &H_1 : \mu_i \neq \mu_j \text{ für mind. ein } (i,j) \end{aligned} \tag{17.3}$$

und analog bei Zugrundelegung des Modells (17.2)

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_s = 0 \text{ gegen} \\ &H_1 : \alpha_i \neq 0 \text{ und } \alpha_j \neq 0 \text{ für mind. ein } (i,j). \end{aligned} \tag{17.4}$$

17.2 Durchführung einer einfaktoriellen Varianzanalyse

Um die Hypothesen (17.3) resp. (17.4) zu testen, benötigt man die Daten der s Zufallsstichproben. Diese kann man übersichtlich in tabellarischer Form zusammenstellen. In Tabelle 17.2 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 $y_{i\cdot}$ sowie der Mittelwert $\overline{y}_{i\cdot}$ der Elemente der i-ten Stichprobe wiedergegeben $(i = 1, \ldots, s)$.

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 (16.16) in zwei Komponenten zerlegen lässt, nämlich in eine Komponente SQ_{zwischen}, die

		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}	$y_{1\cdot}$	$\overline{y}_{1\cdot}$
	2	y_{21}	y_{22}	...	y_{2k}	...	y_{2,n_2}	$y_{2\cdot}$	$\overline{y}_{2\cdot}$
	$\vdots$	$\vdots$		$\ddots$			$\vdots$	$\vdots$	$\vdots$
	i	y_{i1}	y_{i2}	...	y_{ik}	...	y_{i,n_i}	$y_{i\cdot}$	$\overline{y}_{i\cdot}$
	$\vdots$	$\vdots$				$\ddots$	$\vdots$	$\vdots$	$\vdots$
	s	y_{s1}	y_{s2}	...	y_{sk}	...	y_{s,n_s}	$y_{s\cdot}$	$\overline{y}_{s\cdot}$

Tab. 17.2: *Daten bei einer einfaktoriellen Varianzanalyse*

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 17.2 lässt sich anhand der Summe

$$SQ_{\text{Total}} := \sum_{i=1}^{s} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_{..})^2 \tag{17.5}$$

aller quadrierten Abweichungen der Beobachtungswerte y_{ik} vom Gesamtmittelwert $\overline{y}_{..}$ erfassen. [3] Die Quadrierung der Differenzen $y_{ik} - \overline{y}_{..}$ verhindert, dass sich positive und negative Abweichungen vom Gesamtmittel kompensieren. Der Wert $\overline{y}_{..}$ lässt sich errechnen, indem man die Summe der s Elemente der vorletzten Spalte von Tabelle 17.2 durch die Gesamtzahl n aller Beobachtungen dividiert.[4]

In der letzten Spalte von Tabelle 17.2 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 Stichpro-

[3] Das Kürzel SQ steht wieder für „*S*umme der Abweichungs*q*uadrate“ oder „*s*um of s*q*uares“.

[4] Alternativ kann man die Werte der letzten Spalte von Tabelle 17.2 mit den jeweiligen Stichprobenumfängen gewichten und dann die Summe der gewichteten Stichprobenmittelwerte durch n teilen (vgl. Exkurs 5.1).

ben – unter Ausblendung der Variation innerhalb der Gruppen – bietet es sich daher an, von den Abweichungen $\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}$ vom Gesamtmittelwert auszugehen, diese zu quadrieren, die Quadrate mit den jeweiligen Stichprobenumfängen zu gewichten und aufzusummieren:

$$SQ_{\text{zwischen}} := \sum_{i=1}^{s} n_i \cdot (\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2. \tag{17.6}$$

Die nicht durch die Variation von Faktorstufen erklärte Reststreuung kann für jede Stichprobe $y_{i1}, y_{i2}, \ldots, y_{i;n_i}$ durch die Abweichungen $y_{ik} - \overline{y}_{i\cdot}$ der Stichprobenelemente vom Stichprobenmittelwert $\overline{y}_{i\cdot}$ beschrieben werden. Für die s Stichproben hat man also

$$SQ_{\text{Residual}} := \sum_{i=1}^{s} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_{i\cdot})^2. \tag{17.7}$$

Mit diesen Bezeichnungen gilt die zu (16.16) analoge Streuungszerlegung

$$SQ_{\text{Total}} = SQ_{\text{zwischen}} + SQ_{\text{Residual}}. \tag{17.8}$$

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. – im Falle von SQ_{Residual} – mit $n-s$ Freiheitsgraden.

Die aus den Daten der Tabelle 17.2 errechneten Stichprobenmittelwerte $\overline{y}_{i\cdot}$ und der Gesamtmittelwert $\overline{y}_{\cdot\cdot}$ lassen sich wieder – vgl. (13.3) und (14.6) – als Realisationen von Zufallsvariablen $\overline{Y}_{i\cdot}$ resp. $\overline{Y}_{\cdot\cdot}$ auffassen und zur unverzerrten Schätzung der in Tabelle 17.1 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

$$\widehat{\mu}_i = \overline{Y}_{i\cdot}; \quad \widehat{\mu} = \overline{Y}_{\cdot\cdot}; \quad \widehat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} \tag{17.9}$$

gegeben. Auch die Varianz σ^2 des in Tabelle 17.1 veranschaulichten Modells lässt sich erwartungstreu schätzen , wobei man entweder von der Streuungskomponente SQ_{zwischen} aus (17.6) oder aber von der Komponente SQ_{Residual} aus (17.7) ausgehen kann. Im erstgenannten Fall erhält man – vgl. (13.4) und (13.9) – eine unverzerrte Schätzung, indem man die aus s Einzeltermen bestehende Summe SQ_{zwischen} nicht durch s, sondern

Erwartungstreue Schätzung der Varianz

durch $s-1$ dividiert:

$$\widehat{\sigma}^2 = \frac{1}{s-1} \cdot SQ_{\text{zwischen}}. \tag{17.10}$$

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:

$$\widehat{\sigma}^2 = \frac{1}{n-s} \cdot SQ_{\text{Residual}}. \tag{17.11}$$

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 (17.10) und (17.11), verwendet also die hier mit F bezeichnete Teststatistik

$$F := \frac{\frac{1}{s-1} \cdot SQ_{\text{zwischen}}}{\frac{1}{n-s} \cdot SQ_{\text{Residual}}} = \frac{n-s}{s-1} \cdot \frac{SQ_{\text{zwischen}}}{SQ_{\text{Residual}}}. \tag{17.12}$$

Dieser Quotient hat den Vorteil, dass er unter der Nullhypothese H_0 aus (17.3) bzw. (17.4) 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 (17.3) resp. (17.4) 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 12.7). Man führt also einen **F-Test** zum Signifikanzniveau α durch.

Beispiel 17.2: 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 (17.3), 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 (17.12) einmal ohne Softwareunterstützung ausgeführt zu haben.

Bei der Leistungsmessung am Ende der Geometrieunterrichtseinheit gab es für die drei Stufen des Faktors „Unterrichtsform“ folgende Einzelergebnisse:

		Element-Nr der Stichprobe										$\sum$	**Mittelwert**
Gruppe	1	59	48	65	38	74	43	62	42	62	58	551	55,1
	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. 17.3: *Punktzahlen beim Geometrieabschlusstest*

Aus den Daten errechnet man den Gesamtmittelwert $\overline{y}_{..}$ als Summe der drei Elemente der vorletzten Spalte, wenn man diese noch durch die Anzahl $n = 29$ der Beobachtungen teilt. Man erhält so

$$\overline{y}_{..} = \frac{551 + 574 + 630}{29} = \frac{1755}{29} \approx 60{,}517.$$

Für den Anteil SQ_{zwischen} der Streuung zwischen den drei Gruppen an der Gesamtstreuung SQ_{Total} folgt nach (17.6)

$$SQ_{\text{zwischen}} = 10 \cdot (55{,}1 - \overline{y}_{..})^2 + 10 \cdot (57{,}4 - \overline{y}_{..})^2 + 9 \cdot (70{,}0 - \overline{y}_{..})^2 \approx 1199{,}94.$$

Für die Reststreuung SQ_{Residual} errechnet man dann, z. B. unter Einsatz eines Tabellenkalkulationsprogramms, mit (17.7)

$$SQ_{\text{Residual}} = \sum_{i=1}^{3} \sum_{k=1}^{n_i} (y_{ik} - \overline{y}_{i.})^2 \approx 3153{,}30.$$

Für die Testgröße (17.12) 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 17.1 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.

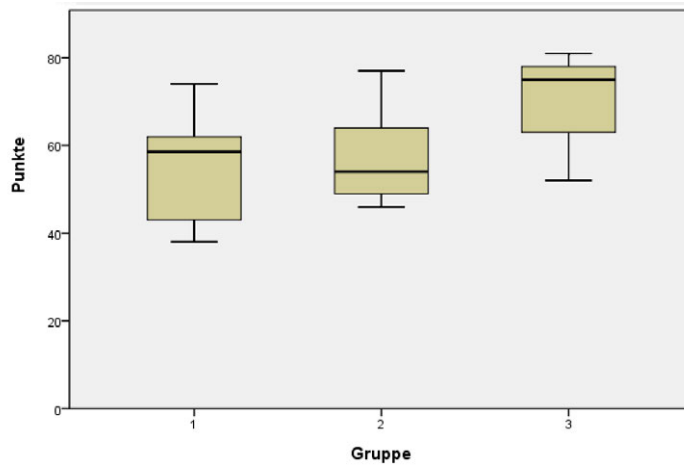


Abb. 17.1: *Boxplots für die Punktzahlen beim Geometrieabschlusstest*

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 17.1 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.

17.3 Ausblick auf die 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 (17.1) eine Darstellung, die sich auf $s \cdot r$ Faktorstufenkombinationen bezieht:

$$Y_{ijk} = \mu_{ij} + U_{ijk} \qquad \begin{array}{l} i = 1, \ldots, s; \\ j = 1, \ldots, r; \\ k = 1, \ldots, n_{ij}, \end{array} \tag{17.13}$$

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**:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + U_{ijk} \qquad \begin{array}{l} i = 1, \ldots, s; \\ j = 1, \ldots, r; \\ k = 1, \ldots, n_{ij}. \end{array} \tag{17.14}$$

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.

Zerlegung der Streuung zwischen den Gruppen

In Tabelle 17.1 ist also jeder Wert y_{ik} im grau markierten Tabelleninneren durch r Werte $y_{ij1}, y_{ij2}, \ldots, y_{ijr}$ zu ersetzen. Die Streuungszerlegung (17.8) 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 17.2 veranschaulicht dies.

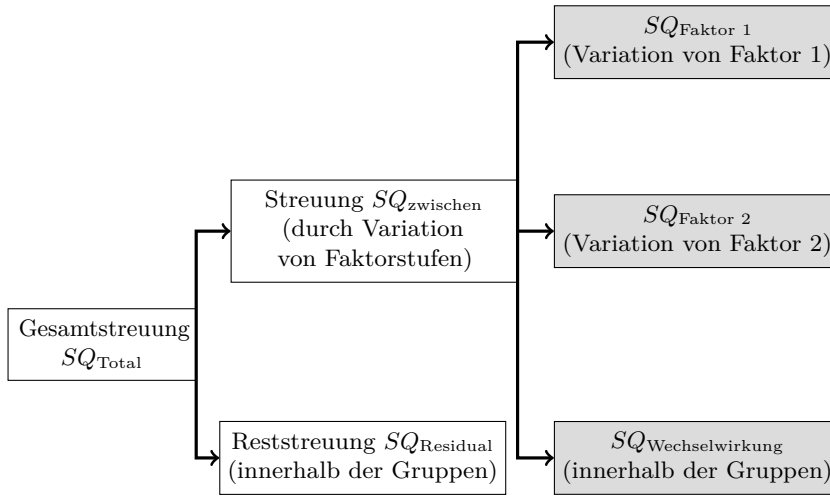


Abb. 17.2: *Streuungszerlegung bei der zweifaktoriellen Varianzanalyse (Zerlegung im einfaktoriellen Fall: ohne grau gerasterte Komponenten)*

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.

Bezüglich einer ausführlicheren Behandlung der ein- und zweifaktoriellen Varianzanalyse sei auf Fahrmeir / Künstler / Pigeot oder Schlittgen (2012, Kapitel 17) verwiesen.

Teil III

Anhänge

Lernziele zu Teil III

Der letzte Teil dieses Manuskripts enthält neben einer knapp gehaltenen Einführung in die Matrizenrechnung 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

- wissen, was Vektoren und Matrizen sind und Standardoperationen mit ihnen ausführen können;
- in der Lage sein, Quantile und Werte von Verteilungsfunktionen aus den zur Verfügung gestellten Tabellen abzulesen;
- 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;
- mit einigen mathematischen Symbolen und Schreibweisen vertraut sein, die in der Statistik häufiger verwendet werden.

18 Grundzüge der Matrizenrechnung

Bei der Behandlung multivariater Verfahren – etwa bei der Analyse des multiplen Regressionsmodells und bei der mehrfaktoriellen Varianzanalyse – kann man mit Einführung von Matrizen und Vektoren mehrere Gleichungen zu einer einzigen Gleichung zusammenfassen und damit eine besonders übersichtlichere Darstellung erreichen.

Vorschau auf das Kapitel

Da Matrizen und Vektoren nicht jedem Leser vertraut sein dürften, werden sie in diesem Kapitel zunächst definiert. Dabei werden auch wichtige Spezialfälle erwähnt, z. B. die Einheitsmatrix oder der Nullvektor. Anschließend erfolgt eine Vorstellung elementarer Operationen mit Matrizen und Vektoren, u. a. die Addition, Multiplikation und die Inversion von Matrizen. Die Operationen werden anhand von Beispielen illustriert.

18.1 Grundbegriffe

In der Mathematik und anderen Wissenschaften, u. a. in der *Physik*, der *Ökonometrie*, der *Statistik* oder auch – bei der Anwendung multivariater Verfahren – in der *Psychologie*, werden häufig Vektoren und Matrizen verwendet, um mathematische Sachverhalte kompakter und übersichtlicher darzustellen. Einen n Elemente umfassenden Satz $x_1, x_2, ..., x_n$ reeller Zahlen kann man z. B. zu einem n-Tupel zusammenfassen. Wenn man ein solches n-Tupel von reellen Zahlen vertikal anordnet, erhält man einen **Spaltenvektor**, den man in Lehrbüchern meist mit einem fett gesetzten lateinischen oder griechischen Kleinbuchstaben kennzeichnet, hier z. B. $\mathbf{x}$. Wenn man das n-Tupel horizontal anordnet, also eine Anordnung $(x_1, x_2, ..., x_n)$ verwendet, spricht man von einem **Zeilenvektor**. Die Überführung eines Spaltenvektors in einen Zeilenvektor wird auch als *Transponieren* des Vektors bezeichnet und durch eine hochgestellten Strich gekennzeichnet:

Spalten- und Zeilenvektoren

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)' = \mathbf{x}'. \tag{18.1}$$

Wenn im Folgenden von einem Vektor die Rede ist, ohne dass explizit spezifiziert wird, ob es um einen Spalten- oder Zeilenvektor geht, ist stets ein Spaltenvektor gemeint. Spezielle Vektoren sind der nur aus

Spezialfälle (Vektoren)

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_18

Nullen bestehende **Nullvektor 0** und der nur aus Einsen bestehende **Einsvektor 1**:

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Will man die Anzahl n der in einem Vektor zusammengefassten Elemente betonen, spricht man genauer von einem n-Spaltenvektor oder von einem Spaltenvektor der Dimension n. Durch einen Vektor $\mathbf{x} = (x_1; x_2; x_3)'$ der Dimension 3 ist z. B. ein Punkt im dreidimensionalen Raum definiert. Reelle Zahlen, die ja die Elemente eines Vektors konstituieren, heißen auch **Skalare**.

Bildung von Matrizen

Hat man nicht nur einen, sondern k Datensätze $\mathbf{x}_j = (x_{1j}, ..., x_{nj})'$ $(j = 1, 2, ..., k)$ des Umfangs n und stellt man die Elemente der k Spaltenvektoren nebeneinander, resultiert ein als **Matrix** bezeichnetes rechteckiges Schema mit Tabellenstruktur. Matrizen werden üblicherweise mit fetten lateinischen oder griechischen Großbuchstaben abgekürzt:

$$\mathbf{X} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nk} \end{pmatrix}}_{n \times k} = (x_{ij})_{i=1,..,n;j=1,...,k}. \quad (18.2)$$

Eine Matrix mit n Zeilen und k Spalten wird als (n x k)-Matrix angesprochen und man nennt n x k *Dimension* der Matrix. In Gleichung (18.2) ist die Dimension unterhalb der Matrix ausgewiesen. In (18.2) ist ferner das im Schnittpunkt der i-ten Zeile und der j-ten Spalte der Matrix stehende Element x_{ij} durch Rasterung betont. Man nennt dieses Element auch das Matrixelement in der Position (i, j). Anstelle von (18.2) schreibt man kürzer $\mathbf{X} = (x_{ij})$, wenn sich der Laufbereich der Indizes i (Anzahl der Zeilen) und j (Anzahl der Spalten) aus dem Kontext erschließt. Im Unterschied zu einer beliebigen Tabelle, etwa einer Zusammenstellung von Adressen, können mit den Elementen einer Matrix einfache Rechenoperationen durchgeführt werden (Addition, Subtraktion, Multiplikation).

Vektoren sind Spezialfälle von Matrizen – ein Zeilenvektor lässt sich als Matrix mit nur einer Zeile und ein Spaltenvektor als Matrix mit nur einer Spalte interpretieren. Eine Matrix, deren Elemente alle Nullen sind, heißt

Nullmatrix. Ein weiterer Spezialfall ist der Fall, dass bei einer Matrix die Anzahl der Zeilen und Spalten übereinstimmen ($n = k$). In diesem Falle liegt eine *quadratische* Matrix vor. Ist $\mathbf{X}$ eine quadratische Matrix, so ist deren Dimension schon durch Angabe entweder der Zeilen- oder der Spaltenanzahl eindeutig bestimmt. Manchmal wird die Dimension einer quadratischen Matrix über einen tiefgestellten Index ausgewiesen, etwa $\mathbf{X}_n$ bei einer quadratischen Matrix $\mathbf{X}$ mit n Zeilen.

