

Wolfgang Ludwig-Mayerhofer |
Uta Liebeskind | Ferdinand Geißler

Statistik

Eine Einführung
für Sozialwissenschaftler

Mit Online-
Materialien

Grundlagentexte Soziologie

Herausgegeben von
Martin Diewald | Klaus Hurrelmann

Der Juventa Verlag hat eine lange Tradition in der Publikation sozialwissenschaftlicher Texte. Bereits in den 1960er Jahren wurden mit der Reihe „Grundfragen der Soziologie“ (hrsg. von Dieter Claessens) programmatische Akzente gesetzt. Die Reihe hatte einen prägenden Einfluss auf die damals noch in den Anfängen stehende Disziplin Soziologie.

Die Reihe „Grundlagentexte Soziologie“ knüpft an diese Tradition an. Die Soziologie hat sich seitdem in Deutschland als theoretisch und empirisch reichhaltiges wissenschaftliches Fach etabliert. Es fehlt ihr aber an Einführungstexten und Übersichtsbänden für den Lehrbetrieb in Universitäten, Fachhochschulen, Fachschulen und anderen Bildungseinrichtungen.

Dieser Herausforderung stellt sich die Reihe „Grundlagentexte Soziologie“. Von fachlich gut ausgewiesenen Wissenschaftlerinnen und Wissenschaftlern werden Texte vorgelegt, die die wichtigsten theoretischen Ansätze des Faches, methodische Zugänge und gesellschaftswissenschaftliche Analysen präsentieren. Die Bände sind so zugeschnitten, dass sie sich als Basislektüre für Vorlesungen, Seminare und andere Lehrveranstaltungen mit einführendem Charakter eignen, dabei aber gleichzeitig auf der Höhe der aktuellen Entwicklung des Faches sind.

Die Reihe „Grundlagentexte Soziologie“ wird gemeinsam herausgegeben von Martin Diewald (Universität Bielefeld, Fakultät für Soziologie) und Klaus Hurrelmann (Hertie School of Governance, Berlin).

Wolfgang Ludwig-Mayerhofer |
Uta Liebeskind | Ferdinand Geißler

Statistik

Eine Einführung für Sozialwissenschaftler

BELTZ JUVENTA

Die Autorin / die Autoren

Wolfgang Ludwig-Mayerhofer, Jg. 1954, Dr. phil., ist Professor an der Philosophischen Fakultät, Seminar für Sozialwissenschaften, der Universität Siegen. Seine Arbeitsschwerpunkte sind Bildungsforschung, soziale Ungleichheit, Sozialpolitik sowie sozialwissenschaftliche Forschungsmethoden.

Uta Liebeskind, Jg. 1978, Dr. phil., ist wissenschaftliche Mitarbeiterin am Deutschen Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW) Hannover. Ihre Arbeitsschwerpunkte sind Bildungs- und Arbeitsmarktforschung, Hochschulforschung und Methoden der empirischen Sozialforschung.

Ferdinand Geißler, Jg. 1985, Dipl.-Soz., ist wissenschaftlicher Mitarbeiter am Institut für Sozialwissenschaften an der Humboldt-Universität zu Berlin. Seine Arbeitsschwerpunkte sind Methoden der empirischen Sozialforschung, Bildungsforschung sowie intergenerationale Beziehungen.

Umfangreiche Materialien zum Buch finden Sie als kostenlosen Download unter: www.beltz.de.

Bibliografische Information der Deutschen Nationalbibliothek

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

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

© 2014 Beltz Juventa · Weinheim und Basel

www.beltz.de · www.juventa.de

Druck und Bindung: Beltz Bad Langensalza GmbH, Bad Langensalza

Druck nach Typoskript

Printed in Germany

ISBN 978-3-7799-2613-9

Vorwort

Die Befassung mit Statistik ist heute fester Bestandteil jeglichen sozialwissenschaftlichen Studiums. Mit diesem Buch unternehmen wir den Versuch einer Einführung in die Statistik, die auf allzu tiefe mathematische Grundlagen verzichtet, ohne deshalb vereinfachend zu sein. Wir können und wollen Ihnen, liebe LeserInnen, also die intensive Auseinandersetzung mit der Statistik nicht ersparen, sondern Ihnen den Weg dorthin eröffnen. Zusätzlich zum Verständnis der wichtigsten Kenngrößen und Verfahren wollen wir auch deren Umsetzung mit Hilfe gängiger Statistik-Software vermitteln, nämlich Stata und IBM SPSS Statistics. Da die meisten Einführungsbücher entweder das eine (Vermittlung von statistischem Wissen) oder das andere (Umsetzung mit Software) zum Gegenstand haben, sei vorsichtshalber darauf hingewiesen, dass der Schwerpunkt dieses Buchs eindeutig auf dem ersten Aspekt liegt. Was die Anwendung der Software angeht, so setzen wir voraus, dass Sie sich hierüber irgendwo (in der Regel im Rahmen Ihres Studiums) erste Grundfertigkeiten verschafft haben; wir zeigen dann vorrangig die Umsetzung der konkreten Verfahren, die wir vorstellen.

Der Umfang des Stoffes dürfte in etwa dem entsprechen, was heute in einem sozialwissenschaftlichen Bachelor-Studium vermittelt wird. Konkret heißt das, dass wir uns ausführlich mit der univariaten und bivariaten Statistik befassen und das Buch mit einem ersten Einblick in die lineare Regression als Verfahren der Modellierung von Zusammenhängen beschließen. Dazwischengeschaltet ist eine Einführung in die wichtigsten Prinzipien und Verfahren des statistischen Schließens. Der in den Sozialwissenschaften so wichtige Bereich der multivariaten Modellierung wird also nur sehr rudimentär abgehandelt; hierzu müssen weiterführende Werke herangezogen werden.

Da ein Schwerpunkt des Buches auf der Anwendung bzw. Umsetzung der Verfahren anhand konkreter und fast durchgängig echter Beispiele liegt, stellen wir im Internet die verwendeten Datensätze und weitere Informationen bereit. Gehen Sie dazu auf die Webseite des Verlages www.beltz.de und geben Sie bei der Suche *Ludwig-Mayerhofer Statistik* ein. Auf diese Weise dürfte die passende Seite auch dann gefunden werden, wenn Webmaster ihrer Lieblingsbeschäftigung, der Änderung von URLs, nachgegangen sind. Auf dieser Webseite finden Sie alle relevanten Informationen zu Daten und sonstigen Materialien.

Wie alle mittleren und größeren Projekte im Leben hätte auch dieses nicht ohne die Hilfe zahlreicher Personen und Institutionen zu einem guten Ende gebracht werden können. In der Vorbereitungsphase half uns Alexandra

Wicht bei der Materialsammlung; darüber hinaus danken wir ihr für die Erstellung von Abbildung 5.7 sowie für hilfreiche Kommentare zu einzelnen Kapiteln und allerlei Rat in typographischen oder anderen Fragen. In einem sehr frühen Stadium der Entstehung dieses Buches hat Lena Ellenberger die Recherche nach geeigneten Daten unterstützt. Ganz grundlegend war die Hilfe von Frank Doepper, der die Umsetzung von Layout-Vorgaben mit L^AT_EX einrichtete, eine Aufgabe, die uns bei aller Liebe zu diesem famosen Satzprogramm überfordert hätte.

Dem DIW bzw. der dort angesiedelten SOEP-Arbeitsgruppe sind wir für die Überlassung eines absolut anonymisierten Datenauszugs aus dem SOEP zu Dank verpflichtet. Namentlich Jan Goebel danken wir für seine Langmut im sich über viele Monate hinziehenden Prozess der Absprachen zur Datenüberlassung. In gleicher Weise danken wir Karl Ulrich Mayer für seine Unterstützung der Idee, die GLHS-Daten für unser Buch zu verwenden, und Christian Brzinsky-Fay (WZB) für die effiziente Hilfe bei der Erstellung des Datensatzes. Dank geht außerdem an Stephanie Schneider für die Zusammenstellung der OECD-Daten.

Eva Geiger unterstützte uns bei der Erarbeitung der Software-Beispiele in SPSS. Mark Felker, Anna Frese, Carina Leser, Justyna Stasizs, Linda Schwarz und André Schulze lasen einzelne oder alle Kapitel und gaben detaillierte und hilfreiche Rückmeldungen. Nico Stawarz und Robert Paul Stephan testeten viele der Software-Beispiele. Frank Seiß achtete darauf, dass wir die Regeln der Rechtschreibung, der Grammatik und des Satzbaus einhielten (einige dezente Abweichungen gehen auf unsere eigene Kappe) und rupfte auch sonst einige sprachliche Unkräuter aus. Ihnen allen gebührt hierfür herzlicher Dank, den wir durch den Hinweis ergänzen, dass wir für mögliche Fehler in jedem Fall selbst geradestehen.

Dankbar notieren wir auch die finanzielle Unterstützung, die wir von der Universität Siegen (Bereitstellung von SHK-Mitteln) sowie vom DZHW Hannover (für das Lektorat des Textes) erhielten.

