Leseprobe aus:

Steffen-M. Kühnel, Dagmar Krebs

Statistik für die Sozialwissenschaften

Mehr Informationen zum Buch finden Sie auf rowohlt.de.

1 Einführung

1.1 Statistik und Sozialwissenschaften

Gibt es Unterschiede im Wahlverhalten von Männern und Frauen? Hat die Ungleichheit in Deutschland in den letzten 15 Jahren abgenommen, zugenommen, oder ist sie gleichgeblieben? Wie bewerten die Bürger die Europäische Union? Dies sind Beispiele für Fragestellungen, die sich nicht ohne statistische Analysen beantworten lassen. Für die Sozialwissenschaften ist Statistik eine Hilfswissenschaft, die für die Analyse empirischer Daten benötigt wird. Die Datenanalyse ihrerseits hat die Funktion, eine Verbindung zwischen der Realität und den theoretischen Überlegungen in den Sozialwissenschaften herzustellen.

Ob Aussagen über die Realität relevante Informationen liefern, hängt von der jeweiligen inhaltlichen Fragestellung ab. So dürfte die Haarfarbe von Befragten in den meisten sozialwissenschaftlichen Untersuchungen irrelevant sein. Wenn es jedoch in einer empirischen Untersuchung beispielweise um Modephänomene ginge, wäre es denkbar, dass die Frage, ob die Haare einer Person in einer bestimmten Farbe getönt sind, eine bedeutsame Information darstellt. Es ist also eine inhaltliche Frage, ob die Informationen, die für eine statistische Datenanalyse gesammelt werden, tatsächlich von Belang sind oder nicht. Die Statistik kann hierüber keine Aussagen machen. Diese prinzipielle Einschränkung sollte stets beachtet werden. Auch eindrucksvolle Tabellen und schöne Grafiken können völlig belanglos sein.

Selbst wenn die vorliegenden Daten relevante Informationen für eine Fragestellung enthalten, kann die Datenanalyse in die Irre führen. Dies zeigt ein Beispiel aus der Umfrageforschung: In einer für die Bundesrepublik repräsentativen Umfrage aus dem Jahre 1991 war ein deutlicher Zusammenhang zwischen der Haltung zum Schwangerschaftsabbruch und dem Vorhandensein ei-

nes Telefonanschlusses in der Wohnung beobachtbar. Bei Befragten ohne Telefonanschluss trat eine liberale Haltung zum Schwangerschaftsabbruch deutlich häufiger auf als bei Befragten, die über einen Telefonanschluss verfügten. Begünstigt Telefonbesitz eine restriktive Haltung gegenüber Schwangerschaftsabbruch? Der Zusammenhang zwischen Telefonanschluss und Haltung zum Schwangerschaftsabbruch war in Deutschland im Jahr 1991 ein empirischer Fakt. Trotzdem ist die Schlussfolgerung, dass Telefonbesitz die Haltung zum Schwangerschaftsabbruch beeinflusst oder umgekehrt die Haltung zum Schwangerschaftsabbruch die Bereitschaft fördert, sich an das Telefonnetz anzuschließen, offensichtlich unsinnig. Der empirische Zusammenhang kam dadurch zustande, dass 1991 im Gebiet der ehemaligen DDR die Telefondichte wesentlich geringer war als im Westen und dass gleichzeitig bei Befragten im Osten häufiger eine liberale Haltung zum Schwangerschaftsabbruch zu beobachten war als bei Befragten im Westen.

Allein mit den Methoden der Statistik kann die fehlerhafte Interpretation dieser empirisch beobachtbaren Beziehung zwischen Telefonbesitz und Haltung zum Schwangerschaftsabbruch als Ursache-Wirkungs-Beziehung nicht ausgeschlossen werden. Zwar verschwindet der Zusammenhang zwischen Telefonbesitz und Haltung zum Schwangerschaftsabbruch, wenn die Analyse in den neuen und in den alten Bundesländern getrennt durchgeführt wird. Man muss aber erst einmal auf die Idee kommen, eine getrennte Analyse für Ost und West zu rechnen. Zudem muss in den Daten die Information enthalten sein, ob eine Person aus dem Osten oder aus dem Westen kommt.