Spezialfälle (Matrizen)

Sind bei einer quadratischen Matrix alle Elemente x_{ij} mit $i \neq j$ Null, spricht man von einer **Diagonalmatrix**. Die Elemente $x_{11}, x_{22}, ..., x_{nn}$ konstituieren die **Hauptdiagonale** einer quadratischen Matrix. Ein Sonderfall einer Diagonalmatrix ist die i. Allg. mit $\mathbf{I}$ oder – bei Ausweis der Dimension – mit $\mathbf{I}_n$ abgekürzte **Einheitsmatrix**. Für diese gilt, dass die Elemente auf der Hauptdiagonalen alle den Wert 1 haben:

$$\mathbf{I} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}}_{n \times n}. \tag{18.3}$$

18.2 Operationen mit Matrizen und Vektoren

Wie Vektoren lassen sich auch Matrizen transponieren. Die zur Matrix $\mathbf{X}$ aus (18.2) gehörende *transponierte* Matrix $\mathbf{X}'$ entsteht durch Vertauschen der Zeilen und Spalten von $\mathbf{X}$:

$$\mathbf{X} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}}_{n \times k} \xrightarrow{\text{Transponieren}}$$

$$\mathbf{X}' = \underbrace{\begin{pmatrix} x_{11} & x_{21} & \dots & x_{i1} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{i2} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{ik} & \dots & x_{nk} \end{pmatrix}}_{k \times n}.$$

Eine Matrix $\mathbf{X}$ mit der Eigenschaft

$$\mathbf{X} = \mathbf{X}' \tag{18.4}$$

heißt *symmetrisch*. Man erkennt leicht, dass nur quadratische Matrizen symmetrisch sein können. Ein Beispiel einer symmetrischen Matrix ist die Einheitsmatrix $\mathbf{I}$.

Eine besonders einfache Operation ist die Multiplikation einer Matrix mit einer reellen Zahl λ (lies: *lambda*). Diese erfolgt, indem man jedes Element einer Matrix $\mathbf{X} = (x_{ij})$ einzeln mit dem Skalar λ multipliziert:

$$\lambda \cdot \mathbf{X} = \lambda \cdot (x_{ij}) = (\lambda \cdot x_{ij}). \tag{18.5}$$

Addition von Matrizen

Einfach ist auch die Addition von Matrizen $\mathbf{A} = (a_{ij})$ und $\mathbf{B} = (b_{ij})$ gleicher Dimension. Hier werden die an gleicher Position stehenden Elemente addiert, d. h. es ist

$$\mathbf{A} + \mathbf{B} = \mathbf{C} = (c_{ij}) \qquad \text{mit} \qquad c_{ij} = a_{ij} + b_{ij}. \tag{18.6}$$

Für Matrizen ungleicher Dimension ist die Addition nicht erklärt. Da Vektoren als spezielle Matrizen zu interpretieren sind, gelten die vorstehenden Ausführungen insbesondere auch für Vektoren.

Beispiel 18.1: Multiplikation mit Skalaren; Addition von Matrizen

Nachstehend wird illustriert, wie man gemäß (18.5) eine (2×3)-Matrix mit 4,5 resp. einen aus zwei Elementen bestehenden Vektor mit der reellen Zahl -3 multipliziert:

$$4{,}5 \cdot \begin{pmatrix} 1 & 1{,}8 & 4 \\ 3 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 4{,}5 \cdot 1 & 4{,}5 \cdot 1{,}8 & 4{,}5 \cdot 4 \\ 4{,}5 \cdot 3 & 4{,}5 \cdot 0 & 4{,}5 \cdot 2 \end{pmatrix} = \begin{pmatrix} 4{,}5 & 8{,}1 & 18 \\ 13{,}5 & 0 & 9 \end{pmatrix}$$

$$(-3) \cdot \begin{pmatrix} 2x \\ -y \end{pmatrix} = \begin{pmatrix} (-3) \cdot 2x \\ (-3) \cdot (-y) \end{pmatrix} = \begin{pmatrix} -6x \\ 3y \end{pmatrix}.$$

Beispielhaft ausgeführt sei auch die Anwendung von (18.6) anhand der Addition zweier (2×3)-Matrizen sowie der Addition zweier quadratischer Matrizen mit zwei Zeilen. Die Elemente der letztgenannten Matrizen sind so spezifiziert, dass als Summe die (2×2)-Einheitsmatrix resultiert:

$$\begin{pmatrix} 1 & 1{,}8 & 4 \\ 3 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 1{,}2 & 2 \\ 1 & 5{,}4 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 3 & 6 \\ 4 & 5{,}4 & 5 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 4 \\ 3 & -1 \end{pmatrix} + \begin{pmatrix} -1 & -4 \\ -3 & 2 \end{pmatrix} = \begin{pmatrix} 2-1 & 4-4 \\ 3-3 & -1+2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2.$$

Auch die Multiplikation von Matrizen ist nur unter bestimmten, die Dimension der Matrizen betreffenden Vorausetzungen möglich. Das Produkt zweier Matrizen $\mathbf{A}$ und $\mathbf{B}$ ist nur dann erklärt, wenn die Anzahl der Spalten von $\mathbf{A}$ mit der Anzahl der Zeilen von $\mathbf{B}$ übereinstimmt. Hat etwa die Matrix $\mathbf{A}$ die Dimension $(n \times k)$ und $\mathbf{B}$ die Dimension $(k \times m)$, so ist die Matrix $\mathbf{C} := \mathbf{A} \cdot \mathbf{B}$ von der Dimension $(n \times m)$:

$$\underbrace{\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix}}_{n\times k} \cdot \underbrace{\begin{pmatrix} b_{11} & \dots & b_{1l} & \dots & b_{1m} \\ b_{21} & \dots & b_{2l} & \dots & b_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{k1} & \dots & b_{kl} & \dots & b_{km} \end{pmatrix}}_{k\times m}$$

$$= \underbrace{\begin{pmatrix} c_{11} & \dots & c_{1l} & \dots & c_{1m} \\ \vdots & \vdots & \vdots & \vdots & \\ c_{i1} & \dots & c_{il} & \dots & c_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{n1} & \dots & c_{nl} & \dots & c_{nm} \end{pmatrix}}_{n\times m}$$

Produkt zweier Matrizen

Das vorstehend durch Rasterung betonte Element c_{il} der $(n \times m)$-Produktmatrix $\mathbf{C}$, also das im Schnittpunkt der i-ten Zeile und der l-ten Spalte stehende Element von $\mathbf{C}$ ergibt sich, indem man die ebenfalls in der obigen Gleichung gerastert dargestellten k Elemente der i-ten Zeile von $\mathbf{A}$ $(i = 1, ..., n)$ und die k Elemente der l-ten Spalte von $\mathbf{B}$ $(l = 1, \dots, m)$ gliedweise miteinander multipliziert und aufsummiert:

$$\underbrace{\mathbf{A}}_{n\times k} = (a_{ij}), \quad \underbrace{\mathbf{B}}_{k\times m} = (b_{jl}) \quad \Rightarrow \quad \mathbf{A} \cdot \mathbf{B} = \underbrace{\mathbf{C}}_{n\times m} = (c_{il}) \tag{18.7}$$

$$\text{mit} \quad c_{il} = \sum_{j=1}^{k} a_{ij} \cdot b_{jl}.$$

Neben dem Produkt $\mathbf{A} \cdot \mathbf{B}$ aus einer $(n \times k)$-Matrix $\mathbf{A}$ und einer $(k \times m)$-Matrix $\mathbf{B}$ ist $\mathbf{B} \cdot \mathbf{A}$ nur dann ebenfalls erklärt, wenn $n = m$ gilt. Aber selbst wenn sowohl $\mathbf{A} \cdot \mathbf{B}$ als auch $\mathbf{B} \cdot \mathbf{A}$ beide definiert sind, stimmen die Produkte i. Allg. – anders als bei der Multiplikation zweier reeller Zahlen – *nicht* überein. Es gibt noch weitere Unterschiede zwischen Operationen mit reellen Zahlen einerseits und Matrizen andererseits. Ist z. B. das Produkt zweier reeller Zahlen Null, kann man stets darauf schließen, dass mindestens eine der beiden Zahlen Null ist. Bei zwei Matrizen hingegen kann der Fall auftreten, dass ihr Produkt die Nullmatrix ergibt, ohne dass eine der beiden Ausgangsmatrizen eine Nullmatrix war.

Die Multiplikation einer Matrix $\mathbf{A}$ mit einem Spaltenvektor $\mathbf{x}$ oder einem Zeilenvektor $\mathbf{x}'$ ergibt sich als Spezialfall der Multiplikation (18.7) von Matrizen. Definiert sind im Falle einer $(n \times k)$-Matrix $\mathbf{A}$ nur das Produkt $\mathbf{A} \cdot \mathbf{x}$ mit einem Spaltenvektor $\mathbf{x}$ der Dimension k und das Produkt $\mathbf{x}' \cdot \mathbf{A}$ mit einem Zeilenvektor der Dimension n. Im ersten Fall resultiert ein Spaltenvektor mit n Elementen, im zweiten Fall ein Zeilenvektor mit k Elementen. Letzterer ist hier nur aus Platzgründen als transponierter Spaltenvektor dargestellt:

$$\mathbf{A} \cdot \mathbf{x} = \underbrace{\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ik} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix}}_{n \times k} \cdot \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_k \end{pmatrix}}_{k \times 1}$$

$$= \underbrace{\begin{pmatrix} a_{11} \cdot x_1 + a_{12} \cdot x_2 + \ldots + a_{1k} \cdot x_k \\ \vdots \\ a_{i1} \cdot x_1 + a_{i2} \cdot x_2 + \ldots + a_{ik} \cdot x_k \\ \vdots \\ a_{n1} \cdot x_1 + a_{n2} \cdot x_2 + \ldots + a_{nk} \cdot x_k \end{pmatrix}}_{n \times 1}$$

$$\underbrace{(x_1, x_2, \ldots, x_n)}_{1 \times n} \cdot \underbrace{\begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1k} \\ a_{21} & \dots & a_{2j} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nk} \end{pmatrix}}_{n \times k}$$

$$= \underbrace{\begin{pmatrix} x_1 \cdot a_{11} + x_2 \cdot a_{21} + \ldots + x_n \cdot a_{n1} \\ \vdots \\ x_1 \cdot a_{1j} + x_2 \cdot a_{2j} + \ldots + x_n \cdot a_{nj} \\ \vdots \\ x_1 \cdot a_{1k} + x_2 \cdot a_{2k} + \ldots + x_n \cdot a_{nk} \end{pmatrix}'}_{1 \times k}$$

Beispiel 18.2: Multiplikation von Matrizen

Betrachtet seien ein Spaltenvektor $\mathbf{x}$ der Dimension 4 sowie eine (2×4)-Matrix $\mathbf{A}$ und eine (4×3)-Matrix $\mathbf{B}$:

$$\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 3 \end{pmatrix} \qquad \mathbf{A} = \begin{pmatrix} 3 & -2 & 6 & 1 \\ 1 & 4 & 2 & 0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 7 & 1 & -1 \\ 1 & 3 & 1 \\ -2 & 0 & 1 \\ 2 & 1 & 2 \end{pmatrix}.$$

Für $\mathbf{A} \cdot \mathbf{x}$ verifiziert man

$$\mathbf{A} \cdot \mathbf{x} = \begin{pmatrix} 3 & -2 & 6 & 1 \\ 1 & 4 & 2 & 0 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \cdot 2 - 2 \cdot 1 + 6 \cdot 0 + 1 \cdot 3 \\ 1 \cdot 2 + 4 \cdot 1 + 2 \cdot 0 + 0 \cdot 3 \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}.$$

Für das Produkt $\mathbf{A} \cdot \mathbf{B}$ der obigen Matrizen erhält man die nachstehende (2×3)-Matrix :

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 9 & -2 & 3 \\ 7 & 13 & 5 \end{pmatrix}.$$

Für die Matrizen $\mathbf{A}$ und $\mathbf{B}$ ist das Produkt $\mathbf{B} \cdot \mathbf{A}$ nicht erklärt, weil die Dimensionen der Matrizen nicht miteinander verträglich sind.

Nachstehend sind zwei andere Matrizen $\mathbf{A}$ und $\mathbf{B}$ wiedergegeben, für die sowohl $\mathbf{A} \cdot \mathbf{B}$ als auch $\mathbf{B} \cdot \mathbf{A}$ erklärt sind. Die beiden Produktterme stimmen aber hier nicht überein:

$$\mathbf{A} = \begin{pmatrix} -1 & 2 \\ 1 & 2 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} -1 \cdot 1 + 2 \cdot 3 & -1 \cdot 2 + 2 \cdot 4 \\ 1 \cdot 1 + 2 \cdot 3 & 1 \cdot 2 + 2 \cdot 4 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 7 & 10 \end{pmatrix}$$

$$\mathbf{B} \cdot \mathbf{A} = \begin{pmatrix} 1 \cdot (-1) + 2 \cdot 1 & 1 \cdot 2 + 2 \cdot 2 \\ 3 \cdot (-1) + 4 \cdot 1 & +3 \cdot 2 + 4 \cdot 2 \end{pmatrix} = \begin{pmatrix} 1 & 6 \\ 1 & 14 \end{pmatrix} \neq \mathbf{A} \cdot \mathbf{B}$$

Inversion von Matrizen

Nicht nur bei der Addition, sondern auch bei der Multiplikation zweier quadratischer Matrizen $\mathbf{A}$ und $\mathbf{B}$ kann der Fall auftreten, dass das Ergebnis der Operation die Einheitsmatrix $\mathbf{I}$ ist. Wenn eine quadratische Matrix $\mathbf{B}$ die Eigenschaft hat, dass das Produkt $\mathbf{A} \cdot \mathbf{B}$ die Einheitsmatrix ist, nennt man sie die **Inverse** zur Matrix $\mathbf{A}$ und schreibt $\mathbf{A}^{-1}$ (lies: *Inverse* der Matrix $\mathbf{A}$). Mit der Schreibweise wird angedeutet, dass es sich um eine Verallgemeinerung der Kehrwertbildung bei reellen Zahlen handelt. Für die Inverse $\mathbf{A}^{-1}$ einer quadratischen Matrix $\mathbf{A}$ ist neben $\mathbf{A} \cdot \mathbf{A}^{-1}$ stets auch $\mathbf{A}^{-1} \cdot \mathbf{A}$ erklärt und es gilt

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}. \tag{18.8}$$

Für die numerische Bestimmung der Inversen einer quadratischen Matrix empfiehlt sich bei größeren Matrizen die Heranziehung geeigneter Software. Bei einer (2×2)-Matrix und auch noch bei einer (3×3)-Matrix $\mathbf{A}$ ist die Bestimmung der Inversen im Prinzip noch per Hand möglich. Bei einer (2×2)-Matrix $\mathbf{A}$ kann man z. B. die Elemente der Inversen $\mathbf{B} := \mathbf{A}^{-1}$ über den Ansatz

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{18.9}$$

bestimmen. Der Ansatz führt auf ein lineares Gleichungssystem – hier auf ein System mit 4 Gleichungen zur Bestimmung der 4 Elemente b_{11}, b_{12}, b_{21} und b_{22}. Ein solches Gleichungssystem ist nicht immer lösbar und entsprechend ist nicht für jede quadratische Matrix $\mathbf{A}$ eine Inverse $\mathbf{A}^{-1}$ erklärt. In der Praxis kann man für die Matrizeninversion Statistiksoftware verwenden, z. B. die freie Software *R*.

Bedingung für Invertierbarkeit

Es sei ohne Beweis angeführt, dass eine quadratische Matrix genau dann invertierbar ist, wenn die Vektoren, die ihre Spalten und ihre Zeilen definieren, *linear unabhängig* sind. Letzteres bedeutet, dass sich keine Zeile oder Spalte als Linearkombination einer anderen Zeile resp. Spalte darstellen lässt. Eine solche quadratische Matrix wird als *reguläre* Matrix oder auch als Matrix *mit vollem Rang* angesprochen. Es sind also nur reguläre Matrizen invertierbar.

Beispiel 18.3: Invertierbarkeit von Matrizen

Nachstehend sind zwei (2×2)-Matrizen $\mathbf{A}$ und $\mathbf{B}$ sowie eine (3×3)-Matrix $\mathbf{C}$ wiedergegeben:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ -2 & -4 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \\ 2 & 4 & 1 \end{pmatrix}.$$

Die Spalten- und Zeilenvektoren der Matrix $\mathbf{A}$ sind jeweils linear unabhängig, d. h. die Matrix ist invertierbar. Die Inverse kann man mit gängiger Mathematik- oder Statistiksoftware, z. B. MATHEMATICA, JMP von SAS, SPSS sowie *R* oder auch manuell nach (18.9) bestimmen. Man erhält für $\mathbf{A}^{-1}$ eine Matrix, die in der ersten Zeile 1 und -1 und in der zweiten Zeile die Elemente 0 und $-0{,}5$ enthält. Dass diese Matrix wirklich die Inverse von $\mathbf{A}$ ist, lässt sich leicht überprüfen:

$$\mathbf{A} \cdot \mathbf{A^{-1}} = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 \\ 0 & 0{,}5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I_2}.$$

Die Matrix **B** is hingegen nicht invertierbar. Offenbar geht die zweite Zeile von **B** aus der ersten Zeile durch Multiplikation mit -2 hervor, d. h. die Zeilen sind hier nicht linear unabhängig.

Für die Matrix **C** kann man zeigen, dass sie – wie die Matrix **A** – regulär ist und somit auch eine Inverse besitzt. Die Bestimmung der inversen Matrix $\mathbf{C}^{-1}$ mit Papier und Bleistift ist zwar noch durchführbar, wurde aber hier mit SPSS und zusätzlich auch unter Verwendung der freien Statistiksoftware *R* ausgeführt. Der Programmcode bei Verwendung von *R* ist sehr kurz:

```
MATRIX.
COMPUTE  a={1,2,0;0,1,3;2,4,1}.
COMPUTE  c={1,0,0;0,1,0;0,0,1}.
COMPUTE  x=INV(a)*c.
PRINT  x.
END  MATRIX.

[DatenSet0]

Run  MATRIX  procedure:

X
 -11   -2    6
   6    1   -3
  -2    0    1
```

```
R> A <- matrix(c(1, 2, 0, 0, 1, 3, 2, 4, 1), nrow = 3, byrow = TRUE)
R> solve(A)

     [,1] [,2] [,3]
[1,]  -11   -2    6
[2,]    6    1   -3
[3,]   -2    0    1
```

Abb. 18.1: *Computerausdruck zur Bestimmung der Inversen einer (3×3)-Matrix) (links: SPSS; rechts unten: R)*

Dass die ermittelte Matrix wirklich die Inverse von **C** ist, lässt sich unschwer verifizieren:

$$\mathbf{C} \cdot \mathbf{C^{-1}} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \\ 2 & 4 & 1 \end{pmatrix} \cdot \begin{pmatrix} -11 & -2 & 6 \\ 6 & 1 & -3 \\ -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{I_3}$$

Exkurs 18.1: Rang von Matrizen

Der Begriff „Rang“ wird auch im Zusammenhang mit beliebigen Matrizen verwendet. Als *Spaltenrang* einer beliebigen Matrix **A** bezeichnet man die Anzahl der linear unabhängigen Spalten der Matrix, als *Zeilenrang* die Anzahl der linear unabhängigen Zeilen. Man kann zeigen, dass Spalten- und Zeilenrang einer Matrix stets übereinstimmen. Man spricht daher kürzer vom **Rang** der Matrix **A** und verwendet für diesen die Notation $rg\mathbf{A}$. Der Rang einer $(n \times k)$-Matrix **A** kann höchstens so groß sein wie die kleinere der beiden Zahlen n und k, die mit $\min(n; k)$ bezeichnet sei:

$$\mathbf{A} \text{ ist eine } (n \times k)\text{-Matrix } \Rightarrow rg\mathbf{A} \leq \min(n; k).$$

Wenn speziell $rg\mathbf{A} = \min(n;k)$ gilt, sagt man, dass $\mathbf{A}$ vollen Rang hat. Bei einer quadratischen Matrix bedeutet die Eigenschaft vollen Rang zu besitzen, dass sie regulär und damit auch invertierbar ist.