Frank Engelhardt danken wir für die stete Unterstützung im Entstehungsprozess des Buches und für seine Geduld mit der Autorengemeinschaft, die nicht wenige Deadlines zur Fertigstellung des Buches verstreichen lassen musste, weil das Leben es so wollte.

Nun sind wir gespannt auf die Rückmeldungen der Leserinnen und Leser!

Siegen, Hannover und Berlin im Mai 2014

Wolfgang Ludwig-Mayerhofer, Uta Liebeskind und Ferdinand Geißler

Inhaltsverzeichnis

1	Einleitung	11
1.1	Worum es in diesem Buch geht	11
1.1.1	Was ist Statistik?	11
1.1.2	Besonderheiten dieses Buches	13
1.2	Statistik selbst- und mitgemacht: Die Beispiele nachvollziehen	14
1.2.1	Rechnen und rechnen lassen	14
1.2.2	Zu den verwendeten Daten	18
1.3	Zum Geheimnis der Formeln	20
2	Daten, Forschungsdesigns und Stichproben	23
2.1	Daten	23
2.1.1	Skalen- oder Messniveaus	23
2.1.2	Weitere Einteilungen von Merkmalen	26
2.1.3	Begrenzte Daten	27
2.1.4	Übersicht zu Merkmalsarten	30
2.2	Forschungsdesigns und Datenauswertung	31
2.3	Datensätze	33
2.3.1	Struktur von Datensätzen	33
2.3.2	Codierung	36
2.3.3	Fehlende Datenwerte	36
3	Univariate Analyse	41
3.1	Verteilung eines Merkmals	41
3.1.1	Häufigkeitstabellen	41
3.1.2	Graphische Darstellung von Merkmalsverteilungen	45
3.2	Maße der zentralen Tendenz	58
3.2.1	Das arithmetische Mittel	58
3.2.2	Der Median	62
3.2.3	Der Modus	65
3.2.4	Lageregeln	65
3.3	Quantile und einfache Kennzahlen für die Streuung von Daten	67
3.3.1	Quantile	67
3.3.2	Der Box-and-Whisker-Plot (Boxplot)	71
3.4	Maße für Streuung und Form metrischer Variablen	73
3.4.1	Varianz	74
3.4.2	Standardabweichung	76
3.4.3	Schiefe und Wölbung	82
3.5	Empfehlungen für weiterführende Literatur	85

3.6	Software-gestützte Berechnung	85
3.6.1	SPSS	86
3.6.2	Stata	89
4	Von der Stichprobe zur Grundgesamtheit: Statistisches Schließen	93
4.1	Wahrscheinlichkeit und Wahrscheinlichkeitsverteilungen	94
4.1.1	Wahrscheinlichkeit – eine Heranführung	94
4.1.2	Zufallsvariablen	97
4.1.3	Der zentrale Grenzwertsatz und seine Folgen	114
4.1.4	Weitere Verteilungen stetiger Zufallsvariablen	116
4.2	Statistisches Schätzen	120
4.2.1	Punktschätzung	122
4.2.2	Intervallschätzung	125
4.3	Statistisches Testen	136
4.3.1	Die Grundidee von Signifikanztests	138
4.3.2	Die Praxis von Signifikanztests am Beispiel des Testens von Mittelwertunterschieden	139
4.3.3	Nichtparametrische Tests	155
4.3.4	Test eines Parameters gegen einen hypothetischen Wert	160
4.3.5	Statistisches Testen und Konfidenzintervalle	165
4.3.6	Probleme statistischen Testens	166
4.4	Komplexe Stichproben	181
4.4.1	Was sind komplexe Stichproben?	181
4.4.2	Das Problem	185
4.4.3	Zum Umgang mit komplexen Stichproben	187
4.5	Empfehlungen für weiterführende Literatur	193
4.6	Software-gestützte Berechnung	194
4.6.1	SPSS	194
4.6.2	Stata	197
5	Bivariate Analyse	201
5.1	Analyse von Kreuztabellen	202
5.1.1	Bedingte Anteilswerte	203
5.1.2	Zusammenhangsmaße	206
5.1.3	Der χ^2 -Test	212
5.2	Zusammenhang zweier metrischer Merkmale	216
5.2.1	Graphische Veranschaulichung	216
5.2.2	Kovarianz	219
5.2.3	Pearsonscher Korrelationskoeffizient	223
5.3	Zusammenhänge zwischen ordinalskalierten Merkmalen	225
5.4	Varianzanalyse	231
5.4.1	Die einfaktorielte Varianzanalyse	232
5.4.2	Inferenzstatistik: Der F-Test	236

5.5	Empfehlungen für weiterführende Literatur	238
5.6	Software-gestützte Berechnung	238
5.6.1	SPSS	238
5.6.2	Stata	240
6	Regressionsanalyse	243
6.1	Idee der linearen Regressionsanalyse	243
6.2	Aufstellen eines Regressionsmodells in der Praxis	249
6.3	Inferenzstatistik und Regressionsdiagnostik	257
6.4	Empfehlungen für weiterführende Literatur	260
6.5	Software-gestützte Berechnung	261
6.5.1	SPSS	261
6.5.2	Stata	262
	Literaturverzeichnis	265
	Index	269

1. Einleitung

1.1 Worum es in diesem Buch geht

1.1.1 Was ist Statistik?

Erste Orientierung gibt ein Lexikon. Der „Große Brockhaus“ von 1984 (Bd. 21, S. 29) sagt Folgendes: Statistik sei „... im materiellen Sinn die geordnete Menge von Informationen in Form empir. Zahlen (,Statistiken‘); im instrumentalen Sinn (Stat. Methoden) der Inbegriff der Verfahren, nach denen empir. Zahlen gewonnen, dargestellt, verarbeitet, analysiert und für Schlußfolgerungen, Prognosen und Entscheidungen verwendet werden.“

Hier wird eine wichtige Unterscheidung angesprochen: Statistik im *materiellen* Sinn ist die „geordnete Menge von Informationen“, die aus amtlichen Statistiken und vielen anderen Datenquellen als gesellschaftliche Selbstbeschreibung bekannt ist: die Arbeitslosenquote, die Bruttoverdienste aus Erwerbstätigkeit, die hergestellte Menge Bier und deren Wert oder auch die Freiland-Anbaufläche für zum Verkauf bestimmten Brokkoli im Saarland im Jahr 2009.¹

Das vorliegende Buch handelt hingegen (wie die meisten Bücher zur sozialwissenschaftlichen Statistik) von der Statistik im *instrumentalen* Sinn. Es geht also um Statistik im Sinne von Verfahren, mit denen man „empir[ische] Zahlen“ gewinnt, darstellt, verarbeitet, analysiert und zu bestimmten Zwecken verwendet. Mit empirischen Zahlen ist kein Gegensatz zu (unseres Wissens nicht existenten) theoretischen Zahlen gemeint; vielmehr soll damit zum Ausdruck gebracht werden, dass die Zahlen sich auf empirische, also in der Erfahrungswelt gegebene Phänomene beziehen. Ein in der Wissenschaft weitaus geläufigerer Ausdruck hierfür ist: *Daten*. Diese Daten kommen freilich meist in Form von Zahlen daher, was den Brockhaus wohl zu seinem Begriff der empirischen Zahlen geführt hat. Diese oft große Menge von Zahlen gilt es nun, in geeigneter Weise so zu analysieren, dass die darin enthaltene *wesentliche* Information zum Vorschein kommt. Dies geschieht durch Visualisierungen, vor allem aber mittels der Gewinnung zusammenfassender Kenngrößen. Es geht also darum, mit den Daten zu arbeiten, sie zu prüfen, zu erkunden, zusammenzufassen und schließlich so weiterzuverarbeiten, dass die relevante Information möglichst klar herausgearbeitet und

1 Die Zahlen finden Sie beispielsweise im Statistischen Jahrbuch für die Bundesrepublik 2010 auf den Seiten 74, 532, 386 sowie 350. Übrigens: Die Brokkoli-Anbaufläche im Saarland beträgt nur zwei Hektar. Zum Glück gibt es Mecklenburg-Vorpommern, dort ist die Anbaufläche ca. 240-mal so groß.

dargestellt werden kann. All dies geschieht, um aus den Daten Schlussfolgerungen über die soziale Wirklichkeit zu ziehen.

Etwas genauer gesprochen geht die Statistik im instrumentalen Sinn von folgender Ausgangslage aus:

Erstens: Wir haben es mit standardisiert erhobenen Daten zu tun. Ein anderer Ausdruck hierfür ist: Es wird gemessen. Oft führt das Messen direkt zu Zahlen, etwa bei der Angabe des Haushaltseinkommens. In anderen Fällen werden die Daten erst in Zahlen ‚übersetzt‘. Beispielsweise wird in der bekannten „Sonntagsfrage“ bei Umfragen danach gefragt, welche Partei man wählen würde, wenn am nächsten Sonntag Bundestagswahl wäre. Auch hier gibt es aber feste Vorgaben, eben die verschiedenen Parteien, die zur Wahl stehen, und für die Datenanalyse wird die Parteipräferenz meist in Form von Zahlen erfasst, etwa CDU = 1, SPD = 2 usw. Mehr dazu finden Sie in den Abschnitten 2.1.1 und 2.3.2.