Dieses Beispiel zeigt die zentrale Bedeutung inhaltlicher Überlegungen. Genau dies ist gemeint, wenn vom *Primat der Theorie* über die Empirie bei der Erhebung und der statistischen Analyse von Daten gesprochen wird. Empirische Daten als solche sind zwar Fakten. Daten können aber niemals für sich sprechen. Sie müssen stets interpretiert werden. Dies geht nur vor

dem Hintergrund bereits vorliegenden Wissens aus der Literatur und den Forschungsergebnissen empirischer Studien oder auf der Basis von Hypothesen, die aus bisher nicht widerlegten Theorien abgeleitet werden können, und ist zudem nur mit dem prinzipiellen Risiko einer Fehlinterpretation möglich.

1.2 Warum Statistik?

Wozu, so mag man sich dann fragen, benötigen wir überhaupt Statistik, wenn es doch primär auf inhaltliche Überlegungen ankommt und das Risiko eines Fehlschlusses nicht auszuschließen ist? Hierzu zwei Antworten:

1. Es ist zum einen die reine Informationsmenge, die dazu zwingt, Statistik anzuwenden. Wenn in einer Umfrage z. B. 3000 Personen befragt werden und für diese Personen jeweils viele Eigenschaften erfasst werden, dann kann diese Informationsmenge nicht mehr ohne statistische Hilfsmittel bewältigt werden.
2. Datenanalyse bedeutet immer Informationsverdichtung und damit gleichzeitig Informationsverlust. Es muss daher stets eine Entscheidung darüber getroffen werden, *wie* die Informationsverdichtung erfolgen soll. Die Statistik liefert hierzu Entscheidungsregeln. Ohne Beachtung dieser Regeln besteht die Gefahr, durch unangemessene Informationsverdichtung der Daten unzutreffende Schlussfolgerungen zu ziehen.

1.3 Statistische Modelle

Für die statistische Analyse von Daten gibt es eine Vielzahl unterschiedlicher Methoden. Statt von Analysemethoden zu sprechen, bevorzugen wir die Bezeichnung *statistisches Modell*. In einem Modell werden bestimmte Eigenschaften hervorgehoben; andere bleiben dagegen unberücksichtigt. Eine Straßenkarte ist beispielweise ein Modell eines geographischen Gebiets. Die

Karte weist die Straßenführung aus. Die Höhe der an einer Straße liegenden Gebäude oder die Anzahl der in dem Gebiet lebenden Menschen wird dagegen ignoriert. Dies ist sinnvoll, weil Straßenkarten über die Lage von Straßen informieren sollen und nicht über die Höhe von Gebäuden oder die Anzahl der Menschen. Analog informiert ein statistisches Modell über bestimmte Eigenschaften von Daten. Mit Hilfe eines solchen Modells kann z. B. der Zusammenhang zwischen Geschlecht und Wahlverhalten beschrieben werden. Dabei ist es irrelevant, ob es Herr Meier ist, der die CDU wählt, oder Herr Müller. Entscheidend ist allein, dass die Personen «Meier» und «Müller» männlich sind und nicht weiblich. Der individuelle Lebenshintergrund von Herrn Müller oder Herrn Meier ist für die Fragestellung nicht von Interesse. Das statistische Analysemodell abstrahiert daher von den konkreten Personen, über die Daten vorliegen. Dies bedeutet nicht, dass individuelle Erfahrungen in der Statistik grundsätzlich ausgeblendet werden. So werden in statistischen Modellen der Ereignisanalyse biografische Erfahrungen von Personen explizit berücksichtigt und es kann z. B. untersucht werden, ob und welche Auswirkungen die Erfahrung von Arbeitslosigkeit auf die nachfolgende Erwerbstätigkeit hat.