Da in der *Psychologie* häufig multivariate Modelle und Verfahren eingesetzt werden, u. a. multiple Regression und mehrfaktorielle Varianzanalyse, und in diesem Kontext der Einsatz von Vektoren und Matrizen hier eine übersichtliche und kompakte Notation ermöglicht, sind Grundlagen der Matrizenrechnung in einigen Statistiklehrbüchern für Studierende der Psychologie wiedergegeben, etwa bei Eid / Gollwitzer / Schmitt (2013). Eine detailliertere Behandlung von Matrixoperationen, u. a. auch Verfahren zur Bestimmung von Rängen für Matrizen und zur Inversion quadratischer Matrizen, findet man in einführenden Lehrbüchern der linearen *Algebra*, z. B. bei Gramlich (2014).

18.3 Charakterisierung von Zufallsvektoren

Die Elemente eines Vektors sind nicht notwendigerweise reelle Zahlen. Sie können auch den Charakter von Zufallsvariablen besitzen. Die n Störvariablen U_i des Regressionsmodells (16.26) lassen sich z. B. zu einem Zufallsvektor zusammenfassen, der mit $\mathbf{u}$ bezeichnet sei: [1]

$$\mathbf{u} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} = (U_1, U_2, \ldots, U_n)'. \tag{18.10}$$

Erwartungswert eines Zufallsvektors

Ähnlich wie bei der Multiplikation eines nicht-stochastischen Vektors mit einem Skalar wird der **Erwartungswert** eines Zufallsvektors gebildet, indem man den Erwartungswert für jedes einzelne Element des Vektors einzeln bestimmt. Für den Störvariablenvektor $\mathbf{u}$ verifiziert man z. B. bei Berücksichtigung der für das multiple Regressionsmodell getroffenen Annahme (MA3a), dass dessen Erwartungswert $\mathbf{u}$ der Nullvektor ist:

$$E(\mathbf{u}) = \begin{pmatrix} E(U_1) \\ E(U_2) \\ \vdots \\ E(U_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}. \tag{18.11}$$

Kovarianzmatrix eines Zufallsvektors:

- kompakte Schreibweise

Bei einer eindimensionalen Zufallsvariablen U mit Erwartungswert $E(U)$ ist die Varianz $V(U)$ gemäß (11.8) durch $E\left[(U - E(U))(U - E(U))\right]$ definiert. Das Streuungsverhalten eines *Vektors* $\mathbf{u}$ von Zufallsvariablen charakterisiert man ganz analog durch

[1] Fette Großbuchstaben werden in diesem Manuskript nur für Matrizen verwendet.

$$V(\mathbf{u}) := E[\underbrace{(\mathbf{u} - E(\mathbf{u}))}_{n\times 1}\ \underbrace{(\mathbf{u} - E(\mathbf{u}))'}_{1\times n}], \tag{18.12}$$

also als Erwartungswert des Produkts aus dem $(n \times 1)$-Spaltenvektor $\mathbf{u} - E(\mathbf{u})$ und dem durch Transposition aus diesem abgeleiteten $(1 \times n)$-Zeilenvektor. Das Transponieren des zweiten Vektors ist notwendig, damit die Vektoren miteinander multiplizierbar sind. Das in (18.12) in der eckigen Klammer stehende Produkt hat somit die Dimension $(n \times n)$, repräsentiert also eine quadratische Matrix.

Auch der Erwartungswert einer Matrix wird gliedweise gebildet. Man erhält für den Erwartungswert der $(n \times n)$-Matrix aus (18.12) eine ausführlichere Darstellung, aus der man ersieht, dass $V(\mathbf{u})$ eine Matrix ist, die die Varianzen $V(U_i)$ sowie die gemäß (13.11) definierten Kovarianzen $Cov(U_i, U_j)$ der Elemente des Zufallsvektors $\mathbf{u}$ zusammenfasst und als **Kovarianzmatrix** von $\mathbf{u}$ angesprochen wird:

- ausführliche Schreibweise

$$V(\mathbf{u}) = \underbrace{\begin{pmatrix} V(U_1) & Cov(U_1,U_2) & \dots & Cov(U_1,U_n) \\ Cov(U_2,U_1) & V(U_2) & \dots & Cov(U_2,U_n) \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(U_n,U_1) & Cov(U_n,U_2) & \dots & V(U_n) \end{pmatrix}}_{n\times n} \tag{18.13}$$

Die Darstellungen (18.12) - (18.13) gelten ganz allgemein, also für jeden beliebigen Vektor $\mathbf{u}$ von Zufallsvariablen. Für den Vektor der Störvariablen $\mathbf{u}$ des Regressionsmodells (16.26) vereinfachen sich beide Darstellungen noch, weil man hier auf die Annahmen (MA3a) und (MA3b) zurückgreifen kann. Mit (MA3a) gilt $E(\mathbf{u}) = \mathbf{0}$ und damit $V(\mathbf{u}) = E(\mathbf{u}\mathbf{u}')$ und die Elemente $V(U_1), ..., V(U_n)$ auf der Hauptdiagonalen von (18.13) haben alle denselben Wert σ^2. Die anderen Elemente von der Matrix $V(\mathbf{u})$ sind nach (MA3b) Null. Die Matrix $V(\mathbf{u})$ ist demnach hier eine Diagonalmatrix, die als Vielfaches der $(n \times n)$-Einheitsmatrix $\mathbf{I}_n$ darstellbar ist:

$$V(\mathbf{u}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \cdot \mathbf{I}_n. \tag{18.14}$$

19 Tabellenanhang

In diesem Kapitel findet man Werte der Verteilungsfunktion der Binomialverteilung und der Standardnormalverteilung. Für die Standardnormalverteilung sind auch ausgewählte Quantile tabelliert. Für einige weitere Verteilungen, nämlich die χ^2- und die t-Verteilung sowie die F-Verteilung, sind ausschließlich Quantile wiedergegeben.

Die Tabellen dieses Kapitels geben nur ausgewählte Werte der genannten Verteilungsfunktionen bzw. nur ausgewählte Quantile wieder. Es wurden aber interaktive Visualisierungen entwickelt, die einen Zugang zu weiteren Werten vermitteln und eine inhaltliche Interpretation der tabellierten Werte liefern.

Vorschau auf das Kapitel

19.1 Verteilungsfunktion der Binomialverteilung

Interaktives Lernobjekt „Binomialverteilung“

Es sei $X \sim B(n;p)$ eine mit Parametern n und p binomialverteilte Zufallsvariable. Die Wahrscheinlichkeitsfunktion $f(x) = P(X = x)$ lautet

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

und die **Verteilungsfunktion** $F(x) = P(X \leq x)$ ist

$$F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} \qquad x = 0{,}1, \ldots, n.$$

Um das Verhalten von X vollständig zu charakterisieren, benötigt man nur eine der beiden obigen Funktionen; die zweite lässt sich durch die andere ausdrücken.

Interaktives Lernobjekt „Rechnen mit der Binomialverteilung“

In Tabelle 19.1 sind Werte $F(x)$ der Verteilungsfunktion einer $B(n;p)$-verteilten Zufallsvariablen X für $n = 1{,}2, \ldots, 20$ und $p = 0{,}05$ sowie für $p = 0{,}10, 0{,}20, \ldots, 0{,}50$ zusammengestellt. Man entnimmt der Tabelle z. B., dass $F(x)$ im Falle $n = 10$ und $p = 0{,}50$ für $x = 3$ 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 $x = 3$.

Will man einen Wert der Wahrscheinlichkeitsfunktion $f(x)$ einer Binomialverteilung ermitteln, kann man diesen als Differenz von zwei Werten der Verteilungsfunktion $F(x)$ errechnen. Der Wert $f(3)$ der Wahrscheinlichkeitsfunktion der Binomialverteilung mit $n = 10$ und $p = 0{,}50$ an der Stelle $x = 3$ errechnet sich z. B. als $f(3) = F(3) - F(2)$, also als $0{,}1719 - 0{,}0547 = 0{,}1172$.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_19

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	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
2	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
3	1	0,9928	0,9720	0,8960	0,7840	0,6480	0,5000
3	2	0,9999	0,9990	0,9920	0,9730	0,9360	0,8750
3	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
4	1	0,9860	0,9477	0,8192	0,6517	0,4752	0,3125
4	2	0,9995	0,9963	0,9728	0,9163	0,8208	0,6875
4	3	1,0000	0,9999	0,9984	0,9919	0,9744	0,9375
4	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
5	1	0,9774	0,9185	0,7373	0,5282	0,3370	0,1875
5	2	0,9988	0,9914	0,9421	0,8369	0,6826	0,5000
5	3	1,0000	0,9995	0,9933	0,9692	0,9130	0,8125
5	4	1,0000	1,0000	0,9997	0,9976	0,9898	0,9688
5	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
6	1	0,9672	0,8857	0,6554	0,4202	0,2333	0,1094
6	2	0,9978	0,9842	0,9011	0,7443	0,5443	0,3438
6	3	0,9999	0,9987	0,9830	0,9295	0,8208	0,6563
6	4	1,0000	0,9999	0,9984	0,9891	0,9590	0,8906
6	5	1,0000	1,0000	0,9999	0,9993	0,9959	0,9844
6	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
7	1	0,9556	0,8503	0,5767	0,3294	0,1586	0,0625
7	2	0,9962	0,9743	0,8520	0,6471	0,4199	0,2266
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

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
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
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

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	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
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

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	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
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

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
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
19	2	0,9335	0,7054	0,2369	0,0462	0,0055	0,0004
19	3	0,9868	0,8850	0,4551	0,1332	0,0230	0,0022
19	4	0,9980	0,9648	0,6733	0,2822	0,0696	0,0096
19	5	0,9998	0,9914	0,8369	0,4739	0,1629	0,0318
19	6	1,0000	0,9983	0,9324	0,6655	0,3081	0,0835
19	7	1,0000	0,9997	0,9767	0,8180	0,4878	0,1796
19	8	1,0000	1,0000	0,9933	0,9161	0,6675	0,3238

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
19	9	1,0000	1,0000	0,9984	0,9674	0,8139	0,5000
19	10	1,0000	1,0000	0,9997	0,9895	0,9115	0,6762
19	11	1,0000	1,0000	1,0000	0,9972	0,9648	0,8204
19	12	1,0000	1,0000	1,0000	0,9994	0,9884	0,9165
19	13	1,0000	1,0000	1,0000	0,9999	0,9969	0,9682
19	14	1,0000	1,0000	1,0000	1,0000	0,9994	0,9904
19	15	1,0000	1,0000	1,0000	1,0000	0,9999	0,9978
19	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9996
19	17	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	0	0,3585	0,1216	0,0115	0,0008	0,0000	0,0000
20	1	0,7358	0,3917	0,0692	0,0076	0,0005	0,0000
20	2	0,9245	0,6769	0,2061	0,0355	0,0036	0,0002
20	3	0,9841	0,8670	0,4114	0,1071	0,0160	0,0013
20	4	0,9974	0,9568	0,6296	0,2375	0,0510	0,0059
20	5	0,9997	0,9887	0,8042	0,4164	0,1256	0,0207
20	6	1,0000	0,9976	0,9133	0,6080	0,2500	0,0577
20	7	1,0000	0,9996	0,9679	0,7723	0,4159	0,1316
20	8	1,0000	0,9999	0,9900	0,8867	0,5956	0,2517
20	9	1,0000	1,0000	0,9974	0,9520	0,7553	0,4119
20	10	1,0000	1,0000	0,9994	0,9829	0,8725	0,5881
20	11	1,0000	1,0000	0,9999	0,9949	0,9435	0,7483
20	12	1,0000	1,0000	1,0000	0,9987	0,9790	0,8684
20	13	1,0000	1,0000	1,0000	0,9997	0,9935	0,9423
20	14	1,0000	1,0000	1,0000	1,0000	0,9984	0,9793
20	15	1,0000	1,0000	1,0000	1,0000	0,9997	0,9941
20	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9987
20	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998
20	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

19.2 Verteilungsfunktion der Standardnormalverteilung

Ist X eine mit Erwartungswert μ und Varianz σ^2 normalverteilte Zufallsvariable, also $X \sim N(\mu, \sigma^2)$, so lässt sie sich anhand ihrer Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

oder anhand ihrer Verteilungsfunktion $F(x) = P(X \leq x)$ charakterisieren. Die erste Ableitung $F'(x)$ der Verteilungsfunktion und die Dichtefunktion $f(x)$ sind über die Beziehung $F'(x) = f(x)$ verknüpft. Jede normalverteilte Zufallsvariable X lässt sich mit der Transformation

$$Z := \frac{X-\mu}{\sigma}$$

in die als **Standardnormalverteilung** angesprochene Normalverteilung mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$ überführen. Es genügt daher Werte der Verteilungsfunktion der Standardnormalverteilung zu tabellieren. Für diese Funktion hat sich die Bezeichnung $\Phi(z)$ (lies: *Groß-Phi von z*) etabliert und für die Dichtefunktion $\Phi'(z)$ der Standardnormalverteilung die Bezeichnung $\phi(z)$ (lies: *Klein-Phi von z*). Zwischen der Verteilungsfunktion $F(x)$ einer $N(\mu, \sigma^2)$-verteilten Zufallsvariablen und der Verteilungsfunktion $\Phi(z)$ der standardisierten Variablen Z besteht die Beziehung

Interaktives Lernobjekt „Standardnormalverteilung“

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z).$$

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. Dabei ist die letzte Dezimalstelle der Werte z im Tabellenkopf ausgewiesen. Abbildung 12.3 liefert für $z = 1{,}38$ ein Ablesebeispiel und veranschaulicht zudem, dass sich der Wert $\Phi(1{,}38) = 0{,}9162$ als Inhalt der Fläche unter der Dichtekurve $\phi(z)$ bis zur Stelle $z = 1{,}38$ auffassen lässt. Die Beschränkung von Tabelle 19.2 auf nichtnegative Werte von z ist gerechtfertigt aufgrund der Symmetriebeziehung

Interaktives Lernobjekt „Rechnen mit der Standardnormalverteilung“

$$\Phi(z) = 1 - \Phi(-z).$$

Werte der Verteilungsfunktion $\Phi(z)$ sind auch über das erste, ganz oben auf dieser Seite eingestellte interaktive Lernobjekt zugänglich. Bei Verwendung dieses Elements kann $\Phi(z)$ 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. Der Flächeninhalt unterhalb der Dichte im Bereich von $z = 0{,}59$ bis $z = 1{,}65$ ist z. B. durch $\Phi(1{,}65) - \Phi(0{,}59)$ gegeben, also durch $0{,}9505 - 0{,}7224 = 0{,}2281$.

Tab. 19.2: *Werte der Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung*

z	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8079	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177

Fortsetzung nächste Seite

Tab. 19.2: *Werte der Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung*

z	0	1	2	3	4	5	6	7	8	9
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9956	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964

Fortsetzung nächste Seite

Tab. 19.2: *Werte der Verteilungsfunktion $\Phi(z)$ der Standardnormalverteilung*

z	0	1	2	3	4	5	6	7	8	9
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

19.3 Quantile der Standardnormalverteilung

Aus Tabelle 19.2 lassen sich auch **Quantile** ablesen. Ein p-Quantil z_p der Standardnormalverteilung ist durch

$$\Phi(z_p) = p \qquad (0 < p < 1)$$

definiert und markiert den Punkt auf der z-Achse, bis zu dem die Fläche unter der Dichte gerade p ist. Der in Abbildung 12.3 beispielhaft markierte Punkt $z = 1{,}38$ ist also das 0,9162-Quantil der Standardnormalverteilung.

Wenn man Quantile z_p anhand von Tabelle 19.2 ablesen will, findet man aber i. Allg. nicht exakt die in der Praxis am häufigsten verwendeten Quantile. Sucht man etwa das 0,95-Quantil $z_{0,95}$, so stünden bei Verwendung von Tabelle 19.2 nur die Quantile $z_{0,9495} = 1{,}64$ und $z_{0,9505} = 1{,}65$ zur Verfügung, aus denen man $z_{0,95}$ etwas umständlich per Interpolation bestimmen müsste. Aus diesem Grunde sind in der folgenden Tabelle 19.3 einige gebräuchliche Quantile separat tabelliert. 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 (12.25) wiedergegeben ist und sich aus der Symmetrie von Dichte- und Verteilungsfunktion bezüglich $z = 0$ ergibt.

Lernobjekt „Quantile der Standardnormalverteilung“

Mit $z_{0,95} = 1{,}6449$ gilt also z. B. $z_{0,05} = -1{,}6449$. Weitere p-Quantile der Standardnormalverteilung sind bei Verwendung des nebenstehenden interaktiven Lernobjekts zugänglich. Dieses gestattet es sogar, die Quantile im Falle $p < 0{,}5$ direkt abzulesen.

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*

19.4 Quantile der χ^2-Verteilung

In der folgenden Tabelle sind, auf drei Dezimalstellen gerundet, Quantile $\chi^2_{n;p}$ der χ^2-Verteilung mit n Freiheitsgraden für $n = 1$ bis $n = 30$ und ausgewählte Werte p zusammengestellt. Man entnimmt der Tabelle z. B., dass das 0,95-Quantil der χ^2-Verteilung mit $n = 8$ Freiheitsgraden den Wert $\chi^2_{8;0,95} = 15{,}507$ besitzt. Weitere Quantile der χ^2-Verteilung lassen sich anhand des nebenstehenden interaktiven Lernobjekts anzeigen.