Die Daten liegen *zweitens* für eine gewisse Anzahl von Untersuchungsobjekten vor; in sozialwissenschaftlichen Umfragen sind es oft Tausende von Befragten, andere Datensätze können sogar in die Zehn- oder gar Hunderttausende gehen. Aber auch kleine Datensätze von ein paar Dutzend Fällen können in der Forschungspraxis vorkommen.² In der Regel werden pro Untersuchungseinheit mehrere, oft sogar recht viele Merkmale erhoben (z. B. Meinungen zu bestimmten Themen, das Einkommen, Alter, Geschlecht, die Haushaltszusammensetzung usw.). Diese Merkmale werden in der Statistik meist als Variablen bezeichnet, da sie unterschiedliche Ausprägungen oder Werte annehmen können (eben z. B. die unterschiedlichen Parteien, die man nennen kann), also variieren. In der statistischen Analyse ist man letztendlich nicht an den einzelnen Zahlenwerten als solchen interessiert, sondern will die darin enthaltene Information zusammenfassend verdichten.

Drittens handelt es sich bei der vorliegenden Menge von Untersuchungseinheiten häufig, aber durchaus nicht immer, um eine Stichprobe, die durch Zufallsverfahren aus der Gesamtheit aller relevanten Fälle gezogen wurde.

Statistik hat nun im wesentlichen zwei Aufgaben, die sich auch als entsprechende Teilgebiete der Statistik wiederfinden lassen.

- Die *beschreibende* (oder deskriptive) Statistik hat zum Ziel, die in den Daten enthaltene Information in geeigneter Weise zusammenzufassen. Dies geschieht durch Kennzahlen, die entweder ein einzelnes Merkmal (etwa das Einkommen) oder den Zusammenhang von zwei oder mehr Merkmalen charakterisieren. Hiermit befassen sich vor allem die Kapitel 3, 5 und 6. Übrigens: Manche Autoren bzw. Lehrbücher

2 In bestimmten Fällen können auch Daten für ein einzelnes Objekt vorliegen, dann allerdings typischerweise zahlreiche Messwerte, die sich auf die Entwicklung in der Zeit beziehen, wie etwa Daten über Aktienkurse, das Wetter und Ähnliches mehr. Solche Daten werden meist mit Verfahren der sogenannten Zeitreihenanalyse untersucht, auf die wir hier nicht eingehen können.

kennen ein weiteres Teilgebiet der Statistik, die explorative Datenanalyse, die sich besonders mit der genauen Erkundung von Daten befasst. Wir stellen dieses Teilgebiet nicht gesondert vor, sondern integrieren einige Elemente (die wichtigsten sind das Stamm-Blatt-Diagramm und der Box-and-Whisker-Plot) in die deskriptive Statistik.

- Wenn es sich bei den Daten um eine Stichprobe handelt – und nur dann –, stellt sich auch die Frage, wie man anhand der Stichprobe Aussagen über die Grundgesamtheit machen kann. Ist es nicht riskant, etwa anhand einer Umfrage unter 1 000 Erwachsenen eine Aussage über die mehr als 60 Millionen Wahlberechtigten in Deutschland zu machen? Ja, es ist riskant – aber man kann das Risiko (nämlich das Risiko, sich bei der Aussage über die Grundgesamtheit zu irren) berechnen. Dies ist das Thema der *schließenden* Statistik, auch *Inferenzstatistik* genannt, die vor allem in Kapitel 4 vorgestellt, aber auch in den nachfolgenden Kapiteln immer wieder angesprochen wird.

1.1.2 Besonderheiten dieses Buches

Was erwartet Sie in diesem Buch? Welche Gründe könnte es geben, gerade dieses und nicht irgendein anderes einführendes Lehrbuch der Statistik zu lesen?

Eine *erste* Besonderheit: Wir erklären Ihnen wichtige Elemente der Statistik so, dass Sie sie selbst nachrechnen können – und zeigen Ihnen gleichzeitig, wie man die gleichen Ergebnisse mit Hilfe geeigneter Statistik-Programme erzielen kann. Dort, wo die Grenzen des Selbstnachrechnens erreicht oder überschritten werden, beschränken wir uns sogar auf den Nachvollzug mit Hilfe von Software und versuchen lieber, die Grundidee mit Worten und mit Bildern zu erklären. Allerdings bietet unser Buch keine ausführliche Einführung in die Handhabung von Statistik-Software. Aber keine Sorge: Die bekommen Sie ohnehin im Rahmen Ihres Studiums vermittelt. In diesem Buch können Sie dann aber recht schnell die Verbindung zwischen Inhalten und Umsetzung herstellen. Das ist mit den meisten anderen Büchern nicht möglich.

Hinter dieser Idee steht eine *zweite*: Wenn wir davon ausgehen, dass man komplizierte Dinge nicht wirklich selbst ausrechnen können muss, sondern sie nur so weit verstehen sollte, dass man sie mit Sinn und Verstand umsetzen kann – dann können wir auch die Konsequenz ziehen, ein paar komplizierte Dinge in dieses Buch hineinzupacken, die andere Bücher vermeiden. Wir machen das natürlich nicht, um Sie mit unnötigem Ballast zu quälen. Vielmehr geht es um Themen, die für die sozialwissenschaftliche Forschungspraxis von größter Bedeutung sind. Das wichtigste dieser Themen ist das Stichprobendesign; warten Sie dazu, bis Sie bei den Abschnitten 2.2 und vor allem 4.4 angelangt sind. In den meisten Statistik-Lehrbüchern lernen Sie ausschließlich die Grundlagen des Schließens von der Stichprobe auf die

Grundgesamtheit, für die in der Forschungspraxis jedoch häufig gar nicht die Voraussetzungen vorliegen. Nur sehr selten wird Ihnen dies dann in einer Fußnote mitgeteilt. Nun, Sie werden auch in unserem Buch diese Grundlagen lernen, weil sie eben essenziell sind. Wir werden Ihnen aber auch sagen, wo (und wie) Sie diese Grundlagen modifizieren müssen. Aus Platzgründen kann das alles nur in den Grundzügen behandelt werden; aber ganz auf diese für die Forschungspraxis so wichtigen Aspekte verzichten kann man heute nicht mehr.

In der Summe bedeutet das: Unsere Darstellung ist ziemlich stark darauf bezogen, immer direkt die praktische Relevanz der statistischen Verfahren herauszustellen – wobei „praktisch“ hier heißt: Was kann man mit den Verfahren aus Daten herausholen? Wie viele, aber doch längst nicht alle, modernen Bücher zur statistischen Datenanalyse arbeiten wir also durchgängig mit Datensätzen, die wir auch eigens für Sie aufbereitet haben und im Internet zur Verfügung stellen (genauere Angaben in Abschnitt 1.2.2).

1.2 Statistik selbst- und mitgemacht: Die Beispiele nachvollziehen

Dieses Buch lässt sich als Lehrbuch und Nachschlagewerk nutzen, es soll aber gleichzeitig auch ein Arbeitsbuch sein, mit dessen Hilfe Sie die angeführten Beispiele selbst nachvollziehen können. Unser Anspruch ist es, Ihnen zunächst jeweils die statistischen Verfahren transparent zu machen. Darüber hinaus wollen wir Ihnen auch aufzeigen, wie Sie die besprochenen Verfahren mit Statistik-Programmen praktisch anwenden können. Aus diesem Grund schließt jedes Kapitel mit einem Abschnitt zur Software-gestützten Berechnung. In diesem Abschnitt werden jeweils zentrale Beispiele des Kapitels in SPSS und Stata umgesetzt. Auf der *Webseite zum Buch* (siehe Vorwort) stellen wir Ihnen alle Datensätze zum freien Download zur Verfügung, so dass Sie jedes Beispiel aus dem Buch selbst nachvollziehen können. Sie finden dort neben den Daten auch die Umsetzung aller im Buch genutzten Beispiele, teilweise in ausführlicherer Form als hier gezeigt, für beide Software-Pakete.

1.2.1 Rechnen und rechnen lassen

Nach wie vor sind Stift und Papier (und Kopf!) die besten Utensilien, um eine Einführung in die Statistik praktisch nachzuvollziehen. Einsteiger sollten höchstens einen Taschenrechner (gern auch einfach die Handy-Taschenrechner-Funktion) zu Hilfe nehmen, und zwar dann, wenn Rechenoperationen vorzunehmen sind, die die meisten Menschen für komplexere Zahlenbeispiele nicht im Kopf lösen können (etwa das Quadrieren und Wurzelziehen). Die Beispiele zur Einführung eines Verfahrens sind sämtlich so gewählt, dass Sie sie im Kopf bzw. mit Stift, Papier und Taschenrechner

selbst nachrechnen können.³ Auch wir sind häufig so vorgegangen. Lassen Sie sich nicht von rundungsbedingten Abweichungen von den per Statistik-Software erhaltenen Ergebnissen irritieren!