In der Statistikausbildung wird oft zwischen beschreibender (deskriptiver) und schließender (Inferenz-)Statistik unterschieden. Mit Hilfe der deskriptiven Statistik werden Aussagen über vorliegende Daten gemacht. In der Inferenzstatistik wird dagegen von beobachteten Eigenschaften in einer Stichprobe auf die unbekannten Eigenschaften in einer Grundgesamtheit geschlossen, wobei es vor allem um die Minimierung der dabei auftretenden Fehlschlussrisiken geht. Da die Vermeidung von Fehlschlüssen aber auch in der deskriptiven Statistik zentral ist, wird in dieser Einführung nicht strikt zwischen beschreibender und schließender Statistik unterschieden.

1.4 Aufbau des Buches

Die Statistik ist so umfangreich, dass wir in dieser Einführung nicht alle Bereiche der Statistik behandeln können. Wir konzentrieren uns in diesem Lehrbuch zum einen auf Grundlagen, die die Basis für die multivariaten Analysemodelle bilden, die in der sozialwissenschaftlichen Forschung eingesetzt werden. Zum anderen liegt der zweite Schwerpunkt bei der Vorstellung von Anwendungsmöglichkeiten linearer und nichtlinearer Regressionsmodelle. Die Regressionsanalyse ist nicht nur die am häufigsten eingesetzte Analysestrategie in den Sozialwissenschaften. Auch sehr viele spezielle Analysemodelle wie die erwähnte Ereignisanalyse oder auch die Mehrebenenanalyse oder die Panelanalyse sind Regressionsmodelle, die Besonderheiten von Datenkonstellationen berücksichtigen. Andere Modelle wie etwa die Faktorenanalyse, Skalierungsmethoden oder die Analyse von Strukturgleichungsmodellen mit latenten Variablen basieren auf regressionsanalytischen Vorstellungen. Kenntnisse der Regressionsanalyse bilden daher eine fruchtbare Basis für die Einarbeitung in weitere Analysemodelle.

Im *Teil A* (Kapitel 2 und 3) werden zunächst die *Grundlagen* der statistischen Datenanalyse vorgestellt. Ausgangspunkt ist die Erfassung von Informationen in einer Datenmatrix (Kapitel 2). Dabei werden die wichtigen Begriffe «Variable», «Ausprägung» und «Realisation» eingeführt sowie Eigenschaften *univariater Verteilungen* und deren Aufbereitung in Tabellen und Grafiken vorgestellt. Univariat bedeutet, dass jeweils nur ein Merkmal betrachtet wird, z. B. das Alter in einer Gruppe von Befragten. Relativ knapp wird die Operationalisierung theoretischer Konzepte und das Problem des Messens diskutiert.[1] Kapitel 3 behandelt

1 Die in der empirischen Sozialforschung eingesetzten Methoden der Operationalisierung und Messung sind Gegenstand von Einführungen in die empirische Sozialforschung: Diekmann (2012), Kromrey (1998) oder Schnell et al. (2011); speziell zur Durchführung sozialwissenschaftlicher Umfragen mit Interviews siehe Schnell (2012). Zur Messtheorie siehe Luce et al. (1990), Orth (1974), Savage (1992).

die Verdichtung von Information durch das Zusammenfassen von Realisierungen univariater Verteilungen zu *Kennwerten*.

Teil B (Kapitel 4–6) beschäftigt sich mit der Beziehung zwischen *Population* und *Stichprobe*. Anhand des Beispiels einer Zufallsauswahl von zwei aus sechs Haushalten werden die Idee des Zufallsexperiments und die Eigenschaften von Wahrscheinlichkeiten vorgestellt (Kapitel 4). Im Zusammenhang mit der Zufallsauswahl von Stichproben aus einer Population werden die Beziehungen zwischen den empirischen und bekannten Stichprobenkennwerten, den unbekannten Populationskennwerten und den *Wahrscheinlichkeitsverteilungen von Stichprobenkennwerten* bei einfachen Zufallsauswahlen aufgezeigt. In Kapitel 5 geht es um die *Kennwerteverteilungen von Häufigkeiten*, in Kapitel 6 um die *Kennwerteverteilungen von Anteilen und Mittelwerten*. Die drei Kapitel des zweiten Teils bilden damit die theoretische Basis für alle Anwendungen, die in den nachfolgenden Kapiteln diskutiert werden.