Interaktives Lernobjekt „Quantile der χ^2-Verteilung“

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
26	12,198	13,844	15,379	38,885	41,923	45,642
27	12,879	14,573	16,151	40,113	43,195	46,963
28	13,565	15,308	16,928	41,337	44,461	48,278
29	14,256	16,047	17,708	42,557	45,722	49,588
30	14,953	16,791	18,493	43,773	46,979	50,892

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 0,975-Quantil der t-Verteilung mit $n = 8$ Freiheitsgraden den Wert $t_{8;0,975} = 2{,}306$ besitzt. Bei bekanntem p-Quantil ergibt sich das $(1-p)$-Quantil aus der Beziehung

$$t_{n;p} = -t_{n;1-p},$$

die in (12.29) 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. Ab etwa $n = 30$ lassen sich die Quantile der t-Verteilung gut durch die entsprechenden Quantile z_p der Standardnormalverteilung approximieren.

Interaktives Lernobjekt „Quantile der t-Verteilung“

Tab. 19.5: *Quantile der t-Verteilung*

	p						
n	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
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

Fortsetzung nächste Seite

Tab. 19.5: *Quantile der t-Verteilung*

	p						
n	0,800	0,850	0,900	0,950	0,975	0,990	0,995
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 0,99-Quantil der F-Verteilung mit $m = 5$ und $n = 10$ Freiheitsgraden den Wert $F_{5;10;0,99} = 5{,}64$ hat.Weitere Quantile der χ^2-Verteilung lassen sich anhand des nebenstehenden interaktiven Lernobjekts anzeigen.

Interaktives Lernobjekt „Quantile der F-Verteilung“

Tab. 19.6: *Quantile der F-Verteilung ($p = 0{,}95$, $m = 1$ bis $m = 10$)*

	m									
n	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
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

Fortsetzung nächste Seite

Tab. 19.6: *Quantile der F-Verteilung ($p = 0{,}95$, $m = 1$ bis $m = 10$)*

	m									
n	1	2	3	4	5	6	7	8	9	10
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 ($p = 0{,}95$, $m > 11$)*

	m									
n	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
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

Fortsetzung nächste Seite

Tab. 19.6: *Quantile der F-Verteilung ($p = 0{,}95$, $m > 11$)*

	m									
n	11	12	13	14	15	20	30	40	50	100
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 $m = 10$)*

	m									
n	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
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

Fortsetzung nächste Seite

Tab. 19.6: *Quantile der F-Verteilung ($p = 0{,}99$, $m = 1$ bis $m = 10$)*

	m									
n	1	2	3	4	5	6	7	8	9	10
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 ($p = 0{,}99$, $m > 11$)*

	m									
n	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
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

Fortsetzung nächste Seite

Tab. 19.6: *Quantile der F-Verteilung* ($p = 0{,}99$, $m > 11$)

	m									
n	11	12	13	14	15	20	30	40	50	100
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

20.1 Beschreibende Statistik

Aufgabe 2.1 (Grundbegriffe)

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.

Kapitel 2

Was sind hier Grundgesamtheit, statistische Einheit, Merkmal und Merkmalsausprägung? Wie könnte man bezüglich der Grundgesamtheit noch 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)

Geben Sie einige Beispiele für Zeitreihen an, die regelmäßig in den Medien zu finden sind.

Kapitel 3

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?

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_20

Kapitel 4

Aufgabe 4.1 (Ergebnisse der Nationalen Verzehrstudie II für Frauen)

Die nachstehende Tabelle zur Nationalen Verzehrstudie II ist analog zu Tabelle 4.2 angelegt und bezieht sich ebenfalls auf gruppierte Daten. Während Tabelle 4.2 absolute und relative Häufigkeiten für drei Ausprägungsintervalle des Body-Mass-Index (BMI) für Männer in den deutschen Bundesländern zeigte, gibt die folgende Tabelle die entsprechenden Häufigkeiten für die an der Studie beteiligten Frauen wieder (ungewichtete Daten; Quelle: Persönliche Mitteilung des Max-Rubner-Instituts). Bei den BMI-Werten wird erneut nur zwischen drei Ausprägungen a_1, a_2 und a_3 unterschieden (a_1: Unter- oder Normalgewicht, a_2: Übergewicht, a_3: Fettleibigkeit).

Bundesland (weibliche 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 (924)	487	0,527	287	0,311	150	0,162
Bayern (1157)	602	0,520	340	0,294	215	0,186
Berlin (270)	157	0,581	65	0,241	48	0,178
Brandenburg (203)	95	0,468	56	0,276	52	0,256
Bremen (63)	35	0,556	14	0,222	14	0,222
Hamburg (133)	82	0,617	32	0,241	19	0,143
Hessen (446)	229	0,513	135	0,303	82	0,184
Mecklenburg-Vorp. (131)	50	0,382	43	0,328	38	0,290
Niedersachsen (851)	425	0,499	261	0,307	165	0,194
Nordrhein-Westf. (1495)	741	0,496	425	0,284	329	0,220
Rheinland-Pfalz (321)	175	0,545	74	0,231	72	0,224
Saarland (84)	37	0,440	24	0,286	23	0,274
Sachsen (360)	155	0,431	113	0,314	92	0,256
Sachsen-Anhalt (180)	76	0,422	59	0,328	45	0,250
Schleswig-Holstein (263)	133	0,506	87	0,331	43	0,164
Thüringen (209)	88	0,421	68	0,325	53	0,254
Summe: 7090	3567		2083		1440	

a) Stellen Sie die relativen Häufigkeiten in Form gestapelter Säulendiagramme dar. Unterdrücken Sie dabei, analog zu Abbildung 4.5, die Wiedergabe der Häufigkeiten $f(a_1)$. Ordnen Sie die Bundesländer nach zunehmender Größe der Summe $f(a_2) + f(a_3) = 1 - f(a_1)$.

b) Vergleichen Sie für jedes Bundesland die in obiger Tabelle wiedergegebenen Ergebnisse für Frauen mit den in Tabelle 4.2 präsentierten Ergebnissen für Männer. Weisen Sie in einer neu anzulegenden Tabelle in einer Spalte (1) den Anteil der Frauen mit einem BMI-Wert von mindestens 25,0 aus, also die Summe $f(a_2) + f(a_3)$ aus vorstehender Tabelle. Geben Sie dann in einer weiteren Spalte (2) den Anteil der Männer mit dieser Eigenschaft wieder, also die Summe $f(a_2) + f(a_3)$ aus Tabelle 4.2. In einer Spalte (3) können Sie auch den Quotienten q_1 der Werte in (1) und (2) ausweisen. Was beinhaltet ein Wert $q_1 < 1$?

c) Vergleichen Sie die Ergebnisse für Frauen und Männer anschließend bezüglich des Verhältnisses von „schwereren“ und „leichteren“ Fällen von Übergewichtigkeit. Berechnen Sie hierzu zunächst anhand obiger Tabelle das Verhältnis $\frac{f(a_3)}{f(a_2)}$ für Frauen und geben Sie die resultierenden Werte in einer weiteren Spalte (4) der neuen Tabelle wieder. Für die Männer können Sie die analogen Werte anhand von Tabelle 4.2 in einer zusätzlichen Spalte (5) darstellen. Berechnen Sie in einer letzten Spalte (6) den Quotienten q_2 der Werte in (4) und (5). Was beinhaltet $q_2 > 1$?

Aufgabe 4.2 (Gruppierung von Daten und Histogrammerstellung)

Für 80 Arbeitnehmer in Portugal, die im Bereich „Industrie und Dienstleistungen“ tätig sind, wurden für das Referenzjahr 2012 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 4.7, die sich allerdings auf Brutto*jahres*verdienste bezog und daher Intervalle von 5000 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 zweckmäßigerweise 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 4.12 – 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 12 Mal nacheinander mit einem Würfel gewürfelt wurde:

Kapitel 5

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. Bei der Berechnung von Varianz und Standardabweichung ist auch der Rechengang wiederzugeben. Die Ergebnisse sind auf 3 Stellen nach dem Dezimalkomma genau anzugeben.

Aufgabe 5.2 (Quantile und Boxplots)

a) Bestimmen Sie für den in Aufgabe 5.1 veranschaulichten Datensatz mit 12 Werten (Würfelexperiment) die Quartile $x_{0,25}$ und $x_{0,75}$.

b) 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?

c) Wenn noch einmal gewürfelt wird und die Augenzahl 3 erscheint, hat man insgesamt einen Datensatz der Länge $n = 13$. Wie groß ist nun der Interquartilsabstand Q?

Aufgabe 5.3 (Quantile und Boxplots)

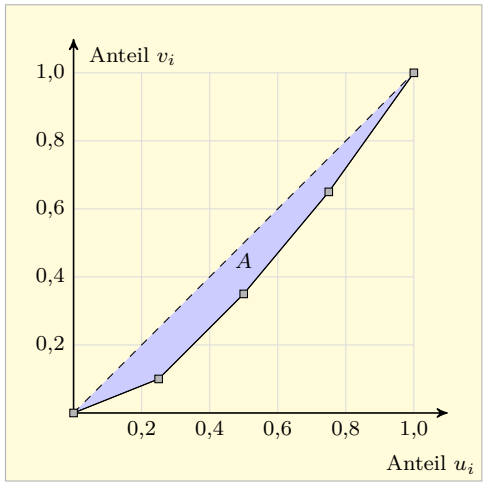
a) Im oberen Teil von Abbildung 4.12 sind die Ergebnisse aus einer Serie von 100 Roulettespielen anhand eines Balkendiagramms dargestellt. Bestimmen Sie für den Datensatz den Median und den Interquartilsabstand.

b) Visualisieren Sie den Datensatz anhand eines Boxplots. Es genügt eine maßstabsgerechte Skizze.

Aufgabe 6.1 (Gini-Koeffizient)

In einer Region konkurrieren vier Energieversorgungsunternehmen. Es seien $x_1 = 20$, $x_2 = 50$, $x_3 = 60$ und $x_4 = 70$ die Umsätze dieser Firmen (in Millionen Euro) im letzten Geschäftsjahr. Die folgende Abbildung zeigt die auf der Basis dieser Daten errechnete Lorenzkurve (Polygonzug).

Kapitel 6



i	u_i	v_i
0	0	0
1	0,25	v_1
2	0,50	v_2
3	0,75	v_3
4	1	1

Die Stützpunkte (u_i, v_i) der Lorenzkurve sind in der Grafik betont. 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.

b) Bestimmen Sie dann den Gini-Koeffizienten G aus (6.5) und den normierten Gini-Koeffizienten G^* aus (6.7).

c) Welchen Inhalt hat die Fläche A, die in der vorstehenden Abbildung betont ist (markierte Fläche zwischen der Verbindungslinie der Punkte $(0;0)$ und $(1;1)$ und der Lorenzkurve)?

Aufgabe 6.2 (Herfindahl-Index)

a) Berechnen Sie mit den Daten aus Aufgabe 6.1 den Herfindahl-Index.

b) Wie groß ist hier die untere Schranke für den Index?

Kapitel 7

Aufgabe 7.1 (Militärausgaben 2014 im Ländervergleich)

Veranschaulichen Sie die Werte in den letzten drei Spalten von Tabelle 7.1 anhand je eines Balkendiagramms. Ordnen Sie die Werte jeder Spalte zuvor nach absteigender Größe und markieren Sie in den drei Grafiken jeweils den Balken für China. Verwenden Sie für die Ländernamen die nachstehenden internationalen Codes:

USA – US; China – CN; Russland – RU; Saudi-Arabien – SA; Frankreich – FR; Großbritannien – UK; Indien –IN; Deutschland – DE; Japan – JP; Brasilien – BR; Israel – IL; Singapur – SG.

Aufgabe 7.2 (Zusammengesetzte Indexzahlen – Medaillenspiegel)

Tabelle 7.1 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 $(1-1-1)$.

a) Wie sähe für die zehn Länder der Tabelle 7.2 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?

b) 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 7.3 (Preisindex)

Aktivieren Sie den Inflationsrechner des Statistischen Bundesamts. Wählen Sie über die Schaltfläche „Güterauswahl“ die Gütergruppe „Pauschalreisen“ aus. Welche Auffälligkeiten beobachten Sie bei der Zeitreihe des Preisindexes für „Pauschalreisen“?

Aufgabe 8.1 (Randverteilungen)

Kapitel 8

Bei einer medizinischen Studie wurde für $n = 360$ Personen erfasst, ob die Beteiligten regelmäßig einen erhöhten Alkoholkonsum hatten (Überschreitung eines gewissen Schwellenwerts, bezogen auf reinen Alkohol) und ob sie Leberfunktionsstörungen aufwiesen (adaptiert aus Toutenburg / Schomaker / Wissmann (2009, Abschnitt 4.2.5)). Es sei X das Merkmal „Alkoholkonsum“ mit den Ausprägungen a_1 (oberhalb des 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 8.2 (Bedingte Häufigkeitsverteilungen)

Interpretieren Sie die Werte für die in Beispiel 8.3 errechneten bedingten Häufigkeiten $f_X(a_5|b_1) \approx 0{,}108$ und $f_Y(b_1|a_2) \approx 0{,}461$.

Aufgabe 9.1 (Zusammenhangsmessung bei Nominalskalierung)

Kapitel 9

a) Berechnen Sie den χ^2-Koeffizienten auf der Basis der Daten aus Aufgabe 8.1. Runden Sie das Ergebnis auf drei Dezimalstellen.

b) Bestimmen Sie dann auch den Koeffizienten Φ und den Kontingenzkoeffizienten V nach Cramér.

Aufgabe 9.2 (Zusammenhangsmessung bei metrischer Skalierung)

Das folgende Beispiel stammt aus Bamberg / Baur /Krapp (2012):

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	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:	$\overline{x} = 7{,}11$	$\overline{y} = 5{,}74$

Berechnen Sie den Korrelationskoeffizienten r nach Bravais-Pearson. Sofern Sie die Rechnung manuell durchführen, können Sie eine zu Tabelle 16.2 analoge Arbeitstabelle verwenden.

Aufgabe 9.3 (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 (Hallen-turnier)	Platzierung (Freiluft-turnier)
A	1	2
B	2	3
C	3	1
D	4	5
E	5	4

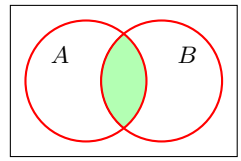
20.2 Wahrscheinlichkeitsrechnung und schließende Statistik

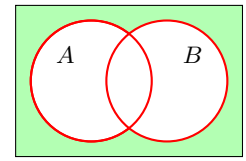
Kapitel 10

Aufgabe 10.1 (Venn-Diagramme)

Zur Veranschaulichung von Ereignissen oder von Mengen lassen sich Venn-Diagramme heranziehen. Diese bestehen aus einem Rechteck, in dem die Ereignisse als Kreise oder Ellipsen dargestellt sind. Das Rechteck repräsentiert eine Grundgesamtheit, von der die eingezeichneten Mengen Teilmengen sind. Es bezeichnen $\overline{A}$ und $\overline{B}$ die Komplementärmengen von A und B, $A \cap B$ deren Schnittmenge und $A \cup B$ die Vereinigungsmenge.

Nachstehend sind zwei Venn-Diagramme abgebildet, die sich auf die Verknüpfung zweier Ereignisse oder Mengen A und B beziehen. Welche der folgenden Aussagen sind richtig?





a) Im ersten dargestellten Venn-Diagramm ist anhand der dunkler gefärbten Fläche die Schnittmenge $A \cap B$ von A und B dargestellt.

b) Im zweiten Venn-Diagramm ist anhand der dunkler gefärbten Fläche die Schnittmenge $A \cap \overline{B}$ aus A und der Komplementärmenge von B dargestellt.

c) Die Schnittmengen $A \cap B$ und $A \cap \overline{B}$ sind disjunkt, d.h. ihre Darstellungen in Venn-Diagrammen überschneiden sich nicht.

d) Die Vereinigung von $A \cap B$ und $A \cap \overline{B}$ liefert A.

e) Die Wahrscheinlichkeiten für ein aus zwei disjunkten Ereignissen zusammensetztes Ereignis A ergibt sich als Summe der Wahrscheinlichkeiten der beiden Ereignisse.

Aufgabe 10.2 (Ereignisse und Ereignisraum)

Eine Münze wird dreimal nacheinander geworfen. Bei jedem einzelnen Wurf sind die möglichen Ausgänge durch „Z“ (Zahl) und „K“ (Kopf) beschrieben. Die möglichen Ausgänge eines dreifachen Münzwurfs (Elementarereignisse) sind entsprechend durch Tripel definiert, die aus den beiden Symbolen „Z“ und „K“ gebildet werden.

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 10.3 (Wahrscheinlichkeiten bei Laplace-Experimenten)

Berechnen Sie für das statistische Experiment „Dreifacher Münzwurf“ die Wahrscheinlichkeiten für die Ereignisse

$$A = \{\text{Bei mindestens zwei Würfen tritt „K“ auf}\}$$
$$\overline{A} = \{\text{Bei höchstens einem Wurf tritt „K“ auf}\}.$$

Setzen Sie eine faire Münze voraus, also gleiche Eintrittswahrscheinlichkeiten für die Ausgänge „Z“ (Zahl) und „K“ (Kopf).

Aufgabe 10.4 (Kombinatorik)

Die FernUniversität ordnet allen Studierenden eine mehrstellige Nummer zu (Matrikel-Nummer), die als Identifikationszwecken verwendet wird. Alternativ könnte man auch Buchstabenkombinationen verwenden.

Wieviele Studierende könnte man maximal anhand von Buchstabenkombinationen unterscheiden, wenn für jede Person genau 5 Großbuchstaben aus der Buchstabenfolge von A bis J verwendet würden, also z. B. BCBJD oder AFGGC?

Aufgabe 10.5 (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 diese 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“ $\overline{A}$) sowie „Erreichter akademischer Grad“ (Ausprägungen „Hochschulabschluss mit Promotion“ B und „Hochschulabschluss ohne Promotion“ $\overline{B}$) ausweisen. Berechnen Sie die Wahrscheinlichkeiten aus den Aufgabenteilen a - c auch anhand dieser Tafel.