Jenseits der Lernsituation wird aber niemand mehr von Ihnen verlangen, statistische Ergebnisse mit Kopf, Stift und Papier zu ermitteln. Die Datensätze, mit denen Sie in Lehre und Forschung konfrontiert werden, sind meistens sehr groß. Dies und die gebotene Genauigkeit bei der Anwendung statistischer Rechenprozeduren (auch bei kleineren Datensätzen) macht die Verwendung von Statistik-Programmen unverzichtbar. Deswegen zeigen wir für viele Auswertungsbeispiele des Buches auch, wie sie mit Hilfe von Statistik-Software zu berechnen sind. Natürlich wollen wir Sie dadurch anregen, weitere Auswertungen selbst vorzunehmen.

Zu den hier verwendeten Statistik-Programmen

Wir arbeiten in diesem Lehrbuch mit zwei der gängigsten Statistik-Pakete: Stata⁴ und IBM SPSS Statistics⁵, kurz SPSS. Damit sind zwei Programme gewählt, die in der sozialwissenschaftlichen Statistik-Ausbildung derzeit am stärksten präsent sind. SPSS war lange Zeit *das* Auswertungsprogramm schlechthin, weil es alle gängigen Verfahren zur Verfügung stellte, die in der empirischen Sozialforschung benötigt wurden. In Forschungsprojekten werden Sie mittlerweile aber häufiger Stata antreffen. Stata hat sehr schnell neuere Verfahren und komplexe Schätzverfahren aufgegriffen, die in der sozialwissenschaftlichen Forschungspraxis mehr und mehr zur Anwendung kommen. SPSS zieht bei der Implementation aktueller Analyseverfahren nach, hat aber in der sozialwissenschaftlichen Grundlagenforschung stark an Bedeutung verloren.

In der Lehre und so auch in den Computer-Pools der Universitäten ist SPSS allerdings noch recht präsent. Ein wichtiger Grund dafür ist sicherlich, dass viele Auftragsforschungsinstitute mit SPSS arbeiten. SPSS hat sich in

-
- 3 „Qualifiziertes“ Kopfrechnen ermöglichen Tabellenkalkulationsprogramme wie Calc (aus LibreOffice, Apache OpenOffice oder Ähnlichem) oder auch Microsoft Excel; die Programme lassen sich gut zum Nachvollziehen von Rechenschritten nutzen, die an allen Elementen einer Stichprobe wiederholt werden müssen. Zudem enthalten sie einige Funktionen für einfache statistische Verfahren (für eine Einführung in die statistische Datenanalyse mit Excel siehe Monka et al. 2008). Für die Auswertung großer Datensätze und die Anwendung komplizierterer Auswertungsverfahren sind diese Programme allerdings nicht geeignet, zumal da Excel einige gravierende Fehler enthält (die sich teilweise auch in den Calc-Varianten finden).
 - 4 Wir verwenden für das Lehrbuch Stata 13. Die vorgestellten Prozeduren funktionieren aber in der Regel auch in weit zurückliegenden Vorgängerversionen.
 - 5 Wir arbeiten mit Version 20, mittlerweile ist bereits 22 auf dem Markt. In den Jahren 2009 und 2010 wurde SPSS im Zuge einer Firmenumstellung unter dem Namen „PASW/SPSS“ vertrieben. In der offiziellen Bezeichnung der aktuellen SPSS-Version ist man von diesem neuen Produktnamen wieder abgerückt; der offizielle Name ist IBM SPSS Statistics. Die vorgestellten SPSS-Prozeduren funktionieren auch in den mit „PASW/SPSS“ betitelten Vorgängerversionen von IBM SPSS Statistics.

den letzten Jahren klar in Richtung Marktforschung spezialisiert. Mit SPSS umgehen zu können, heißt also auch, sich auf das Berufsleben jenseits der Grundlagenforschung vorzubereiten. Die Grundzüge des Arbeitens sind aber ohnehin bei beiden Programmen gleich, so dass man schnell von dem einen auf das andere umsteigen kann.

Arbeiten mit dem Lehrbuch und Statistik-Software

Wie schon im Vorwort erwähnt: Dieses Buch ist keine Einführung in das Arbeiten mit Statistik-Software. Wir setzen voraus, dass Sie zumindest die groben Abläufe des Umgangs mit SPSS oder Stata bereits beherrschen oder die entsprechenden Fähigkeiten parallel zum Durcharbeiten der Beispiele erwerben. Die Handhabung der Software ist nicht schwer. Man kann sie sich so vorstellen: Sie haben einen Datensatz und wollen ihn auswerten, also z. B. den Durchschnitt eines Merkmals berechnen. Dazu übermitteln Sie an die Software einen Befehl (gewissermaßen einen schriftlichen Auftrag), und die Software antwortet Ihnen, wiederum schriftlich, indem sie das gewünschte Ergebnis übermittelt. Dies geschieht in einem eigenen Fenster am PC.

Nun gibt es zwei Arten, wie man an Statistik-Software Befehle übermitteln kann:

„*Klick-Modus*“: Dieser entspricht dem heute dominierenden Umgang mit Software: Befehle werden durch Anklicken von Menü-Elementen, Pop-up-Fenstern usw. erzeugt. Wir wissen nicht, wie häufig dieser Modus in der sozialwissenschaftlichen Lehre vermittelt wird; die Dominanz entsprechender Bücher für SPSS lässt befürchten, dass das nicht ganz selten geschieht. Dennoch ist dieser Modus für seriöses wissenschaftliches Arbeiten gänzlich ungeeignet und außerdem außerordentlich umständlich und zeitraubend, jedenfalls dann, wenn man nicht nur mal kurz in einen Datensatz hineinschnuppern, sondern die Daten zumindest teilweise richtig auswerten möchte. Daher spielt er in diesem Buch keine Rolle.

„*Befehl-Modus*“: Hier benötigen Sie neben Daten- und Ausgabefenster ein drittes Fenster, in welches Sie die Befehle explizit, sozusagen im Klartext, hineinschreiben. Das geht im Grunde ganz einfach; die Befehlsprache ist geradezu militärisch knapp. Nehmen wir an, ein Lehrer möchte die Durchschnittsnote seiner Klasse in Mathematik haben. Er ruft einfach „Durchschnitt Mathenote“ – und schon liefert sein braver Diener das Ergebnis! Bei Statistik-Software rufen wir nicht, sondern schreiben, und die Software versteht nur Englisch. Daher schreiben wir etwa „mean mathenote“ (die Mathematik-Note darf einen deutschen Namen haben, nur der Befehl muss englisch sein, und zwar genau der Befehl, den die Software versteht).

Das Fenster, in das man die Befehle hineinschreibt, ist eine eigene Datei, die man speichern und wiederverwenden kann (das ist der entscheidende Unterschied zum Klick-Modus!). Diese Datei heißt in SPSS „Syntax-Datei“ oder „Syntax-File“, in Stata „Do-File“.

Für das konkrete Arbeiten mit Daten und Software raten wir zu folgendem Vorgehen:

1. Legen Sie auf Ihrem Rechner ein eigenes Verzeichnis zum Arbeiten mit dem Lehrbuch an. Wo dieses Verzeichnis liegt, ist gleichgültig.
2. Gehen Sie auf die Website www.beltz.de und geben Sie bei der Suche *Ludwig-Mayerhofer Statistik* ein. Dann werden Sie zu einer Webseite geführt, auf der Sie weitere Angaben zu den Daten finden. Eventuell müssen Sie zum Download aller Dateien noch ein oder zwei weitere Webseiten aufsuchen, die auf der Webseite beim Verlag verlinkt sind. Laden Sie nun Daten sowie gegebenenfalls weitere Materialien in das Arbeitsverzeichnis, das Sie im vorherigen Schritt angelegt hatten.

3. Schreiben Sie in Ihre Syntax-Datei bzw. Ihr Do-File einen Befehl, der auf das Arbeitsverzeichnis verweist. Ein Beispiel: Nehmen wir an, Ihr Verzeichnis heißt: C:\Users\Sibel\Documents\StatLehrbuch (so könnte ein typischer Verzeichnispfad unter *Windows* aussehen). Dann schreiben Sie einfach folgenden Befehl, der das Verzeichnis zum Arbeitsverzeichnis macht:

In SPSS: `cd "C:\Users\Sibel\Documents\StatLehrbuch"`.

In Stata: `cd "C:\Users\Sibel\Documents\StatLehrbuch"`

Bei *Mac-UserInnen* taucht am Anfang des Verzeichnispfades üblicherweise kein Laufwerksbuchstabe auf; auch verwenden sie statt des Backslashes den normalen Schrägstrich. Das gleiche gilt unter *Unix*.

