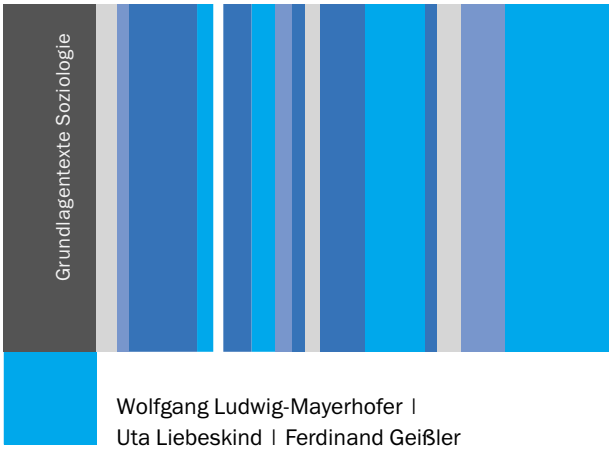

Wolfgang Ludwig-Mayerhofer |
Uta Liebeskind | Ferdinand Geißler

Statistik

Eine Einführung
für Sozialwissenschaftler

Mit Online-Materialien

BELTZ JUVENTA

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: Stata[4] und IBM SPSS Statistics[5], 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 Befehlssprache 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 Laufwerkbuchstabe auf; auch verwenden sie statt des Backslashs 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_llg` (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 Forschergemeinschaft

zur Verfügung steht. Daher können Sie SOEPinfo[6], 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_llg` 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 OECD[8]; 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

$$\sum x_i \cdot 2 \tag{1.1}$$

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 Summensymbol 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 erhielte, 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

$$\sum_{i=1}^{n} x_i \cdot 2 \tag{1.2}$$

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

$$\sum_{i=2}^{n} x_i \cdot 2 \tag{1.3}$$

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.