Aufgabe 10.6 (Bedingte Wahrscheinlichkeiten und Vierfeldertafel)

Sie haben nun das Rüstzeug dafür zu überprüfen, ob Sie in Beispiel 1.3 die Größenordnung für die Wahrscheinlichkeit eines falsch-positiven Befundes intuitiv richtig eingeschätzt haben. Bezeichnen Sie das Ereignis „Eine zur Grundgesamtheit gehörende Frau hat Krebs“ mit A und den Fall „Screeningbefund ist positiv“ mit B. Die Komplementärereignisse sind mit $\overline{A}$ resp. $\overline{B}$ anzusprechen.

a) Zerlegen Sie zunächst gedanklich die N Frauen umfassende Grundgesamtheit G in zwei Teilpopulationen G_1 und G_2, wobei G_1 alle N_1 Frauen mit A und G_2 die $N_2 = N - N_1$ Frauen mit $\overline{A}$ umfasse. Zeichnen Sie ein Balkendiagramm, das die Anzahl der positiven Befunde in beiden Gruppen ausweist, wobei der Gesamtumfang N beider Gruppen als Bezugsgröße dient (relative Anzahl, z. B. ausgedrückt in % von N).

b) Berechnen Sie die Wahrscheinlichkeit $P(B)$ dafür, dass eine zufällig aus der Gesamtpopulation ausgewählte Frau mit einem positiven Befund konfrontiert wird.

c) Bestimmen Sie die Wahrscheinlichkeit $P(\overline{A}|B)$ dafür, dass eine Frau mit positivem Screeningbefund gesund ist, also ein Fehlalarm erfolgte.

d) Wie groß wären die Wahrscheinlichkeiten $P(B)$ und $P(\overline{A}|B)$, wenn man bei der betrachteten Grundgesamtheit G für die Erkrankungswahrscheinlichkeit $P(A)$ nicht den in Beispiel 1.1 angenommenen Wert 0,008, sondern 0,006 voraussetzte?

Aufgabe 10.7 (Bedingte Wahrscheinlichkeiten und Baumdiagramm)

a) Wie häufig man bei dem in der vorstehenden Aufgabe 10.6 genannten Beispiel mit Fehlentscheidungen zu rechnen hat, also mit falsch-positiven Befunden (Fehlalarme) oder falsch-negativen Befunden (unterbliebene Alarme), lässt sich besonders anschaulich – auch ohne Bayes-Formel und leichter nachvollziehbar – unter Verwendung eines Baumdiagramms für absolute Häufigkeiten visualisieren.

   Zeichnen Sie ein zu Abbildung 8.2 analoges Baumdiagramm, also ein Diagramm, das sich auf absolute Häufigkeiten bezieht und die relativen Häufigkeiten nur als Zusatzinformation wiedergibt. Unterteilen Sie die Grundgesamtheit zunächst nach der Binärvariablen „Gesundheitsstatus“ (Ausprägungen „erkrankt“ und „gesund“) und danach beide Teilmengen jeweils noch nach der Binärvariablen „Screeningbefund“ (Ausprägungen „positiv“ und „negativ“). Gehen Sie von $N = 100\,000$ Teilnehmerinnen aus. Bestimmen Sie anhand des Diagramms und des Laplace-Ansatzes (10.5) erneut – nun auf andere Weise – die Wahrscheinlichkeiten, die in Teil a - c von Aufgabe 10.6 zu bestimmen waren.

b) Geben Sie die vier absoluten Häufigkeiten am Ende des Baumdiagramms (dritte Ebene) in einer Vierfeldertafel für absolute Häufigkeiten wieder. Die Tafel soll auch die Randverteilungen für die beiden Binärvariablen „Gesundheitsstatus“ und „Screeningbefund“ ausweisen. Berechnen Sie nochmals – jetzt anhand der Vierfeldertafel – die Wahrscheinlichkeiten aus Teil b - c von Aufgabe 10.6.

Aufgabe 10.8 (Wahrscheinlichkeit bei einem Würfelspiel)

Bei dem bekannten Spiel *Mensch ärgere Dich nicht* darf jeder Spieler zu Beginn drei Mal würfeln. Sobald dabei eine Sechs gewürfelt wird, darf eine Spielfigur starten, also auf den Rundparcours gesetzt werden. Spieler A beginnt. Wie groß ist die Wahrscheinlichkeit P dafür, dass bei Spieler A auch nach dem dritten Wurf mit dem Würfel noch keine Sechs gefallen ist? Setzen Sie voraus, dass der verwendete Würfel fair ist, also gleiche Eintrittswahrscheinlichkeiten für die einzelnen Augenzahlen aufweist.

Kapitel 11

Aufgabe 11.1 (Würfeln mit zwei Würfeln)

Es werde mit zwei „fairen“ Würfeln gewürfelt, also solchen mit gleicher Eintrittswahrscheinlichkeit für jede Augenzahl, und die Summe X der beiden Augenzahlen festgestellt.

a) Welche Ausprägungen sind für die Zufallsvariable X möglich? Welche Eintrittswahrscheinlichkeiten besitzen die Ausprägungen?

b) Welchen Wert hat die Verteilungsfunktion $F(x)$ der Augesumme X an den Stellen $x = 0{,}5$, $x = 3$, $x = 3{,}5$ und $x = 6$?

c) Berechnen Sie auch den Erwartungswert von X.

Aufgabe 11.2 (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

a) mindestens zwei Kugelschreiber gewonnen werden?

b) genau einmal „Grün“ auftritt, also ein einwöchiges Freiabonnement gewährt wird?

Aufgabe 11.3 (Hypergeometrische Verteilung)

In Österreich und der Schweiz wird das Lottospiel „6 aus 45“ gespielt, nicht „6 aus 49“ wie in Deutschland.

a) Berechnen Sie den Erwartungswert für die Anzahl X der Richtigen.

b) Wie groß ist hier die Wahrscheinlichkeit des Ereignisses „6 Richtige“?

Aufgabe 11.4 (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?

Kapitel 12

Aufgabe 12.1 (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 Fahrgast 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 12.2 (Normalverteilung und Standardnormalverteilung)

a) 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.

b) 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 12.3 (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 $\mu = 2950$ und Standardabweichung $\sigma = 120$ modellieren lässt.

a) Wie groß ist die Wahrscheinlichkeit, dass ein Neugeborenes nicht mehr als 2800 Gramm wog?

b) Wie groß ist die Wahrscheinlichkeit für ein Gewicht zwischen 2800 und 3250 Gramm?

c) Was beinhaltet das 0,1-Quantil der Normalverteilung mit $\mu = 2950$ und Varianz $\sigma^2 = 120^2$ und welchen Wert hat es hier?

Anmerkung zu Teil a: Die gesuchte Wahrscheinlichkeit $P(X \leq 2800)$ stimmt mit $P(X < 2800)$ überein, wie man aus (12.8) mit $x_0 = 2800$ ersieht. Es ist also für das Ergebnis unerheblich, ob man bei der Aufgabenformulierung „nicht mehr als 2800 Gramm“ oder „weniger als 2800 Gramm“ verwendet.

Aufgabe 12.4 (Quantile von t- und Standardnormalverteilung)

Bei einem Test werde eine Teststatistik T eingesetzt, die unter bestimmten Voraussetzungen (bei Gültigkeit der Nullhypothese des Tests) einer t-Verteilung mit $n = 10$ Freiheitsgraden folgt.

a) Geben Sie einen Wert an, den eine Ausprägung der Testgröße T mit Wahrscheinlichkeit $\alpha = 0{,}05$ nicht überschreitet.

b) 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 13.1 (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 13

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 (13.12) heranzieht, kann man den dort auftretenden Term $E(XY)$ analog zu (11.6) ermitteln als Summe der vier möglichen Ausprägungen von XY, wobei jeder Summand jeweils mit seiner Eintrittswahrscheinlichkeit gewichtet wird, also mit einer der Wahrscheinlichkeiten p_{11}, p_{12} , p_{21} resp. p_{22}.

Aufgabe 14.1 (Schätzung von Erwartungswert und Varianz)

Bei einer medizinischen Studie, an der 24 Patienten beteiligt waren, wurden auch deren Gewicht X ermittelt. Es ergaben sich folgende Werte, jeweils auf volle kg gerundet (angelehnt an Toutenburg / Schomaker / Wissmann (2009, Abschnitt 10.4)):

Kapitel 14

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 $\widehat{\mu}$ für den Erwartungswert μ der Normalverteilung.

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.

Anmerkung: Geben Sie Ihre Ergebnisse auf drei Dezimalstellen genau an. Sie können bei der Lösung dieser Aufgabe anstelle eines Taschenrechners auch eine Statistik-Software oder EXCEL heranziehen.

Aufgabe 14.2 (Konfidenzintervalle bei geschätzter Varianz)

a) Bestimmen Sie mit den Daten aus Aufgabe 14.1 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.

b) Interpretieren Sie Ihr Ergebnis.

Aufgabe 15.1 (einseitiger Gauß-Test)

Kapitel 15

Die nachstehende Aufgabe ist adaptiert aus Caputo / Fahrmeir / Künstler / Lang / Pigeot / Tutz (2009, Kapitel 10):

Bei einer Studie zum Thema „Frauen und 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 $\overline{x} = 26$ (Altersangaben in vollen Jahren).

a) 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.

b) Was beinhalten bei diesem konkreten Test die Fehler 1. und 2. Art?

Aufgabe 15.2 (einseitiger Gauß-Test)

Wenn man in Beispiel 15.3 eine weitere Stichprobe zieht, wird man möglicherweise zu einer anderen Testentscheidung kommen. Man kann aber die Wahrscheinlichkeit für die Ablehnung der Nullhypothese anhand von (15.12) berechnen, ohne hierfür Stichprobendaten zu benötigen.

a) Berechnen Sie für den linksseitigen Test (15.5) mit $\mu_0 = 2$ kg und $\alpha = 0{,}05$ aus Beispiel 15.3 die Wahrscheinlichkeit einer Verwerfung der Nullhypothese für den Fall, dass μ den Wert $\mu = 2{,}002$ kg hat.

b) Wie groß ist diese Wahrscheinlichkeit für $\mu = 1{,}997$?

c) Skizzieren Sie den vollständigen Verlauf der Gütefunktion $G(\mu)$ des linksseitigen Tests aus Aufgabenteil a.

Aufgabe 15.3 (zweiseitiger Gauß-Test)

Betrachten Sie wie in Beispiel 15.3 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, der Hersteller aber aus Kostengründen auch an einer Überwachung und Abstellung von Sollwertüberschreitungen interessiert.

a) Anhand einer Stichprobe von 10 Tüten wurde für das Füllgewicht der Mittelwert $\overline{x} = 2{,}007$ kg ermittelt. Um den Interessen von Verbraucher und Hersteller gleichermaßen zu entsprechen, soll über einen zweiseitigen Test (15.1) mit $\mu_0 = 2$ kg 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.

b) Führen Sie den Test auch mit $\alpha = 0{,}01$ durch.

Kapitel 16

Aufgabe 16.1 (Kleinst-Quadrat-Schätzung)

Im Vorfeld von Herzkatheteruntersuchungen 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 beiden Variablen „Körpergewicht“ und „Körpergröße“ können anhand des schon in Beispiel 4.2 verwendeten Body-Mass-Indexes zusammengeführt werden, dessen Wert eine schnelle 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	x_i	y_i
1	26	170
2	23	150
3	27	160
4	28	175
5	24	155
6	25	150

Gehen Sie davon aus, dass die Werte x_i und y_i über eine lineare Regression (16.1) 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 $\widehat{\beta}$ und $\widehat{\alpha}$ auf zwei Stellen nach dem Dezimalkomma genau aus.

Aufgabe 16.2 (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 $(i = 1, 2, ..., 8)$, d. h. es sind hier außer dem Nettoeinkommen keine Daten für andere denkbare Einflussvariablen aufgeführt:

i	x_i	y_i
1	1,9	3,0
2	2,7	2,5
3	3,1	4,5
4	4,0	3,5
5	3,9	4,0
6	3,4	3,0
7	2,9	4,0
8	2,1	3,5

a) Berechnen Sie unter Annahme des einfachen linearen Regressionsmodells (16.1) die KQ-Schätzungen für die Regressionskoeffizienten β und α.

b) Quantifizieren Sie anhand des Bestimmtheitsmaßes R^2 aus (16.17) die Anpassungsgüte der Regressionsgeraden. Interpretieren Sie das Ergebnis.

Aufgabe 16.3 (Kleinst-Quadrat-Schätzung)

In Beispiel 16.1, das sich auf das einfache Regressionsmodell bezog ($k = 1$), wurden die KQ-Schätzformeln (16.6) und (16.7) auf einen sehr kleinen Datensatz angewendet. Leiten Sie die dabei errechneten Schätzwerte $\widehat{\beta} = 0{,}125$ und $\widehat{\alpha} = 0{,}25$ erneut her, nun aber unter Verwendung der KQ-Schätzformel (16.35) für das multiple Regressionsmodell. Notieren Sie die Formel (16.35) zunächst für den Spezialfall $k = 1$.

21 Lösungen zu den Übungsaufgaben

21.1 Beschreibende Statistik

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)

Kapitel 3

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).

Lösung zu Aufgabe 3.2 (geschichtete Zufallsauswahl)

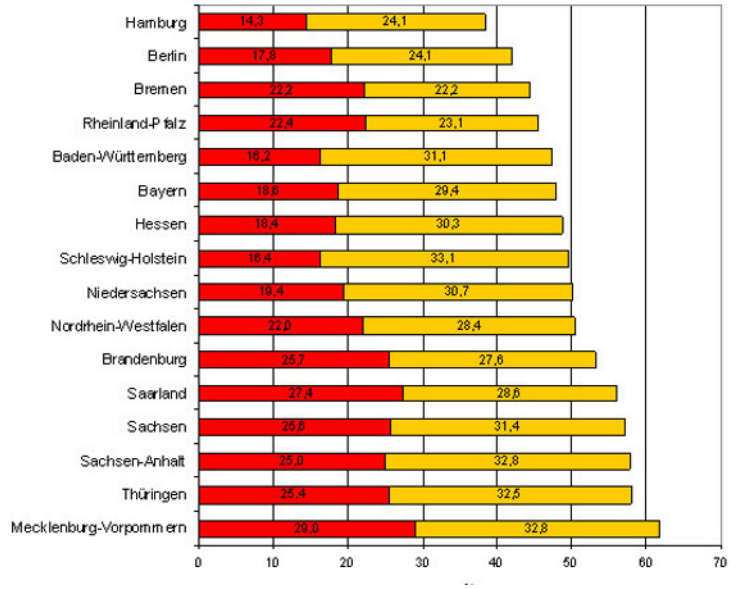
Bei proportionaler Schichtung entfallen $\frac{270}{600} \cdot 120 = 54$ Studierende auf Schicht 1, $\frac{180}{600} \cdot 120 = 36$ auf Schicht 2 und $\frac{150}{600} \cdot 120 = 30$ auf Schicht 3.

Lösung zu Aufgabe 4.1 (Nationale Verzehrstudie II)

Kapitel 4

a) Die folgende Abbildung zeigt die in Prozentwerten wiedergegebenen relativen Häufigkeiten für Frauen. Die numerischen Werte der beiden dargestellten Teilhäufigkeiten sind jeweils eingeblendet. Die Bundesländer sind nach zunehmender Gesamtlänge geordnet.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_21



b) Bei den Frauen liegt der Anteil der Personen mit einem BMI-Wert von 25 und mehr, also derjenigen, die entweder als übergewichtig (BMI von 25,0 bis unter 30,0) oder gar als fettleibig (BMI von 30,0 und mehr) eingestuft sind, in allen Bundesländern niedriger als bei den Männern. Man verifiziert dies durch Vergleich der Balken (jeweils Gesamtlänge beider Balkenabschnitte) aus Abbildung 4.5 oder anhand eines Vergleichs der Werte in den Spalten (1) und (2) der nachstehenden Tabelle. Dort ist aus den Werten der Spalte (3), die den Quotienten q_1 aus (1) und (2) wiedergibt, das bessere Abschneiden der Frauen quantifiziert. Ein Wert $q_1 < 1$ beinhaltet, dass bei den Frauen der Anteil der leicht oder deutlich Übergewichtigen (BMI-Wert von mindestens 25,0) niedriger als bei den Männern liegt. Man erkennt z. B., dass q_1 in Berlin und Hamburg besonders deutlich unterhalb von 1 liegt.

c) In den Spalten (4) und (5) der Tabelle wird für Frauen und Männer jeweils ausgewiesen, wie innerhalb der Personengruppe mit einem BMI-Wert von mindestens 25,0 der Quotient der relativen Häufigkeiten $f(a_3)$ und $f(a_2)$ ausfällt, also für jedes Geschlecht einzeln das Verhältnis zwischen dem Anteil der Fettleibigen und dem der nur schwächer Übergewichtigen. Man sieht, dass dieses Verhältnis bei den Frauen ungünstiger ausfällt, wenn man von Schleswig-Holstein absieht. Ein Wert $q_2 > 1$ besagt, dass das Verhältnis von stärker zu leichter Übergewichtigen bei den Männern günstiger ausfällt. In Bremen und im Saarland liegt q_2 auffällig deutlich oberhalb von 1.

Bundesland	$f(a_2)+f(a_3)$		q_1	$\frac{f(a_3)}{f(a_2)}$		q_2
	Frauen	Männer		Frauen	Männer	
	(1)	(2)	(3)	(4)	(5)	(6)
Baden-Württemberg	0,473	0,688	0,688	0,521	0,427	1,219
Bayern	0,480	0,661	0,726	0,633	0,479	1,321
Berlin	0,419	0,661	0,634	0,739	0,386	1,915
Brandenburg	0,532	0,689	0,772	0,928	0,591	1,569
Bremen	0,444	0,613	0,724	1,000	0,310	3,228
Hamburg	0,384	0,616	0,623	0,593	0,333	1,780
Hessen	0,487	0,694	0,702	0,607	0,437	1,390
Mecklenburg-Vorpommern	0,618	0,678	0,912	0,884	0,551	1,603
Niedersachsen	0,501	0,678	0,739	0,632	0,503	1,255
Nordrhein-Westfalen	0,504	0,672	0,750	0,775	0,427	1,815
Rheinland-Pfalz	0,455	0,679	0,670	0,970	0,380	2,551
Saarland	0,560	0,662	0,846	0,958	0,271	3,540
Sachsen	0,569	0,682	0,836	0,815	0,516	1,581
Sachsen-Anhalt	0,578	0,691	0,836	0,762	0,446	1,710
Schleswig-Holstein	0,495	0,684	0,724	0,495	0,551	0,899
Thüringen	0,579	0,692	0,837	0,782	0,514	1,520

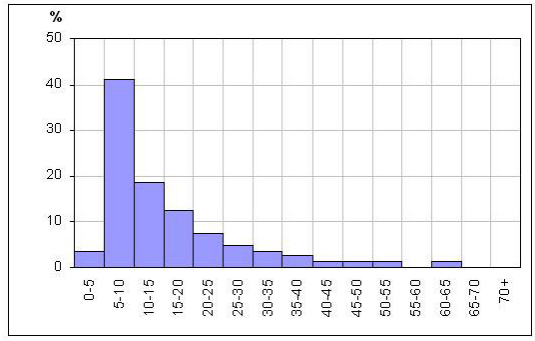
Lösung zu Aufgabe 4.2 (Gruppierung von Daten, Histogramme)

a) Merkmalsträger: Arbeitnehmer
Merkmal: Bruttoverdienst / Stunde (in EUR).

b)

Nr. der Klasse	Klassen-grenzen	Klassenbesetzungshäufigkeit absolut	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 $3, 4, \ldots 18$.

b) Die empirische Verteilungsfunktion kann höchstens 16 Sprünge aufweisen.