Falls Sie noch zu den EinsteigerInnen gehören: Dass der SPSS-Befehl mit einem Punkt endet und der Stata-Befehl nicht, hat System: SPSS erkennt das Ende des Befehls am Punkt, Stata daran, dass die Zeile zu Ende ist. Sie können (und sollen) also alle Befehle genau so eingeben, wie Sie sie bei uns sehen.

4. *Nach* diesen Befehl schreiben Sie nun den Befehl, mit dem Sie die Daten holen. Wenn Sie z. B. mit den Daten der GLHS arbeiten wollen (mehr dazu gleich im nächsten Abschnitt), schreiben Sie

In SPSS: `GET FILE "glhsteach.sav"`.

In Stata: `use "glhsteach.dta"`

5. Nun können Sie für die Analyse direkt die Befehle verwenden, die Sie jeweils am Ende der folgenden Kapitel finden, die Sie aber auch auf der oder den Webseiten zum Buch herunterladen können.

Noch ein Hinweis zu Stata: Für manche Analysen werden Befehle benötigt, die standardmäßig nicht in Stata vorhanden sind. Hierzu werden sogenannte *user-written Ado-Files* benötigt. Dabei handelt es sich um Do-Files, welche von anderen Anwendern geschrieben worden sind, und die online kostenlos zur Verfügung stehen. Wenn für eine Analyse in diesem Buch ein solches

zusätzliches Ado-File benötigt wird, werden wir Sie jeweils darauf aufmerksam machen und erklären, wie und/oder wo das Ado-File heruntergeladen werden kann.

Zur weiteren Beschäftigung mit Stata empfehlen wir das Buch von Kohler und Kreuter (2012), das mittlerweile als das Standardeinführungswerk in Stata gilt. Das Buch führt Schritt für Schritt und anschaulich in die Datenanalyse mit Stata ein, ohne dass dabei besondere statistische Vorkenntnisse vorausgesetzt werden. Darüber hinaus finden sich im Internet sehr gute Quellen. Darunter ist zum einen der empfehlenswerte Stata-Bereich auf der Webseite der University of California, Los Angeles (UCLA) zu nennen (<http://www.ats.ucla.edu/stat/stata/>). Zum anderen finden Sie einen Stata-Guide von Wolfgang Ludwig-Mayerhofer unter <http://wlm.userweb.mwn.de/wlmstata.htm>.

Zur weiterführenden Beschäftigung mit SPSS können Sie das Buch von Akremi et al. (2011) verwenden. Zwar sind einige Elemente wirklich „für Fortgeschrittene“ (wie es im Titel heißt), aber auch Einsteiger in die statistische Datenanalyse werden von diesem Buch mehr profitieren als von den vielen teuren Büchern, die nur den weitgehend sinnlosen „Klick-Modus“ erklären. Zur statistischen Datenanalyse mit SPSS (unter Verwendung der Syntax) können Sie mit Gewinn den SPSS-Guide von Wolfgang Ludwig-Mayerhofer im Internet nutzen (<http://wlm.userweb.mwn.de/wlmspss.htm>). Die bereits oben erwähnte Webseite der University of California, Los Angeles (UCLA) enthält auch einen umfassenden Bereich zur Arbeit mit SPSS (<http://www.ats.ucla.edu/stat/spss/>).

1.2.2 Zu den verwendeten Daten

Wir arbeiten im Buch mit drei verschiedenen Datensätzen, die Ihnen über die Website zum Buch zugänglich sind. Diese Datensätze (bzw. weitere Datensätze aus diesen Quellen) finden sämtlich in der sozialwissenschaftlichen Forschungspraxis häufig Verwendung, so dass Sie ganz nebenbei auch mit der Datenstruktur einiger wichtiger Datenquellen für Sozialwissenschaftlerinnen vertraut werden.

Der SOEP-Datensatz: Der Datensatz `soep_11g` (der Nachsatz steht für die Namen von Autorin und Autoren dieses Buches) ist ein kleiner Auszug aus Daten des Sozio-oekonomischen Panels (kurz: SOEP), eines sehr wichtigen Sekundärdatensatzes in der deutschsprachigen empirischen Sozialforschung. Im SOEP werden im Längsschnitt, genauer: als Panel-Befragung, Daten zur sozialen und ökonomischen Lebenssituation von Personen und Haushalten erhoben, aber auch Einstellungen und Wertvorstellungen erfragt. Im Buch arbeiten wir vor allem mit personenbezogenen Daten. Wir haben die Daten aus Datenschutzgründen leicht verfremdet; die Variablennamen entsprechen aber bis auf zentrale Merkmale wie Geschlecht und Alter denen des *Scientific Use Files*, also des SOEP-Datensatzes, der auch der Forschungsgemeinschaft

zur Verfügung steht. Daher können Sie SOEPinfo⁶, das Informationssystem zum SOEP benutzen, um weiterführende Informationen zu den im Datensatz befindlichen Merkmalen zu recherchieren. Wir arbeiten mit Daten der Welle U des SOEP, also mit Daten, die im Jahr 2004 erhoben wurden. Die Datenstruktur des Gesamtdatensatzes ist recht komplex, was unserem fertig präparierten Lehrbuch-Datensatz nicht mehr anzusehen ist. Auf den Seiten der Arbeitsgruppe SOEP des DIW in Berlin (siehe Fußnote 6) finden Sie die vollständige Dokumentation zum SOEP.

Daten der deutschen Lebensverlaufsstudie (GLHS): Diese Studie, aus der wir den Datensatz glhsteach erstellt haben, wurde in der Zeit von den frühen 1980er Jahren bis in die 2000er Jahre hinein am Max-Planck-Institut für Bildungsforschung unter der Leitung von Prof. Karl Ulrich Mayer durchgeführt. Im Unterschied zum SOEP, wo die Längsschnittinformation nach und nach aus den jährlichen Erhebungen zusammengesetzt wird, beruht die Lebensverlaufsstudie oder German Life History Study (GLHS) auf retrospektiven Befragungen, in denen die Untersuchungspersonen möglichst genaue Angaben über ihre Bildungs-, Erwerbs- und Familiengeschichten machten. Befragt wurden etwa 8 500 Personen in West- und fast 3 000 Personen in Ostdeutschland, die Letzteren vor allem mit Blick auf ihr Leben in der DDR. Wir verwenden nur die Daten aus Westdeutschland, da die Informationen aus der DDR nicht immer unmittelbar vergleichbar sind. Die westdeutschen Daten beziehen sich auf insgesamt acht Geburtskohorten, anhand derer die Entwicklung der Lebenschancen von Menschen im historischen Verlauf nachvollzogen werden kann. Wurden in den frühen Erhebungen immer drei beisammen liegende Geburtsjahrgänge zusammengefasst (z. B. Menschen der Jahrgänge 1919–1921 oder 1929–1931), wurden später nur Personen eines einzigen Jahrgangs (z. B. 1964 oder 1971) ausgewählt. Wir haben aus dem reichhaltigen Schatz der Daten einige wenige Variablen konstruiert, die das eigentliche Ziel der Daten, Verläufe zu analysieren, zwar nicht angemessen berücksichtigen, aber doch größtenteils auf den Lebensverlauf der Menschen bezogen bleiben (z. B.: Was hat man bis zu einem bestimmten Zeitpunkt im Leben erreicht?). Zur Anonymisierung für unser Lehrbuch wurde außerdem eine 50-Prozent-Stichprobe aus der Gesamtstichprobe der Westdeutschen gezogen. Eine Veröffentlichung zu den ersten drei Kohorten stammt von Blossfeld (1987); hiervon haben wir eine Definition von „Bildungsjahren“ (Jahren, die man typischerweise für bestimmte Bildungsabschlüsse benötigt) übernommen, allerdings noch etwas verfeinert. Mehr Informationen zu den Daten findet man im Internet.⁷

6 http://www.diw.de/de/diw_02.c.222725.de/soepinfo.html

7 <http://www.mpib-berlin.mpg.de/de/forschung/beendete-bereiche/bildung-arbeit-und-gesellschaftliche-entwicklung/deutsche-lebensverlaufsstudie>

Daten der OECD: Die OECD (Organization for Economic Co-operation and Development) sammelt seit langer Zeit verschiedenste Daten, die sie in Zeitreihen über ihre Webseite zur Verfügung stellt. Die Daten beziehen sich immer auf Länder, nicht auf Individuen. Aus dem großen Datenfundus haben wir den sehr kleinen Datensatz `oecd_11g` zusammengestellt, den wir häufig heranziehen, um Ihnen Rechenwege direkt vor Augen zu führen. Klein ist der Datensatz zum einen, weil wir nur 21 Länder der westlichen Welt ausgewählt haben, zum anderen wegen der recht kleinen Zahl von Variablen, die sich etwa auf Frauenerwerbstätigkeit oder auf Kinderbetreuung beziehen. Man muss bei diesen Daten beachten, dass die OECD manchmal nachträglich noch Korrekturen durchführt, so dass es sein kann, dass die von uns zusammengestellten Daten, die sich hauptsächlich auf 2006 und 2007 beziehen, heute vereinzelt etwas andere Werte annehmen. Unsere Daten enthalten auch einige Lücken (fehlende Werte, siehe Abschnitt 2.3.3), die auf den OECD-Seiten längst geschlossen sind, die wir aber absichtlich belassen haben, um Merkmale mit unterschiedlichen Fallzahlen zur Verfügung zu haben – ein Phänomen, das ständig vorkommt, wenn man Daten auswertet. Zugang zu den Daten bekommt man auf den Internetseiten der OECD⁸; einzelne Daten zu finden, erfordert allerdings einige Geduld.