Kapitel 5

Lösung zu Aufgabe 5.1

a) Häufigkeitsverteilung für das Merkmal „Augenzahl“:

Augenzahlen	⚀	⚁	⚂	⚃	⚄	⚅
Absolute Häufigkeit	1	2	3	2	1	3
Relative Häufigkeit	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$

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 $\widetilde{x} = \frac{1}{2} \cdot (3 + 4) = 3{,}5$.

Nach (5.2) erhält man dann $\overline{x} = \frac{1}{12} \cdot 45 = 3{,}75$. Wenn man alternativ von (5.4) ausgeht, ergibt sich dieser Wert wie folgt:

$$\overline{x} = \left(1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{4}\right) = \frac{45}{12} = 3{,}75.$$

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 $\widetilde{x} = 3{,}5$. 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)

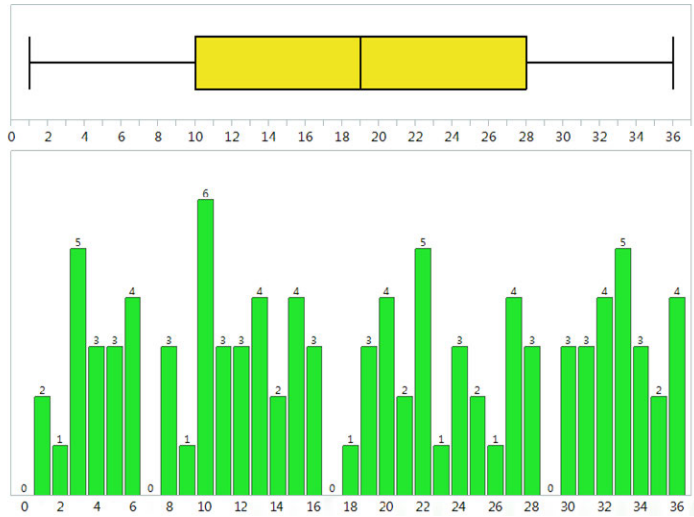
a) Die 100 Werte der Rouletteserie lassen sich aus dem Balkendiagramm im oberen Teil von Abbildung 4.12 ablesen. Der Wert 0 tritt z. B. gar nicht auf, der Ausgang 1 insgesamt zweimal, der Ausgang 2 einmal etc. Man erhält den – hier nur unvollständig wiedergegebenen – Datensatz

$$1, 1, 2, 3, 3, 3, 3, 3, 4, 4, 4, \ldots, 35, 35, 36, 36, 36, 36.$$

Für den Median folgt $\widetilde{x} = x_{0,5} = 19$, für das untere und obere Quartil $x_{0,25} = 10$ bzw. $x_{0,75} = 28$ und für den Interquartilsabstand $Q = 18$.

b) 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 $\widetilde{x} = 19$ liegt genau in der Mitte der Box.

Die folgende Grafik zeigt erneut das Balkendagramm aus Abb. 4.12, nun aber mit der dynamischen Statistiksoftware JMP erzeugt und mit zusätzlich oberhalb des Balkendiagramms eingezeichnetem Boxplot.



Kapitel 6

Lösung zu Aufgabe 6.1 (Gini-Koeffizient)

a) Man erhält für die Ordinatenwerte v_1, v_2 und v_3 mit $p_4 = 200$:

$$v_1 = \frac{p_1}{p_4} = 0{,}1; \quad v_2 = \frac{p_2}{p_4} = 0{,}35; \quad v_3 = \frac{p_3}{p_4} = 0{,}65.$$

Java-Applet „Gini-Koeffizient“

b) 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 (6.5) resultiert

$$G = \frac{1}{4} \cdot \left(\frac{2 \cdot 580}{200} - 1 \right) - 1 = 0{,}2.$$

Nach (6.8) folgt für den normierten Gini-Koeffizienten

$$G^* = \frac{4}{3} \cdot G = \frac{4}{15} \approx 0{,}267.$$

c) Der Inhalt A der markierten Fläche ist durch $A = \frac{G}{2} = 0{,}1$ gegeben.

Lösung zu Aufgabe 6.2 (Herfindahl-Index)

a) Für den Herfindahl-Index H erhält man mit $p_4 = 200$:

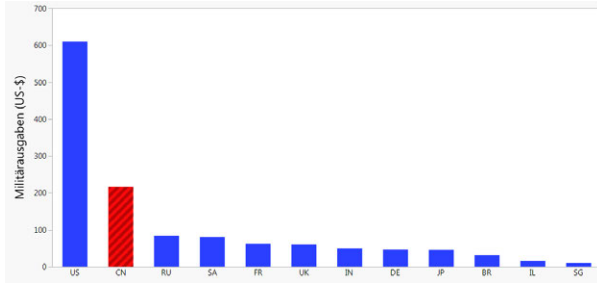
$$H = \frac{1}{p_4^2} \cdot \sum_{i=1}^{4} x_i^2 = \frac{1}{200^2} \cdot (20^2 + 50^2 + 60^2 + 70^2) = 0{,}285.$$

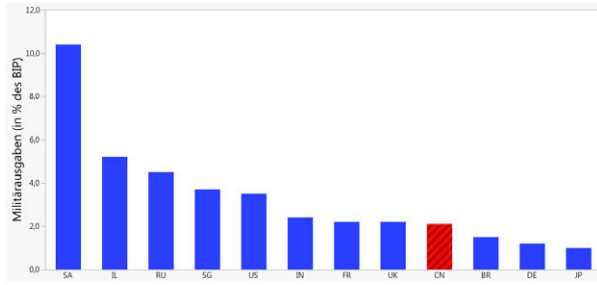
b) Der Index H kann im Fall $n = 4$ nicht kleiner als 0,25 sein.

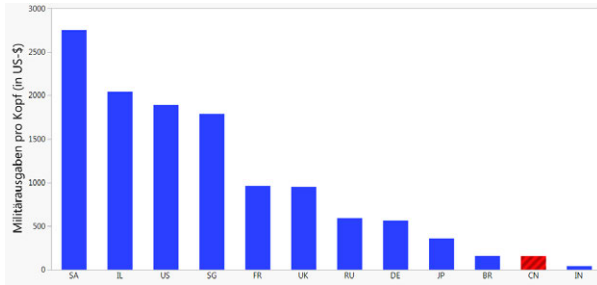
Kapitel 7

Lösung zu Aufgabe 7.1 (Militärausgaben 2014 im Ländervergleich)

Die nachstehenden, mit JMP erzeugten Grafiken zeigen die nach absteigenden Werten geordneten Balkendiagramme. Die Betonung eines Landes (hier: China) macht gut sichtbar, dass sich die drei Ranglisten deutlich unterscheiden.







Lösung zu Aufgabe 7.2 (zusammengesetzte Indexzahlen)

a) Gewichtet man die in Tabelle 7.2 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

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 7.3 (Preisindex)

Man erkennt deutlich, dass sich die Preise für Pauschalreisen stets im Juli /August und noch stärker im Dezember stark nach oben verändern, offenbar aufgrund der höheren Nachfrage nach Reisen in den Sommerferien und um die Weihnachtszeit. Da die Sommerferien von Bundesland unterschiedlich terminiert sind, verteilt sich die höhere Nachfrage hier auf einen längeren Zeitraum.

In Zeiten schwächerer Nachfrage, vor allem im Januar und November (vor und nach den Weihnachtsferien) und auch im April / Mai (nach den Osterferien), senken die Reiseveranstalter die Preise, um die Nachfrage anzukurbeln und bessere Auslastungen zu erzielen.

Lösung zu Aufgabe 8.1 (Randverteilungen)

Die Randverteilungen erhält man mit Aufsummieren der Zeilen resp. Spalten:

Kapitel 8

	b_1	b_2	Zeilensummen
a_1	62	96	158
a_2	14	188	202
Spaltensummen	76	284	n

Lösung zu Aufgabe 8.2 (Bedingte Häufigkeitsverteilungen)

Der Wert $f_X(a_5|b_1) = \frac{54}{501} \approx 0{,}108$ sagt aus, dass von den Personen in der Stichprobe, die männlichen Geschlechts ($Y = b_1$) waren, 10,8 % die Grünen favorisierten ($X = a_5$). Das Ergebnis $f_Y(b_1|a_2) = \frac{100}{217} \approx 0{,}461$ beinhaltet, dass von den Personen, die sich für die SPD ($X = a_2$) entschieden hatten, 46,1 % Männer waren ($Y = b_1$).

Lösung zu Aufgabe 9.1 (Zusammenhangsmessung; Nominalskala)

Kapitel 9

a) Man erhält mit den Werten der in Aufgabe 8.1 wiedergegebenen Vierfeldertafel bei Anwendung von (9.7) 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 (9.3)

$$\Phi = \sqrt{\frac{55{,}571}{360}} \approx 0{,}393.$$

Das Cramérsche Zusammenhangsmaß V aus (9.5) ist bei einer Vierfeldertafel wegen $M - 1 = 1$ mit dem Φ-Koeffizienten identisch, d. h. es gilt $V = \Phi \approx 0{,}393$.

Lösung zu Aufgabe 9.2 (Zusammenhangsmessung; metr. Skala)

Wenn man eine Arbeitstabelle anlegt, erhält man folgende Werte:

i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(y_i - \overline{y})(x_i - \overline{x})$
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 (9.11) liefert

$$r = \frac{\sum_{i=1}^{10} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{10} (y_i - \overline{y})^2}} = \frac{26{,}31}{\sqrt{93{,}69} \cdot \sqrt{95{,}94}} \approx 0{,}278.$$

Dieser Wert beinhaltet schwache Korrelation.

Lösung zu Aufgabe 9.3 (Zusammenhangsmessung; Ordinalskala)

Der Rangkorrelationskoeffizient kann nach (9.16) bestimmt werden, weil kein Rangplatz doppelt besetzt ist. Für die Anwendung von (9.16) sind die Rangplatzdifferenzen d_i und deren Quadrate zu ermitteln:

Mannschaft i	Rang (Hallenturnier)	Rang (Freiluftturnier)	Rangdifferenz d_i	Quadrierte Rangdifferenz 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}:

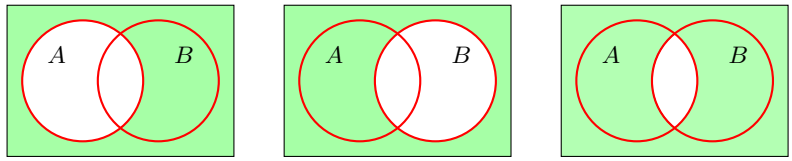
$$r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^{5} d_i^2}{5 \cdot (5^2 - 1)} = 1 - \frac{6 \cdot 8}{120} = 0{,}6.$$

21.2 Wahrscheinlichkeitsrechnung und schließende Statistik

Kapitel 10

Lösung zu Aufgabe 10.1 (Venn-Diagramme)

Nur Aussage B ist unzutreffend. Dass C und D zutreffend sind, aber B nicht richtig ist, erkennt man leichter, wenn man zunächst $\overline{A}$ und $\overline{B}$ einzeln visualisiert. Die entsprechenden Venn-Diagramme sind nachstehend an erster und zweiter Stelle wiedergegeben. Das dritte Venn-Diagramm zeigt die Vereinigungsmenge von $\overline{A}$ und $\overline{B}$. Die dort dunkel markierte Fläche stimmt nicht mit der Fläche überein, die im zweiten Venn-Diagramm der Aufgabe dunkel markiert war.



Lösung zu Aufgabe 10.2 (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 $A = \{$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 10.3 (Laplace-Wahrscheinlichkeiten)

Das Ereignis A umfasst 4 der 8 Elementarereignisse des dreifachen Münzwurfs. Jedes Tripel ist aufgrund der Annahme einer fairen Münze gleichwahrscheinlich. Es gilt also nach der Formel (10.5) für Laplace-Experimente $P(A) = \frac{4}{8} = 0{,}5$ und damit nach (10.2) auch $P(\overline{A}) = 0{,}5$.

Hinweis: 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 der Verteilungsfunktion $F(1) = 0{,}5$ der B(3;0,5)-Verteilung, d. h. es ist $P(X \geq 2) = 1 - 0{,}5 = 0{,}5$. Die Vorteile der Verwendung der Binomialverteilung werden allerdings erst bei größeren Werten von n oder im Falle $p \neq 0{,}5$ besser sichtbar.

Lösung zu Aufgabe 10.4 (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 (10.1) 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 $10^5 = 100000$.

Lösung zu Aufgabe 10.5 (Bedingte Wahrscheinlichkeiten)

a) Die Wahrscheinlichkeit $P(\overline{A})$ dafür, dass die zufällig ausgewählte Person keine Vollzeitbeschäftigung hat, ist nach (10.5) gegeben durch

$$P(\overline{A}) = \frac{160 - 64}{160} = \frac{96}{160} = 0{,}6.$$

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 vollzeitbeschäftigten Lehrpersonal zufällig ausgewählte Person promoviert ist, ergibt sich nach (10.13)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0{,}25}{1 - 0{,}6} = \frac{0{,}25}{0{,}4} = 0{,}625.$$

Dasselbe Ergebnis ließ sich auch mit (10.11) 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 (10.16) 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, bei der die Vorgaben dieser Aufgabe kursiv gesetzt sind:

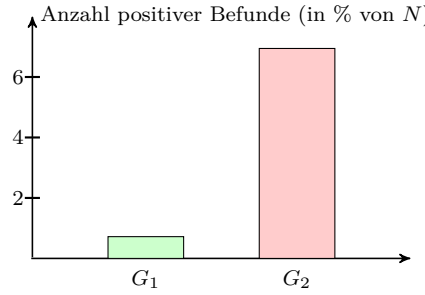
	mit Promotion (B)	ohne Promotion ($\overline{B}$)	Zeilensummen
Vollzeit (A)	*40*	24	*64*
Teilzeit ($\overline{A}$)	20	76	96
Spaltensummen	*60*	100	*160*

Die Wahrscheinlichkeit $P(\overline{A})$ ergibt sich hieraus als $\frac{96}{160} = 0{,}6$. Für die Wahrscheinlichkeit $P(A \cap B)$ erhält man sofort den Quotienten $\frac{40}{160} = 0{,}25$ und für $P(B|A)$ den Wert $\frac{40}{64} = 0{,}625$.

Lösung zu Aufgabe 10.6 (Bedingte Wahrscheinlichkeiten)

In Beispiel 1.3 wurde angenommen, dass bei einer Grundgesamtheit von N Frauen einer bestimmten Altersklasse 0,8 % Brustkrebs haben, wobei im Zuge eines Screenings 90 % dieser Fälle entdeckt werden (Gruppe G_1 mit N_1 Frauen). Es wurde ferner unterstellt, dass bei der Gruppe G_2 mit $N_2 = N - N_1$ gesunden Frauen in 7 % aller Fälle ein Fehlalarm erfolgt.

a) In der Gruppe G_1 mit $N_1 = 0{,}008 \cdot N$ erkrankten Frauen ist in $0{,}9 \cdot N_1 = 0{,}0072 \cdot N$ Fällen ein positiver Befund zu erwarten (korrekter Befund). In der Gruppe G_2, die $N_2 = 0{,}992 \cdot N$ gesunde Frauen umfasst, ist in $0{,}07 \cdot N_2 = 0{,}06944 \cdot N$ Fällen ein positiver Befund zu erwarten (falscher Befund). Nachstehend ist die relative Anzahl der positiven Befunde für beide Gruppen anhand eines Balkendiagramms visualisiert:



b) Für die Wahrscheinlichkeit $P(B)$ eines positiven Befunds bei einer zufällig aus der Grundgesamtheit G ausgewählten Frau verifiziert man

$$P(B) = \frac{0{,}0072 \cdot N + 0{,}06944 \cdot N}{N} = 0{,}0072 + 0{,}06944 = 0{,}07664.$$

c) Für die Wahrscheinlichkeit $P(\overline{A}|B)$ eines falsch-positiven Befunds ergibt sich nach dem Satz von Bayes aus (10.15)

$$P(\overline{A}|B) = \frac{P(B|\overline{A}) \cdot P(\overline{A})}{P(B)} = \frac{0{,}07 \cdot 0{,}992}{0{,}07664} \approx 0{,}906.$$

Bei ca. 90,6 % (!) aller positiven Befunde ist der Befund falsch-positiv. Vermutlich haben Sie einen weitaus niedrigeren Wert erwartet. Das Balkendiagramm macht das Ergebnis verständlicher – der zweite Balken ist etwa 9,6-mal so lang wie der erste. Wenn man die Gesamtlänge beider Balken mit 100 % ansetzt, entfallen auf den zweiten Balken 90,6 % und auf den ersten etwa 9,4 %.

d) Verwendet man anstelle von $P(A) = 0{,}008$ bei ansonsten unveränderten Vorgaben den Wert $P(A) = 0{,}006$, erhält man

$$P(B) = \frac{0{,}0054 \cdot N + 0{,}06958 \cdot N}{N} = 0{,}0054 + 0{,}06958 = 0{,}07498$$

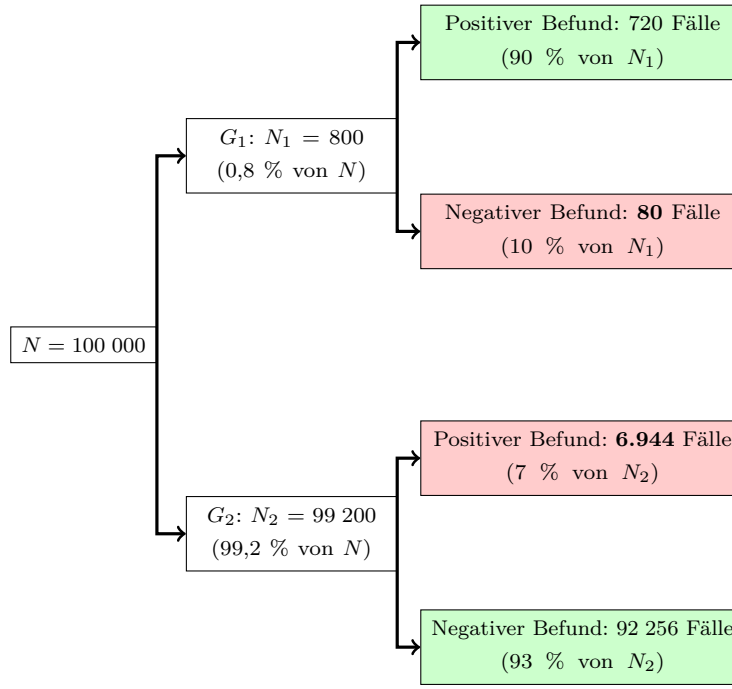
und damit

$$P(\overline{A}|B) = \frac{P(B|\overline{A}) \cdot P(\overline{A})}{P(B)} = \frac{0{,}07 \cdot 0{,}994}{0{,}07498} \approx 0{,}928.$$

Der Anteil der Fehlalarme erhöht sich nun also auf ca. 92,8 %. Zeichnete man auch hier, analog zu Aufgabenteil a, ein Balkendiagramm, wäre der zweite Balken jetzt fast 13,9-mal so lang wie der erste Balken.