1.3 Zum Geheimnis der Formeln

Statistik hat viel mit Formeln zu tun, und das ist einer der Gründe, warum viele Studierende Vorbehalte gegen die Statistik haben. Wir verraten Ihnen hier aber gleich zu Beginn das Geheimnis der Formeln: Es gibt gar keines. Wer den grundsätzlichen Aufbau von Formeln kennt, wird Formeln lesen und verstehen können wie eine Musikerin Noten. In diesem Abschnitt wollen wir Ihnen den Aufbau von Formeln transparent machen.

Statistik ließe sich salopp auch als Wissenschaft des gekonnten Zusammenfassens bezeichnen. Um Rechenoperationen zusammenzufassen, die für alle Stichprobenelemente wiederholt werden sollen, werden griechische Großbuchstaben als Symbole benutzt. Im Rahmen dieses Lehrbuchs brauchen wir eigentlich nur das Summenzeichen \sum (das Symbol ist das große Sigma des griechischen Alphabets). Es sagt, wenn es ohne Erweiterungen benutzt wird, aus: „Summiere alles, was hinter mir steht, und zwar für alle Stichprobenelemente.“

Beispiel 1.1: Für eine Variable X , die bei drei Personen jeweils unterschiedliche Werte aufweist, nämlich den Wert 4 bei Person 1, 12 bei Person 2 und 7 bei Person 3, bedeutet

$$\sum x_i \cdot 2 \quad (1.1)$$

8 <http://stats.oecd.org/index.aspx>

nichts anderes, als dass der x -Wert jeder Person mit 2 zu multiplizieren ist, und diese Produkte dann aufsummiert werden über alle drei Personen. Schreibt man diese Rechnung explizit auf, so würde man notieren:

$$x_{\text{Person 1}} \cdot 2 + x_{\text{Person 2}} \cdot 2 + x_{\text{Person 3}} \cdot 2 = 4 \cdot 2 + 12 \cdot 2 + 7 \cdot 2 = 46$$

Für drei Personen mag diese Schreibweise noch ausreichend übersichtlich sein. Wenn aber der Stichprobenumfang 10 übersteigt, ist die extensive Schreibweise ungeeignet, weswegen wir (in Übereinstimmung mit fast allen weiteren Statistik-Lehrbüchern) mit dem abkürzenden Summen-Symbol arbeiten.

Sie haben gesehen, dass beim Notieren der Rechnung ohne das Summen-symbol kleine Subskripte aufgetaucht sind. Der Ausdruck $x_{\text{Person 3}}$ bedeutet nichts anderes als den x -Wert, den Person 3 aufweist. Die Personen, oder allgemeiner: die Merkmalsträger (Stichprobenelemente), erhalten in Formeln Indices. Der Index steht stellvertretend für die Zahl, die das Stichprobenelement erhielt, wenn man die Stichprobenelemente einmal abzählte. Der Ausdruck x_i steht also für den x -Wert des i -ten Elements in der Stichprobe.

Oft wird das Summenzeichen auch um diesen Index ergänzt. Man nennt ihn Laufindex, denn er zeigt an, über welche Stichprobenelemente die Summenbildung laufen soll. Unter dem Summensymbol wird festgehalten, bei welchem Element die Summenbildung starten soll, darüber steht, bis zu welchem Stichprobenelement der Index laufen soll, bei welchem Element also die Summenbildung beendet werden soll. Wenn wir die oben genannte Summe tatsächlich aus allen auftretenden x -Werten bilden wollen, also beginnend vom ersten Stichprobenelement bis zum letzten, kann man auch explizit notieren

$$\sum_{i=1}^n x_i \cdot 2 \quad (1.2)$$

Der Laufindex i läuft hier also von 1 bis n , also bis zum letzten Stichprobenelement, dessen Indexwert i dem Stichprobenumfang n entspricht. Möchte man hingegen die Summe im oben genannten Beispiel erst von der zweiten Person an bilden, dann notiert man

$$\sum_{i=2}^n x_i \cdot 2 \quad (1.3)$$

Nun werden, bliebe man beim Beispiel von oben, in dem für drei Personen Messwerte vorliegen, nur die x -Werte von Person 2 und 3 jeweils mit 2 multipliziert und dann addiert. Im Lehrbuch wird dieser Fall nicht auftreten, wir summieren hier immer über alle Elemente einer (Sub-)Stichprobe auf und notieren deswegen nicht immer explizit den Laufindex am Summensymbol.

Wichtig ist die Sache mit dem Laufindex aber doch, denn: Nicht nur Merkmalsträger können mit einem stellvertretenden Index adressiert werden, sondern alle in der Mehrzahl auftretenden Bestandteile statistischer Maßzahlen. So werden z. B. dort, wo es nötig ist, auch die unterschiedlichen Merkmalsausprägungen einer gruppierenden Variablen mit einem Index bezeichnet.

Beispiel 1.2: Hier sehen Sie einen Teil der Formel 5.14, die in Abschnitt 5.4.1 vorgestellt wird:

$$\sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y})^2 \quad (1.4)$$

Was diese Formel genau besagt, ist momentan nicht wichtig. Wir achten nur auf die Laufindizes i und j : Sie stehen hier für Gruppen von Stichprobenelementen (i) und für einzelne Stichprobenelemente (j); die Indices laufen jeweils von 1 bis r bzw. m , also von der ersten Gruppe bis zur letzten (wegen r = Gesamtzahl der Gruppen) und vom jeweils ersten bis zum jeweils letzten Stichprobenelement pro Gruppe (wegen m = Gesamtzahl der Stichprobenelemente pro Gruppe). Der Ausdruck $\sum_{i=1}^r \sum_{j=1}^m$ bedeutet damit: „Summiere alles, was hinter mir steht, auf für alle Stichprobenelemente pro Gruppe, und zwar von der ersten bis zur letzten Gruppe.“ Die Summenbildung läuft also wie auch in Formel 1.2 über alle Stichprobenelemente, es wird aber durch das doppelte und mit Indices versehene Summenzeichen darauf hingewiesen, dass sich die n Stichprobenelemente aufteilen in m Gruppen, wobei jede Gruppe r Stichprobenelemente enthält.

Sie sehen: i und j sind beliebige Indexbuchstaben. Sie stehen mal für Stichprobenelemente, mal für Gruppen, mal für Merkmalsausprägungen, mitunter auch für Spalten und Zeilen in einer Tabelle (siehe z. B. Abschnitt 5.1). Wofür genau sie stehen, ist immer im Kontext der Formel anzugeben.

Ein Letztes noch zum Innenleben der Formeln: Statistische Formeln sind „mehrsprachig“, wie Sie gesehen haben. In Formeln tauchen große und kleine Buchstaben des griechischen und des lateinischen Alphabets auf. Hinter der Wahl der Buchstaben verbergen sich Konventionen der statistischen Notationsweise: Meistens werden Parameter der Grundgesamtheit – das sind statistische Kennzahlen in der Grundgesamtheit – mit griechischen Kleinbuchstaben bezeichnet. Wenn man diese Kennzahlen nicht direkt errechnet, sondern schätzt – und das ist ein prominentes Unterfangen in der Statistik, siehe Kapitel 4 –, dann versieht man den griechischen Buchstaben mit einem Dach, also etwa $\hat{\mu}$ (gesprochen: „mü Dach“) als Schätzwert für den Mittelwert in der Grundgesamtheit. Als Symbole für Rechenoperationen werden griechische Großbuchstaben gewählt. Für Variablen- und Indexbezeichnungen nutzt man meist lateinische Klein- und Großbuchstaben.

2. Daten, Forschungsdesigns und Stichproben

Bevor es nun tatsächlich losgeht, sind wichtige Begriffe rund um die Eigenschaften von Daten und Merkmalen zu klären, die statistisch auszuwerten sind. Von der Art und Weise, wie statistische Daten zustande gekommen sind, hängt es nämlich ganz entscheidend ab, auf welche Weise sie ausgewertet werden können.

Dafür werden wir zunächst diejenigen Begriffe aus der grundständigen Methodenausbildung kurz wiederholen (und einige weitere einführen), die mit der statistischen Datenauswertung in direktem Zusammenhang stehen. Zunächst wird es um das Messen und Klassifizieren von Eigenschaften, später um Forschungsdesigns und ihren Einfluss auf die statistischen Auswertungsmöglichkeiten gehen.