Lösung zu Aufgabe 10.7 (Baumdiagramm und Vierfeldertafel)

a) Unter den in Aufgabe 10.6 genannten Voraussetzungen erhält man bei Wahl von $N = 100\,000$ das nachstehende Baumdiagramm. In der Grafik sind die Anzahlen fett betont, die als Fehlentscheidungen zu interpretieren sind (80 Fälle von Nicht-Entdeckung pro 100.000; 6 944 Fälle von Fehlalarmen pro 100 000). Diese Werte sind mit den in Teil a der Lösung zu Aufgabe 10.6 errechneten Werten kompatibel – dort war N lediglich noch nicht näher spezifiziert.



Java-Applet „Screening-risiken“

Auch die in Teil b - c der Lösung von Aufgabe 10.6 ermittelten Ergebnisse können aus der Grafik erschlossen werden. Auf 100 000 Teilnehmerinnen entfallen 720 korrekt-positive und 6 944 falsch-positive Befunde, insgesamt also 7 664 positive Befunde. Die Wahrscheinlichkeit eines positiven Befundes, gleich ob korrekt oder falsch, hat nach dem Laplace-Ansatz (10.5) den auch in Lösungsteil b errechneten Wert 0,07664, also etwa 7,7 %. Das Ergebnis zu Lösungsteil c ergibt sich analog aus dem Laplace-Ansatz, hier als Quotient der Zahlen 6 944 (Anzahl der gesunden Frauen mit positivem Befund) und 7 664 (Anzahl aller Frauen der Grundgesamtheit mit positivem Befund), d. h. als 0,906 (ca. 90,6 %).

b) Die Vierfeldertafel für die absoluten Häufigkeiten der Binärvariablen „Gesundheitsstatus“ und „Screeningbefund“ (einschließlich der Randverteilungen) sieht im Falle $N = 100\,000$ wie folgt aus:

	positiver Befund (B)	negativer Befund ($\overline{B}$)	Zeilen-summen
erkrankt (A)	720	*80*	800
gesund ($\overline{A}$)	*6.944*	92 256	99 200
Spaltensummen	7 664	92 336	100 000

Die Häufigkeiten, die sich auf Fehlentscheidungen beziehen, sind hier kursiv gesetzt (80 falsch-negative und 6 944 falsch-positive Befunde). Aus der Tabelle erkennt man unmittelbar, dass die Wahrscheinlichkeit $P(B)$ eines positiven Befunds bei einer zufällig aus der Grundgesamtheit ausgewählten Frau durch $\frac{7\,664}{100\,000} = 0{,}07664$ gegeben ist. Für die Wahrscheinlichkeit $P(\overline{A}|B)$ eines falsch-positiven Befundes erhält man sofort den Wert $\frac{6\,944}{7\,664} \approx 0{,}906$.

Lösung zu Aufgabe 10.8 (Wahrscheinlichkeit bei einem Würfelspiel)

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 einzelnen Wurf als

$$P = \left(\frac{5}{6}\right)^3 = \frac{125}{216} \approx 0{,}5787.$$

Kapitel 11

Lösung zu Aufgabe 11.1 (Würfeln mit zwei Würfeln)

a) Die Augensumme X beim Würfeln mit zwei Würfeln hat mindestens den Wert 2 und höchstens den Wert 12.Für die Eintrittswahrscheinlichkeiten $f(x)$ gilt nach (11.1)

$$f(x) = \begin{cases} \frac{1}{36} \approx 0{,}0277 & \text{für } x = 2 \text{ und für } x = 12; \\ \frac{1}{18} \approx 0{,}0556 & \text{für } x = 3 \text{ und für } x = 11; \\ \frac{1}{12} \approx 0{,}0833 & \text{für } x = 4 \text{ und für } x = 10; \\ \frac{1}{9} \approx 0{,}1111 & \text{für } x = 5 \text{ und für } x = 9; \\ \frac{5}{36} \approx 0{,}1388 & \text{für } x = 6 \text{ und für } x = 8; \\ \frac{1}{6} \approx 0{,}1667 & \text{für } x = 7; \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

Die Funktion $f(x)$ ist symmetrisch bezüglich $x = 7$.

b) Für die gemäß (11.3) 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 $F(6) = F(3) + f(4) + f(5) + f(6) = \frac{5}{12} \approx 0{,}41667$.

Interaktives Lernobjekt „Augensummen“ (mit Modell)

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 (11.13) den Wert 7.

Lösung zu Aufgabe 11.2 (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:

Interaktives Lernobjekt „Rechnen mit der Binomialverteilung“

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. Mit Tabelle 19.1 resultiert $F(1) = 0{,}7383$. Die hier 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

$$P(X \geq 2) = 1 - P(X \leq 1) = 0{,}2617.$$

b) Die Wahrscheinlichkeit $P(X = 1)$ *genau einmal* die Farbe „Grün“ zu erreichen errechnet sich als Differenz der Werte $F(1) = P(X \leq 1) = 0{,}7383$ und $F(0) = P(X \leq 0) = P(X = 0) = 0{,}3164$ der Verteilungsfunktion der genannten Binomialverteilung. Man erhält $0{,}7383 - 0{,}3164 = 0{,}4219$.

Lösung zu Aufgabe 11.3 (Hypergeometrische Verteilung)

Die Anzahl X der Richtigen beim Spiel „6 aus 45“ ist H(n; M; N)-verteilt mit $n = 6$, $M = 6$ und $N = 45$.

a) Für $\mu = E(X)$ folgt nach (11.24), dass $\mu = 6 \cdot \frac{6}{45} = 0{,}8$.

b) Die Anzahl der möglichen Ausgänge beim Spiel „6 aus 45“ ist nach Tabelle 10.1 – siehe dort den Fall „Ziehen ohne Zurücklegen und ohne Berücksichtigung der Anordnung“ – gegeben durch

$$\binom{45}{6} = \frac{45!}{39! \cdot 6!} = \frac{45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 8145060.$$

Da von den 8145060 möglichen Ausgängen, die alle gleichwahrscheinlich sind, nur ein einziger „6 Richtige“ beinhaltet, gilt nach (10.5) für die Wahrscheinlichkeit $f(6) = P(X = 6)$

$$f(6) = \frac{1}{8145060} \approx 0{,}12277 \cdot 10^{-6}.$$

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 11.3).

Lösung zu Aufgabe 11.4 (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 (11.26) 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 10.1 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 (10.5) für die gesuchte Wahrscheinlichkeit den Wert $\frac{8}{15} \approx 0{,}533$.

Kapitel 12

Lösung zu Aufgabe 12.1 (Rechteckverteilung)

Der Fahrgast 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 (12.6) durch

$$f(x) = \begin{cases} \frac{1}{20} & \text{für } 0 \le x \le 20 \\ 0 & \text{für alle sonstigen } x. \end{cases}$$

gegeben. Für den Erwartungswert errechnet man dann mit (12.12) den Wert $E(X) = 10$, der sich als mittlere Wartezeit bei zufälligem Eintreffen an der Bushaltestelle interpretieren lässt.

Lösung zu Aufgabe 12.2 (Normalverteilung)

a) Für die $N(3;4^2)$-verteilte Zufallsvariable X gilt mit (12.23)

$$P(3 \le X \le 7) = \Phi(1) - \Phi(0) = 0{,}8413 - 0{,}5 = 0{,}3413.$$

b) Mit (12.20) – (12.23) und Tabelle 19.2 folgt:

$$\begin{aligned} P(Z \le 2{,}9) &= \Phi(2{,}9) = 0{,}9981 \\ P(0 \le Z \le 2{,}3) &= \Phi(2{,}3) - \Phi(0) = 0{,}9893 - 0{,}5 = 0{,}4893 \\ P(-1{,}3 \le Z \le 0) &= \Phi(0) - [1 - \Phi(1{,}3)] = 0{,}4032 \\ P(-0{,}8 \le Z \le 0{,}8) &= \Phi(0{,}8) - [1 - \Phi(0{,}8)] = 0{,}5762 \\ P(-1{,}3 \le Z \le 1{,}2) &= \Phi(1{,}2) - [1 - \Phi(1{,}3)] = 0{,}7881. \end{aligned}$$

Interaktives Lernobjekt „Rechnen mit der Standardnormalverteilung“

Anmerkung: Wenn Sie R installiert haben, können Sie Aufgabe 12.2 besonders einfach lösen. Zur Berechnung von Wahrscheinlichkeiten des Typs (12.23) ist in der R Konsole nach dem Zeichen > nur das Kommando pnorm(b,μ,σ) – pnorm(a,μ,σ) einzugeben, wobei für den Term in Aufgabenteil a die Werte $b = 7$, $a = 3$ sowie $\mu = 3$, $\sigma = 4$ einzusetzen sind und z. B. zur Berechnung des letzten Terms in Aufgabenteil b die Werte $b = 1{,}2$, $a = -1{,}3$, $\mu = 0$ und $\sigma = 1$. Mehrere Berechnungen lassen sich zusammen ausführen. Für die beiden genannten Wahrscheinlichkeiten erhält man

```
> pnorm(7,3,4) - pnorm(3,3,4)
[1] 0.3413447
> pnorm(1.2,0,1) - pnorm(-1.3,0,1)
[1] 0.7881298
```

Mit „pnorm(.)“ ist die Verteilungsfunktion der Normalverteilung bezeichnet.

Lösung zu Aufgabe 12.3 (Normalverteilung)

a) Nach (12.21) gilt für die Verteilungsfunktion $F(x)$ der $N(2950;120^2)$-verteilten Zufallsvariablen X

$$F(x) = P(X \le 2800) = \Phi\left(\frac{2800 - 2950}{120}\right) = \Phi(-1{,}25).$$

Mit (12.20) und Tabelle 19.2 folgt:

$$\Phi(-1{,}25) = 1 - \Phi(1{,}25) = 1 - 0{,}8944 = 0{,}1056.$$

Die Wahrscheinlichkeit dafür, dass ein Neugeborenes ein Geburtsgewicht von höchstens 2800 Gramm aufwies, betrug 10,56 %.

Interaktives Lernobjekt „Standardnormalverteilung“

b) Mit (12.23) verifiziert man, dass

$$\begin{aligned} P(2800 \le X \le 3250) &= \Phi\left(\frac{3250 - 2950}{120}\right) - \Phi\left(\frac{2800 - 2950}{120}\right) \\ &= \Phi(2{,}5) - \Phi(-1{,}25). \end{aligned}$$

Erneuter Rückgriff auf (12.20) 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 2800 Gramm und 3250 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 (12.26) verknüpft. Mit $z_{0,1} = -z_{0,9} = -1{,}2816$ aus Tabelle 19.3 errechnet man den Wert

$$x_{0,1} = 2950 + z_{0,1} \cdot 120 = 2950 - 1{,}2816 \cdot 120 \approx 2796{,}2.$$

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 2796,2 Gramm.

Lösung zu Aufgabe 12.4 (Quantile)

a) 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 (12.29) und Tabelle 19.5 erhält man $t_{10;0,05} = -t_{10;0,95} = -1{,}812$.

b) 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 (12.23) 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 %) kleiner als $-2{,}228$ oder größer als 2,228 ist, beträgt die entsprechende Wahrscheinlichkeit bei einer standardnormalverteilten Zufallsvariablen nur etwa $1 - 0{,}974 = 0{,}026$, d. h. 2,6 %. Dies zeigt (vgl. auch Abbildung 12.6), dass die Dichte der Standardnormalverteilung die der t-Verteilung mit nur 10 Freiheitsgraden noch nicht gut approximiert.

Interaktives Lernobjekt „Quantile der t-Verteilung“

Kapitel 13

Lösung zu Aufgabe 13.1 (Kovarianz zweier Zufallsvariablen)

a) 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.

b) Die Kovarianz von X und Y kann nach (13.12) 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 nach (13.12) 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 schon direkt aus (13.13) erschließen können.

Kapitel 14

Lösung zu Aufgabe 14.1 (Schätzung von Erwartungswert, Varianz)

a) Ein unverzerrter Punktschätzer $\widehat{\mu}$ ist nach (14.6) durch die Ausprägung $\overline{x}$ des in (13.3) eingeführten Stichprobenmittelwerts gegeben. Man errechnet $\overline{x} \approx 66{,}792$.

b) Aus (14.9) ersieht man, dass die korrigierte Stichprobenvarianz s^{*2} aus (13.5) für die Varianz σ^2 der Normalverteilung eine unverzerrte Schätzung liefert. Die Summe (13.5) umfasst hier 24 Quadratterme, sollte also unter Verwendung einer Statistik-Software wie SPSS oder mit 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 14.2 (Konfidenzintervalle)

a) Das Konfidenzintervall ergibt sich aus (14.16) mit $\alpha = 0{,}05$ und 23 Freiheitsgraden:

$$KI = \left[\overline{X} - t_{23;0,975} \cdot \frac{S^*}{\sqrt{24}}; \overline{X} + t_{23;0,975} \cdot \frac{S^*}{\sqrt{24}} \right].$$

Interaktives Lernobjekt „Quantile der t-Verteilung“

Setzt man für $\overline{X}$ 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:

$$KI = \left[66{,}792 - 2{,}069 \cdot \frac{13{,}429}{\sqrt{24}}; 66{,}792 + 2{,}069 \cdot \frac{13{,}429}{\sqrt{24}} \right] \approx [61{,}1; 72{,}5].$$

b) 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 μ bei wiederholter Berechnung von Konfidenzintervallen in $(1-\alpha) \cdot 100$ % der Fälle von den Intervallen überdeckt wird.

Lösung zu Aufgabe 15.1 (einseitiger Gauß-Test)

a) Die Testvariable ist durch (15.2) gegeben, wobei dort $\mu_0 = 25$, $\sigma = 3$, $n = 49$ sowie $\overline{x} = 26$ einzusetzen ist. Man erhält

$$z = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \frac{26 - 25}{3} \cdot 7 \approx 2{,}333.$$

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 (15.5) aus $\beta = P(\text{Nicht-Verwerfung von } H_0 | \mu = 28)$ errechnen.

Lösung zu Aufgabe 15.2 (einseitiger Gauß-Test)

a) Im linksseitigen Gauß-Test aus Beispiel 15.3 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 Gütefunktion

$$G(\mu) = \Phi\left(-z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma} \cdot \sqrt{n}\right)$$

des Tests ein, so folgt für die Wahrscheinlichkeit $G(2{,}002)$ der Verwerfung der Nullhypothese für $\mu = 2{,}002$

$$G(2{,}002) = \Phi\left(-z_{0,95} - \frac{0{,}002}{0{,}01} \cdot \sqrt{10}\right) \approx \Phi\left(-2{,}277\right).$$

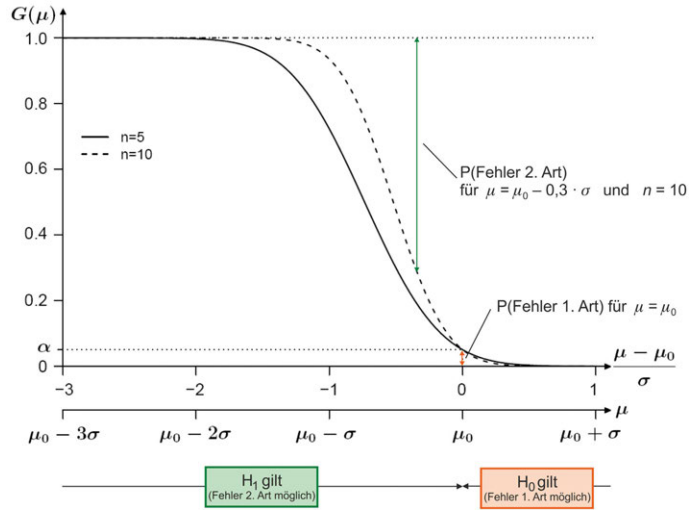
Wegen $\Phi(-2{,}277) = 1 - \Phi(2{,}277)$ folgt mit Tabelle 19.2 dann $G(2{,}002) \approx 0{,}0113$. Eine Ablehnung der Nullhypothese wäre im Falle $\mu = 2{,}002$ wegen $H_0 : \mu \geq 2$ eine Fehlentscheidung (Fehler 1. Art). Die Wahrscheinlichkeit hierfür beträgt also ca. 1,1 %.

b) Für $\mu = 1{,}997$ wäre eine Ablehnung der Nullhypothese hingegen eine korrekte Entscheidung. Sie tritt ein mit einer Wahrscheinlichkeit von

$$G(1{,}997) = \Phi\left(-z_{0,95} - \frac{-0{,}003}{0{,}01} \cdot \sqrt{10}\right) \approx \Phi\left(-0{,}696\right).$$

Mit $\Phi(-0{,}696) = 1-\Phi(0{,}696)$ und Tabelle 19.2 erhält man hier $G(1{,}997) \approx 0{,}242$. 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 Gütefunktionsverlauf für den *rechtsseitigen* Gauß-Test war im oberen Teil von Abbildung 15.4 für $n = 5$ und $n = 10$ und $\alpha = 0{,}05$ wiedergegeben. Für den *linksseitigen* Fall und mit den genannten Werten für n und α ergibt sie sich hieraus durch Spiegelung der Gütefunktion des rechtsseitigen Tests an der vertikalen Geraden $\mu = \mu_0$. Die resultierende Grafik ist nachstehend wiedergegeben. Der hier relevante Fall $n = 10$ ist durch die gestrichelte Kurve repräsentiert.



Setzt man in obiger Abbildung bei der unteren Abszissenachse speziell $\mu_0 = 2$ und $\sigma = 0{,}01$ ein, so kann man die zuvor 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 15.3 (zweiseitiger Gauß-Test)

a) Die zu testenden Hypothesen sind durch (15.1) mit $\mu_0 = 2$ gegeben. Die Ablehnung der Nullhypothese $H_0 : \mu = \mu_0$ erfolgt nach (15.3) genau dann, wenn der Betrag

$$|z| = \left| \frac{\overline{x} - 2}{0{,}01} \cdot \sqrt{10} \right|$$

der Prüfstatistik aus (15.2) den aus Tabelle 19.3 ablesbaren Wert $z_{0{,}975} = 1{,}96$ überschreitet. Mit $\overline{x} = 2{,}007$ 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 $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung zu vergleichen, nach Tabelle 19.3 also mit $z_{0{,}995} = 2{,}5758$. Der Wert dieses Quantils wird von $|z| = 2{,}2136$ nun nicht mehr überschritten, d. h. man wird hier an der Nullhypothese H_0 festhalten, also davon ausgehen, dass keine systematische Unter- oder Überschreitung des Soll-Füllgewichts vorliegt.