2.1 Daten

2.1.1 *Skalen- oder Messniveaus*

Daten werden erhoben, um bestimmte theoretische Konzepte abzubilden: Um etwa das Konzept „sozioökonomische Situation eines Haushaltes“ zu bestimmen, wird man neben anderen auch das Merkmal (die Variable) „Haushaltseinkommen“ erfassen. Kurzum: Konzepte, z. B. „individuelle Gesundheit“ oder „Bildungsniveau“, werden anhand bestimmter beobachtbarer Merkmale oder Indikatoren empirisch greifbar gemacht.

Es gibt meist mehrere Möglichkeiten, ein Konzept empirisch abzubilden: Man kann das Konzept „Körper“ u. a. über den Indikator „Körpergröße“ erfassen. Die Körpergröße lässt sich abbilden, indem man Menschen in „kleine Menschen“, „mittelgroße Menschen“ und „große Menschen“ einteilt. Man kann Körpergröße aber auch ziemlich genau in „Körpergröße, gemessen in cm“ abbilden. Wie Merkmale empirisch erfasst sind, drückt sich aus im Skalenniveau eines Merkmals. Das Skalenniveau, oft auch als Messniveau bezeichnet, legt fest, welche Aussagen über unterschiedliche Ausprägungen eines Merkmals zulässig sind. Dabei gilt, dass auf höherem Skalenniveau auch die Aussagen zulässig sind, die man über unterschiedliche Merkmalsausprägungen auf niedrigerem Skalenniveau machen kann. Umgekehrt ist das nicht möglich. Das heißt etwa im Beispiel der Körpergröße: Wenn zwei paarweise Messungen zur Körpergröße vorliegen – Person A_1 ist „klein“, Person A_2 ist „mittelgroß“, Person B_1 ist 164 cm groß, Person B_2 ist 185 cm

groß –, dann lässt sich zwar jeweils sagen: Die zweite Person ist größer als die erste, allerdings lässt sich nur für B_1 und B_2 zusätzlich angeben, *um wie viel* die zweite Person größer ist als die erste.

Das Skalenniveau eines Merkmals ist unmittelbar relevant für statistische Auswertung: Von ihm hängt ab, welche statistischen Kennzahlen sich für das Merkmal bzw. die Merkmale bestimmen lassen und welche Analyseverfahren man einsetzen darf. Deswegen werden die Skalenniveaus im Folgenden kurz eingeführt.

Nominalskalierte Merkmale: Merkmale, die auf Nominalskalenniveau gemessen wurden, lassen nur Aussagen über die Unterschiedlichkeit von Merkmalsträgern zu. Weist Person A die Merkmalsausprägung „kroatisch“ auf, Person B „griechisch“, dann wissen wir, dass beide sich im Merkmal „Staatsangehörigkeit“ unterscheiden. Um ein „Mehr“ oder „Weniger“ kann es hier nicht gehen; das Merkmal „Staatsangehörigkeit“ lässt sich weder in eine Rangordnung bringen noch gar quantifizieren.

Ordinalskalierte Merkmale: Das Wort „ordinal“ verweist auf den Begriff der Ordnung; Messung auf Ordinalskalenniveau ermöglicht es also, die Messwerte nach der Stärke einer bestimmten Eigenschaft zu ordnen, erlaubt aber keine Aussage über den genauen Abstand zwischen den Messwerten. Der Idealfall von Ordinalskalen sind Rangskalen, die eine mehr oder weniger eindeutige Reihung aller Merkmalsträger erlauben. Das klassische Beispiel sind Ranglisten im Sport; vereinzelt kommt es hier vor, dass zwei Personen oder Mannschaften genau gleich gut sind und daher auch den gleichen Rangplatz zugesprochen bekommen, aber das ist die Ausnahme.

Viel häufiger sind allerdings ordinalskalierte Merkmale, bei denen die Untersuchungseinheiten (meist Personen) sich selbst auf einer Skala mit mehreren, meist zwischen vier und elf, Ausprägungen einstufen. Das Merkmal „Zufriedenheit mit dem Haushaltseinkommen“, das im SOEP in elf Schritten zwischen „niedrig“ und „hoch“ gemessen ist, ist ein Beispiel. Jemand, der in der Befragung einen Wert am linken Ende der Skala, also in der Nähe von „niedrig“ ankreuzt, ist weniger zufrieden mit seinem Haushaltseinkommen als jemand, der eine Antwortmöglichkeit weiter rechts, also näher an „hoch“ wählt. Um *wie viel* mehr jemand zufrieden ist, der statt des Punktes in der Mitte der Skala den Punkt rechts neben der Skalenmitte ankreuzt, lässt sich allerdings auf Basis dieser Messung nicht sagen.

Intervallskalierte Merkmale: Bei intervallskalierten Merkmalen sind die Abstände zwischen zwei Ausprägungen sinnvoll zu interpretieren, nicht aber ihr Verhältnis zueinander. Letzteres liegt daran, dass der Wert 0 bei diesen Merkmalen nicht identisch ist mit dem kleinsten möglichen Wert. Das bekannteste Beispiel ist die Temperatur in Grad Celsius: Man kann nicht sagen, dass es an Tagen mit einer Temperatur von 20° doppelt so warm

ist wie an Tagen mit 10° . Wohl aber kann man sagen, dass die Differenz zwischen 10° und 20° genauso groß ist wie die zwischen 20° und 30° .

Beispiel 2.1: Ein Beispiel für ein intervallskaliertes Merkmal ist das Berufprestige nach Treiman, das in den GLHS-Daten verwendet wird. Der niedrigste Wert, den es hier laut Treiman (1977: 172) gibt, ist -2 . Man kann also nicht sagen, dass ein Prestigewert von 60 doppelt so groß ist wie ein Wert von 30.

Ratio- oder verhältnisskalierte Merkmale: Hier sind nicht nur die Unterschiede zwischen Messwerten inhaltlich interpretierbar, sondern auch Verhältnisse. Die Bedingung hierfür ist einfach: Die Skala muss einen absoluten Nullpunkt haben, unterhalb dessen kein Messwert liegen kann. Ein typischer Fall ist das Einkommen: Menschen können kein Einkommen erzielen und weisen dann beim Einkommen einen Wert von 0 auf; weniger gibt es nicht. Ein Kind, das monatlich 30 € Taschengeld erhält, erhält eineinhalbmal so viel wie ein Kind, das 20 € Taschengeld bekommt.

Intervallskalierte + verhältnisskalierte Merkmale = metrische Merkmale: Für die angewandte Statistik ist die Unterscheidung zwischen intervall- und ratioskalierten Merkmalen meist irrelevant. Beide werden unter dem Begriff *metrische Merkmale* zusammengefasst. Viele statistische Verfahren wurden zuerst für solche Merkmale entwickelt; ganz zentrale Konzepte wie etwa das arithmetische Mittel oder die Varianz (die im nächsten Kapitel erläutert werden) setzen voraus, dass die Merkmale auf metrischem Niveau vorliegen. Der entscheidende „Graben“, wenn man so will, verläuft also zwischen metrischen Variablen und Variablen niedrigeren Skalenniveaus.

Ordinale und metrische Merkmale in der Forschungspraxis: Da in den Sozialwissenschaften Merkmale, die bei strenger Betrachtung nur ordinalskaliert sind, recht häufig vorkommen, gibt es lange Debatten und unterschiedliche Auffassungen darüber, wo genau die Grenze zwischen ordinalen und metrischen Merkmalen verläuft. Likert-skalierte Merkmale¹, die formal betrachtet ordinalskaliert sind, werden in der Analyse oft wie metrische Merkmale behandelt. Meist wird dabei argumentiert, dass die Abstände zwischen den Merkmalsausprägungen als gleich groß (äquidistant) betrachtet werden können. Infolgedessen werden für solche Merkmale z. B. Mittelwerte berechnet,

1 Als likert-skalierte Merkmale werden polytome Merkmale (siehe S. 27) mit mindestens fünf, oft auch sieben oder mehr Ausprägungen bezeichnet, die Intensitäten von Zustimmung zu einer vorgegebenen Aussage messen (z. B. „stimme gar nicht zu“ bis „stimme völlig zu“, oder „trifft gar nicht zu“ bis „trifft voll und ganz zu“). Genau genommen ist die Likert-Skala die *Summe* von Zustimmungswerten zu mehreren solcher Items, die sämtlich ein und dasselbe theoretische Konstrukt messen. Zu Likert-Skalen siehe Schnell et al. (2011: 178 ff.).

auch wenn das für ordinale Merkmale eigentlich nicht zulässig ist. Ein Freibrief zur metrischen Verwendung eigentlich ordinaler Merkmale ist das jedoch nicht: Nicht jedes auf einer fünfteiligen ordinalen Skala klassifizierte Merkmal kann als metrisch betrachtet bzw. wie ein metrisches Merkmal behandelt werden. Ob ein ordinale Merkmal wie ein metrisches Merkmal verwendet werden kann, ist also von Fall zu Fall genau abzuwägen. Manche Sozialwissenschaftler halten dies grundsätzlich für eine Todsünde; wir gehören nicht zu diesen.