Kapitel 16

Lösung zu Aufgabe 16.1 (KQ-Schätzung)

Man kann analog zu Tabelle 16.2 eine Arbeitstabelle anlegen, wenn man die KQ-Schätzungen manuell und nicht – wie in Abbildung 16.3 illustriert – mit Software berechnen will. Mit $\overline{x} = 25{,}5$ und $\overline{y} = 160$ erhält man:

i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
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 $\widehat{\beta}$ von β (Steigung der Regressionsgeraden) folgt dann wegen $s_{xy} = \frac{80}{6}$ und $s_x^2 = \frac{35}{12}$ gemäß (16.6) zunächst

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{40}{3} \cdot \frac{12}{35} = \frac{32}{7} \approx 4{,}57.$$

Hieraus erhält man mit $\overline{x} = 25{,}5$ und $\overline{y} = 160$ nach (16.7) für $\widehat{\alpha}$ (Schnittpunkt der Regressionsgeraden mit der y-Achse)

$$\widehat{\alpha} = 160 - \widehat{\beta} \cdot 25{,}5 \approx 160 - 116{,}54 = 43{,}46.$$

Lösung zu Aufgabe 16.2 (KQ-Schätzung; Bestimmtheitsmaß)

a) Mit $\overline{x} = 3{,}0$ und $\overline{y} = 3{,}5$ resultiert folgende Arbeitstabelle für die manuelle Berechnung der KQ-Schätzungen:

Für die KQ-Schätzung von β folgt dann nach (16.6)

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{0{,}14375}{0{,}5125} \approx 0{,}28$$

und für die KQ-Schätzung von α mit (16.7)

$$\widehat{\alpha} = 3{,}5 - \widehat{\beta} \cdot 3 \approx 3{,}5 - 0{,}84 = 2{,}66.$$

i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$
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

b) Um das Bestimmtheitsmaß zu ermitteln, könnte man die Arbeitstabelle noch um eine Spalte $(y_i - y)^2$ erweitern. Die Spaltensumme wäre 3, d. h. es ist $s_y^2 = 0{,}375$. Nach (16.18) 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 16.3 (KQ-Schätzung)

Die Matrizen $\mathbf{X}$ und $\mathbf{X}'\mathbf{X}$ sowie der Vektor $\mathbf{y}$ haben hier die Gestalt

$$\mathbf{X} = \begin{pmatrix} 1 & 10 \\ 1 & 30 \\ 1 & 50 \end{pmatrix} \qquad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 3 & 90 \\ 90 & 3500 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix}$$

– vgl. auch (16.29) mit $n = 3$ und den Werten aus Tabelle 16.2. Mit $\alpha := \beta_0$ und $\beta := \beta_1$ folgt für die KQ-Schätzung des Vektors β der Regressionskoeffizienten

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} 3 & 90 \\ 90 & 3500 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 12 \\ 460 \end{pmatrix}.$$

Die Inversion der obigen (2×2)-Matrix kann man anhand von (18.9) durchführen oder unter Heranziehung einer Software, etwa der freien Statistiksoftware *R*. Für die Regressionskoeffizienten α und β resultieren erneut die in Beispiel 16.1 schon ohne Verwendung von Matrizen errechneten Schätzwerte $\widehat{\alpha} = 0{,}25$ und $\widehat{\beta} = 0{,}125$:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \frac{35}{24} & -\frac{3}{80} \\ -\frac{3}{80} & \frac{1}{800} \end{pmatrix} \cdot \begin{pmatrix} 12 \\ 460 \end{pmatrix} = \begin{pmatrix} \frac{35}{2} - \frac{69}{4} \\ -\frac{9}{20} + \frac{23}{40} \end{pmatrix} = \begin{pmatrix} 0{,}25 \\ 0{,}125 \end{pmatrix}.$$

22 Verzeichnisse und Internet-Ressourcen

22.1 Literaturverzeichnis

Gesamtverzeichnis

Antuofermo, M. / E. Di Meglio (2012): 23 % of EU citizens were at risk of poverty or social exclusion in 2010, *Statistics in Focus*, Ausgabe 9/2012, Eurostat, Luxemburg.

Asendorpf, J. B. / F. J. Neyer (2012): *Psychologie der Persönlichkeit*, Springer Verlag, 5. Auflage, Berlin - Heidelberg.

Bamberg, G. / F. Baur / M. Krapp (2012): *Statistik*, 17. Auflage, Oldenbourg Verlag, München.

Büning, H. / G. Trenkler (1994): *Nichtparametrische statistische Methoden*, 2. Auflage, de Gruyter Verlag, Berlin.

Burkschat, M. / E. Cramer / U. Kamps (2012): *Beschreibende Statistik – Grundlegende Methoden*, 2. Auflage, Springer Verlag, Berlin - Heidelberg.

Caputo, A. / L. Fahrmeir / R. Künstler / S. Lang / I. Pigeot / G. Tutz (2009): *Arbeitsbuch Statistik*, 5. Auflage, Springer Verlag, Berlin - Heidelberg.

Cramer, E. / U. Kamps (2008): *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik*, 2. Auflage, Springer Verlag, Berlin - Heidelberg.

Eid, M. / M. Gollwitzer / M. Schmitt (2013): *Statistik und Forschungsmethoden*, 3. Auflage, Beltz Verlag, Weinheim - Basel.

Fahrmeir, L. / T. Kneib / S. Lang (2009): *Regression – Modelle, Methoden und Anwendungen*, 2. Auflage, Springer Verlag, Berlin - Heidelberg - New York.

Fahrmeir, L. / R. Künstler / I. Pigeot / G. Tutz (2010): *Statistik*, 7. Auflage, Springer Verlag, Berlin - Heidelberg.

Gigerenzer, G. (2009): *Das Einmaleins der Skepsis*, Berliner Taschenbuch Verlag, Berlin.

Gramlich, G. M. (2014): *Lineare Algebra – Eine Einführung*, 4. Auflage, Hanser Verlag, München.

Kauermann, G. / H. Küchenhoff (2011): *Stichproben – Methoden und praktische Umsetzung mit R*, Springer Verlag, Berlin - Heidelberg.

Ligges, U. (2014): *Programmieren mit R*, 4. Auflage, Springer Verlag, Berlin - Heidelberg.

Mayer-Schönberger, V. / K. Cupier (2013): *Big Data – Die Revolution, die unser Leben verändern wird*, 2. Auflage, Redline Verlag, München.

H.-J. Mittag, *Statistik*, Springer-Lehrbuch, DOI 10.1007/978-3-662-47132-6_22

Mittag, H.-J. (2006): Earnings disparities across European countries and regions, *Statistics in Focus*, Ausgabe 7/2006, Eurostat, Luxemburg.

Mittag, H.-J. (2015): Interaktive Lernobjekte für Tablets und Smartphones, *Stochastik in der Schule*, erscheint in Heft 3/2015, Seeberger Verlag, Neuss.

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.

Randow, G. von (2006): *Das Ziegenproblem. Denken in Wahrscheinlichkeiten*, Rowohlt Verlag, Reinbek.

Rasch, D. / K. D. Kubinger (2006): *Statistik für das Psychologiestudium*, Spektrum 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 (2011): *Methoden der empirischen Sozialforschung*, 9. Auflage, Oldenbourg Verlag, München.

Sedlmeier, P. / F. Renkewitz (2013): *Forschungsmethoden und Statistik in der Psychologie*, 2. Auflage, Pearson Verlag, München.

Steland, A. (2013): *Basiswissen Statistik - Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*, 3. Auflage, Springer Verlag, Berlin - Heidelberg.

Toutenburg, H. / C. Heumann (2008): *Induktive Statistik - Eine Einführung mit 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.

Wilkinson, R./ K. Pickett (2009): *The Spirit Level – Why Greater Equality Makes Societies Stronger*, Penguin Books, London.

Wollschläger, D. (2013): *R kompakt*, Springer Verlag, Berlin - Heidelberg.

Zucchini, W. / A. Schlegel / O. Nenadic / S. Sperlich (2009): *Statistik für Bachelor- und Masterstudenten*, Springer Verlag, Berlin - Heidelberg.

22.2 Kommentierte Liste ausgewählter Lehrbücher

Lehrbücher

Die nachstehende kommentierte Lehrbuchliste soll helfen, bei Bedarf tiefer in einzelne Themenbereiche einzudringen.

- Bamberg, G. / F. Baur / M. Krapp (2012): Statistik, 17. Auflage, Oldenbourg Verlag, München.

Bewährte Einführung in die Statistik für Studierende der Wirtschafts- und der Sozialwissenschaften. Es wird auch ein Überblick über einige Spezialgebiete der Statistik gegeben (z. B. multivariate Verfahren).

- Fahrmeir, L. / R. Künstler / I. Pigeot / G. Tutz (2010): Statistik, 7. Auflage, Springer Verlag, Berlin - Heidelberg.

Didaktisch gelungene und sehr umfassende Einführung in die Statistik, die sich an Studierende der Wirtschafts- und der Sozialwissenschaften wendet.

- 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.

Solide Einführung in die beschreibende Statistik resp. die Inferenzstatistik mit Hinweisen zur Verwendung von Excel *und* SPSS*, primär für Studierende der Wirtschaftswissenschaft.*

- Schlittgen, R. (2012): *Einführung in die Statistik – Analyse und Modellierung von Daten*, 12. Auflage, Oldenbourg Verlag, München.

Sehr umfassende und fundierte Einführung in die Statistik, die die leistungsfähige Programmierumgebung R verwendet. Das Buch enthält Beispiele aus ganz unterschiedlichen Anwendungsbereichen.

- Sedlmeier, P. / F. Renkewitz (2013): Forschungsmethoden und Statistik in der Psychologie, 2. Auflage, Pearson Verlag, München.

Sehr umfassendes, auf die Psychologie zugeschnittenes Lehrbuch mit nur mäßigem Mathematisierungsgrad.

- Toutenburg, H. / C. Heumann (2009): Deskriptive Statistik – Eine Einführung in Methoden und Anwendungen mit SPSS, 7. Auflage, Springer Verlag, Berlin - Heidelberg.
- Toutenburg, H. / C. Heumann (2008): Induktive Statistik – Eine Einführung mit SPSS für Windows, 4. Auflage, Springer Verlag, Berlin - Heidelberg.

Fundierte Einführung in die beschreibende Statistik bzw. die Inferenzstatistik für Studierende der Wirtschafts- und Sozialwissenschaften.

22.3 Online-Ressourcen

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 deutlich sichtbaren Mehrwert auf. Dieser beruht darauf, dass hier direkte Verknüpfungen zu interessanten Web-Adressen sowie zu interaktiven statistischen Experimenten und tongestützten Animationen realisiert wurden. Die HTML5- oder Java-basierten statistischen Experimente und auch die Animationen (Flash) stammen aus mehreren Quellen. Zu nennen ist hier zunächst eine unter

QR-Code (deutschsprachige Statistik-App)

- https://www.hamburger-fh.de/statistik-app

zugängliche **Statistik-App** mit interaktiven Lernobjekten, die man auch über den nebenstehenden QR-Code erreichen kann.

Die Elemente der App sind HTML5-basiert und nicht nur auf Desktops, sondern auch auf Smartphones und Tablets einsetzbar. Unter

QR-Code (englischsprachige Statistik-App)

- http://www.fernuni-hagen.de/jmittag/app

oder über den zweiten QR-Code am Seitenrand findet man eine englischsprachige Fassung der Statistik-App. Letztere ist allerdings nicht so umfangreich und enthält auch keine Handhabungsanleitungen für die einzelnen Lernobjekte. Die Elemente der englischsprachigen App waren mit der 3. Auflage dieses Manuskripts verknüpft und sind nun durch die Elemente der erweiterten deutschsprachigen Version ersetzt worden.

In dieses Lehrbuch sind auch statistische Experimente auf Java-Basis aus einer unter

- http://www.fernuni-hagen.de/jmittag/bibliothek

frei zugänglichen **virtuellen Bibliothek** eingegangen. Ein Nachteil dieser Applikationen (Java-Applets) besteht darin, dass sie nicht auf Smartphones und Tablets lauffähig sind und auch bei Einsatz auf Desktops erhöhten Sicherheitsanforderungen genügen müssen, denen z. B. durch Zertifizierung Genüge getan werden kann. Einige unter

- http://www.fernuni-hagen.de/neuestatistik/applets/appletIndex.htm

eingestellte Java-Applets stammen auch aus einem älteren, vom Bundesministerium für Bildung und Forschung finanzierten **Multimedia-Projekt „Neue Statistik“**. Aus dieser Quelle stammen auch etliche Flash-Animationen, die aber ebenfalls nicht auf mobilen Endgeräten verwendbar sind.

Sonstige Online-Ressourcen

Eine Vielzahl von Statistikvorlesungen und Statistikkkursen steht heute online in der Gestalt von **MOOCs** zur Verfügung. Beispiele zur Statistik sind auf den E-Learning-Plattformen *Coursera* und *EdX* zu finden, etwa unter

- https://www.coursera.org/course/introstats;

- 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 einzelne statistische Verfahren beziehen sowie Sammlungen solcher Ressourcen. Beispiele für letztere sind die virtuellen Bibliotheken **SOCR** (*S*tatistics *O*nline *C*omputational *R*esources) und **CAUSE** (*C*onsortium for the *A*dvancement of *U*ndergraduate *S*tatistics *E*ducation), die unter

- http://www.socr.ucla.edu/
- http://www.causeweb.org

zugänglich sind. Man findet auch zahlreiche Online-Umgebungen zur dynamischen oder interaktiven Visualisierung von Daten. Genannt sei hier die schwedische *Gapminder-Stiftung*, die sich der breitenwirksamen Visualisierung von Daten der internationalen amtlichen Statistik verschrieben hat. Unter

- http://tools.google.com/gapminder/

findet man 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 jeweils die Bevölkerungsgröße wider.

Eine ebenfalls sehr ansprechende Visualisierungsumgebung für ausgewählte Daten der amtlichen Statistik wird von Google unter der Bezeichnung „Public Data Explorer“ bereit gestellt. Man findet hier u. a. Datensätze des US Census Bureau, der Weltbank oder von Eurostat, etwa zur Arbeitslosenquote in Europa oder zur Entwicklung der Staatsschulden in europäischen Ländern. Die OECD bietet ein ähnliches Visualisierungsinstrument als „Factbook eXplorer“ an.

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
- http://data.worldbank.org/indicator/SP.DYN.LE00.IN

studieren. Vor allem in afrikanischen Ländern ist die Lebenserwartung noch immer sehr niedrig, z. T. kaum über 50 Jahre.

Die freie Enzyklopädie „Wikipedia“ umfasst ein informatives Statistikportal, zugänglich via

- http://de.wikipedia.org/wiki/Portal:Statistik,

und seit Anfang 2010 gibt es unter

- http://www.tableausoftware.com/public/

das Portal *Tableau Public* für die interaktive Online-Visualisierung von Daten. Erwähnenswert 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/RSI/#?vis=nuts1.labourmarket.

Ein sehenswertes Video zum Thema „Alterung der Bevölkerung“ ist zugänglich über

- http://wisdom.unu.edu/en/ageing-societies/

und eine als „Weltuhr“ bezeichnete dynamische, permanent aktualisierte Darstellung von Schlüsseldaten unserer heutigen Welt unter

- http://www.poodwaddle.com/worldclock.swf.

Das britische Statistikamt hat alle im Land in den letzten hundert Jahren gesammelten Zensusdaten bis einschließlich des Zensus 2011 in einer ansprechenden Form über eine interaktive Animation verfügbar gemacht, optional mit Ton. Man findet die Darstellung unter

- http://www.ons.gov.uk/ons/interactive/vp1-story-of-the-census/index.html.

Gelungene interaktive Visualisierungen von Daten der deutschen amtlichen Statistik sind auch auf den Internetseiten des Statistischen Bundesamts eingestellt. Erwähnt sei nochmals das schon in Abbildung 7.2 wiedergegebene Preiskaleidoskop, das eine interaktive Visualisierung der Ausgabenanteile des Warenkorbs des „Durchschnittsverbrauchers“ bietet mit farblicher Codierung der Preisveränderungen gegenüber dem Vormonat. Erwähnt sei auch die von der Europäischen Zentralbank unter

- http://www.ecb.int/stats/prices/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/gov/html/dashboard.en.html.

22.4 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 in Tabelle 22.1 mit Aussprachehinweisen wiedergegeben.

Tab. 22.1: *Griechisches Alphabet*

Kleinbuchstabe	Großbuchstabe	Aussprache
α	A	Alpha
β	B	Beta
γ	Γ	Gamma
δ	Δ	Delta
ϵ	E	Epsilon
ζ	Z	Zeta
η	H	Eta
θ	Θ	Theta
ι	I	Iota
κ	K	Kappa
λ	Λ	Lambda
μ	M	Mü
ν	N	Nü
ξ	Ξ	Xi
o	O	Omikron
π	Π	Pi
ρ	P	Rho
σ	Σ	Sigma
τ	T	Tau
y	Y	Ypsilon
ϕ	Φ	Phi
χ	X	Chi
ψ	Ψ	Psi
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 Manuskript häufiger auftretenden Notationen zusammengestellt. Tabelle 22.2 fasst Schreibweisen für Mengen und für Operationen mit Mengen zusammen.

Symbol	Beschreibung
$\{\ldots\}$	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

Tab. 22.2: *Schreibweisen für Mengen und Mengenoperationen*

In Tabelle 22.3 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
$f : A \to B$	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 $a_1, a_2, \ldots, a_n$
$\sum_{i=1}^{\infty} a_i$	Summe der Terme $a_1, a_2, \ldots$
$n!$	Produkt $n \cdot (n-1) \cdot \ldots \cdot 1$ der n ersten natürlichen Zahlen
$\binom{n}{k}$	Binomialkoeffizient; Quotient aus $n!$ und $(n-k)! \cdot k!$
$\sqrt{a}$; $\lvert a \rvert$	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
$Y \sim \chi^2_n$; $Y \sim t_n$	Y ist χ^2_n- 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
$\chi^2_{n;p}$; $t_{n;p}$; $F_{m;n;p}$	p-Quantil der χ^2-, t- und F-Verteilung

Tab. 22.3: *Häufig verwendete Notationen der Mathematik und Statistik*

22.5 Autorenregister

22.6 Sachregister

T

W

Z