Beispiel 2.2: Das Merkmal „Ausgeglichen in den letzten vier Wochen“ im SOEP-Datensatz erfasst, wie oft sich eine befragte Person in letzter Zeit ausgeglichen fühlte. Die Merkmalsausprägungen lauten „immer“ (1), „oft“ (2), „manchmal“ (3), „fast nie“ (4) und „nie“ (5); für diese Ausprägungen stehen im Datensatz die in Klammern genannten Zahlen. Dieses Merkmal (das man auf den ersten Blick leicht mit einem likert-skalierten Merkmal verwechseln könnte) wie ein metrisches Merkmal zu behandeln, wäre fatal, denn: Um wie viel häufiger genau sich jemand ausgeglichen fühlte, der die Frage mit „oft“ beantwortet hat, im Vergleich zu jemandem, der die Frage mit „manchmal“ beantwortet hat, ist völlig unklar. Und anzunehmen, dass der Unterschied von „manchmal“ zu „oft“ ungefähr die gleiche Steigerung an empfundener Ausgeglichenheit bedeutet wie von „oft“ zu „immer“, ist auch nicht plausibel. Damit hat z. B. die Bestimmung eines Durchschnittswerts für dieses Merkmal in der Stichprobe keinen Sinn.

2.1.2 Weitere Einteilungen von Merkmalen

Im Folgenden werden einige übliche Unterscheidungen von Merkmalen eingeführt. Dies geschieht allerdings eher zu nomenklatorischen Zwecken, damit Sie wissen, wovon die Rede ist, wenn diese (häufig verwendeten) Begriffe an anderer Stelle auftauchen. Die Verfahren, die in diesem Buch behandelt werden, werden vor allem durch die bereits eingeführten Skalenniveaus voneinander unterschieden.²

Metrische Merkmale: Stetig oder diskret? Metrische Merkmale werden unterteilt in stetige und diskrete Merkmale. Stetige Merkmale werden auf einem Kontinuum gemessen und können dort (u. U. innerhalb bestimmter Grenzen) unendlich viele Werte aufweisen. Das Merkmal „Körpergröße“ ist ein stetiges Merkmal: Es kann auf einem Kontinuum von ca. 45 bis 251 cm theoretisch jeden denkbaren Wert annehmen.³

2 Für multivariate Analyseverfahren, die nur ausblickartig Gegenstand dieses Buches sind (siehe Kap. 6), sind die folgenden Unterscheidungen jedoch häufig relevant.

3 Der größte derzeit lebende Mensch ist laut Guinness World Records 251 cm groß (siehe <http://www.guinnessworldrecords.com/tallest-man-living/>, Zugriff am 18.04.2014).

Diskrete Merkmale hingegen können innerhalb ihres Wertebereichs nur bestimmte Werte annehmen. Das trifft auf alle Zählraten zu, Daten, die sich eben aus dem Abzählen von Gegenständen, Ereignissen usw. ergeben. Die Zahl der Kinder unter 16 Jahren im Haushalt ist ein Beispiel: Es können nur 0, 1, 2 etc. Kinder im Haushalt leben, nicht aber 2,333 oder 1,7 Kinder.

Zwischen der Natur eines Merkmals und seiner Messung ist zu unterscheiden. Durch den Vorgang des Messens werden alle Merkmale mehr oder weniger diskret, da Messungen immer nur in einem bestimmten Genauigkeitsgrad stattfinden. Die Körpergröße etwa wird meist nur in ganzen Zentimetern angegeben. Zwei Personen mit der Merkmalsausprägung „164 cm“ für Körpergröße sind der Messung nach gleich groß, faktisch ließen sich durch eine noch genauere Messung wahrscheinlich kleine, aber irrelevante Unterschiede feststellen. In der Regel werden Variablen, die „ihrer Natur nach“ stetig sind und für die eine ausreichend große Menge verschiedener Messwerte vorliegt, auch in der Datenanalyse als stetig behandelt.

Dichotome und polytome Merkmale: In der Gruppe der nominalen und ordinalen Merkmale wird unterschieden zwischen dichotomen und polytomen Merkmalen. Dichotome Merkmale sind Merkmale, deren Merkmalsraum in zwei Ausprägungen erschöpft ist. Das Geschlecht wird in den meisten Datensätzen nur dichotom erfasst. Oder: In vielen breit zugänglichen Datensätzen wird die Staatsangehörigkeit aus Datenschutzgründen nur als „deutsch“ oder „nicht deutsch“ ausgewiesen.

Polytome Merkmale haben – das ist denen, die ein wenig Griechisch können, jetzt schon klar – mehr als zwei Merkmalsausprägungen. Die auf einer zehnteiligen Skala klassifizierte allgemeine Lebenszufriedenheit ist ein polytomes Merkmal.

2.1.3 Begrenzte Daten

Manchmal haben Variablen weitere Eigenschaften, deren Bedeutung man gerade am Anfang u. U. noch nicht erkennen kann (meist werden sie auch erst in komplexen Analysen relevant). Auf einige dieser Eigenschaften wollen wir Sie aber schon hier hinweisen.

Das Problem kann man auf den allgemeinen Begriff der „begrenzten Daten“ bringen. „Unbegrenzt“ sind nur *stetige Daten* ohne (relevante) Begrenzung des Wertebereichs nach oben oder unten. In Lehrbüchern über begrenzte Daten (klassisch ist Maddala 1983) spielen dann beispielsweise nominal- und ordinalskalierte Merkmale eine zentrale Rolle, auch wenn sie oft (so auch bei Maddala) als „qualitative“ Variablen einen (auch begrifflichen) Sonderstatus haben. Aber auch metrische Daten können in ihrem Wertebereich begrenzt sein. Einige dieser Beschränkungen seien hier zunächst kurz genannt:

- Daten weisen oftmals eine Grenze auf, oberhalb oder unterhalb derer keine Werte liegen *können*. Denken wir an ratioskalierte Variablen: Diese können keine Werte annehmen, die kleiner als 0 sind. Die möglichen Prozentwerte in unserem OECD-Datensatz sind außerdem (wie Prozentwerte überhaupt) nach oben begrenzt. Viele statistische Verfahren setzen aber voraus, dass zumindest der „abhängigen Variablen“ (der Variablen, deren Werte erklärt werden sollen; siehe Kasten 5.1 auf S. 202) keine Grenzen nach oben oder unten gesetzt sind.
- Datenwerte können auch (gegebenenfalls zusätzlich zum gerade genannten Problem) „intern“ beschränkt sein, indem sie nur bestimmte Werte annehmen können. Beispielsweise ist die Zahl der Kinder, die Frauen gebären können, relativ gering; vor allem kann die Kinderzahl nur ganzzahlige Werte annehmen. Die schon erwähnten „Zählraten“, also Daten, die sich auf die Anzahl von Dingen oder Vorkommnissen (z. B. Zahl der Unfälle, die jemand erleidet, Zahl der Patente, die einer Firma in einem Jahr erteilt wurden) beziehen, spielen hier eine besondere Rolle, weil sie oft charakteristische Verteilungen aufweisen (so sind sie fast nie normalverteilt, ein Begriff, den wir in Abschnitt 3.4 kennenlernen werden). In den GLHS-Daten ist beispielsweise das Merkmal „Zahl der Geschwister“ enthalten.

Wie schon gesagt sind diese Eigenschaften, denen man in der Forschungsliteratur immer wieder begegnet, für eine Einführung in die Statistik noch nicht wichtig. Anders ist das mit einer dritten Beschränkung, die wir unter dem etablierten Begriff *Zensurierung* diskutieren (was aber nichts mit „Zensur“ im üblichen Sinn zu tun hat).

Zensierte Daten

Beginnen wir gleich mit zwei in der Forschungspraxis wichtigen Beispielen:

Erstens: In den letzten Jahrzehnten ist in den Sozialwissenschaften das Interesse an Prozessen oder Abläufen, generell: an zeitlichen Verläufen enorm angestiegen. Eine Möglichkeit, Prozesse zu untersuchen, ist diese: Man untersucht, wie lange Personen oder andere Untersuchungseinheiten in einem bestimmten Zustand verbleiben, oder anders formuliert: wie lange es dauert, bis ein bestimmtes Ereignis eintritt, das einen Zustand beendet. Man fragt also beispielsweise, wie lange es dauert, bis Ehen geschieden werden (Eintritt des Ereignisses „Scheidung“) oder bis Arbeitslose wieder eine Beschäftigung aufnehmen.

Zunächst scheint es so, als handelte es sich hier um ganz „normale“ metrische Daten. In der Praxis ist man aber oft mit dem Problem konfrontiert, dass nicht für alle Personen die Dauer bis zum Eintritt des den Ausgangszustand beendenden Ereignisses bekannt ist. Beispielsweise werden längst nicht alle Ehen geschieden, nicht alle Arbeitslosen nehmen eine Beschäftigung auf. Oft reicht auch die zur Verfügung stehende Beobachtungsdauer nicht aus, um abzuwarten, bis alle Ereignisse eingetreten sind.