In *Teil C* (Kapitel 7–8) wird das *Schätzen* von Populationseigenschaften mit Hilfe von Stichprobendaten und das *Testen* statistischer Hypothesen vorgestellt. An die Beschreibung der Gütekriterien für Schätzer schließt sich die Konstruktion von Konfidenzintervallen für die Schätzung von Populationswerten (Mittelwerten und Anteilen) sowie für die Schätzung von Mittelwert- und Anteilsdifferenzen in der Population (das entspricht einem Gruppenvergleich) bei unabhängigen und abhängigen Messungen an (Kapitel 7).

Die Logik und Vorgehensweise beim statistischen Hypothesentesten ist Gegenstand von Kapitel 8. Als Anwendungsbeispiele werden hier Unterschiede zwischen Populationsanteilen und -mittelwerten in zwei Gruppen geprüft.

In *Teil D* (Kapitel 9–11) wird der Zusammenhang zwischen zwei Variablen mit Hilfe der *Tabellenanalyse* untersucht. Bei allen Zusammenhangsanalysen geht es um die Fragen, ob überhaupt ein Zusammenhang besteht, welches Muster der Zusam-

menhang hat und schließlich wie stark ein Zusammenhang ist. Ausgangspunkt ist der einfachste Fall eines Zusammenhangs zwischen zwei *dichotomen Variablen* mit jeweils zwei Ausprägungen (Kapitel 9). Die Zusammenhangsanalyse wird hier als bivariate gemeinsame Häufigkeitsverteilung in einer Vierfeldertabelle eingeführt, die in Kapitel 10 auf bivariate Kreuztabellen mit *kategorialen Variablen* verallgemeinert wird. Dabei wird zwischen der Analyse von nominalskalierten Variablen und der Analyse ordinaler Variablen unterschieden. Hier wird auch die Logik hierarchischer Modelltests an Beispielen demonstriert.

Kapitel 11 stellt die Logik der *Drittvariablenkontrolle* vor, bei der es darum geht, festzustellen, ob und wie sich der Zusammenhang zwischen zwei Variablen bei Kontrolle einer weiteren (dritten) Variable verändert. Da diese Veränderungen inhaltlich oft nur dann interpretierbar sind, wenn der datengenerierende kausale Prozess bekannt ist, werden die möglichen Zusammenhangskonstellationen linear-additiver Modelle oder linearer Modelle mit Interaktionseffekt hier zunächst an Hand konstruierter Daten diskutiert. Den Abschluss des Kapitels bildet die Drittvariablenkontrolle bei einem empirischen Beispiel.

In *Teil E* (Kapitel 12–16) wird die *multiple Regression* vorgestellt. Kapitel 12 beginnt mit den symmetrischen Zusammenhangsmaßen der Kovarianz und Korrelation zwischen zwei metrischen Variablen, um dann auf die Analyse gerichteter Zusammenhänge mittels (zunächst) bivariater Regression einzugehen. Die Drittvariablenkontrolle in der trivariaten und der Übergang zur multiplen Regression ist Gegenstand von Kapitel 13. Dabei wird auch die Möglichkeit der Modellierung und Analyse nichtlinearer Zusammenhänge mit multiplen Regressionsmodellen diskutiert.

Während Kapitel 14 die für inferenzstatistische Schätzungen und Hypothesentests notwendigen Anwendungsvoraussetzungen der multiplen Regression beschreibt und die Berechnung von Konfidenzintervallen und Anwendung von Tests in Regressions-

analysen vorstellt, zeigt Kapitel 15 in unterschiedlichen stufenweisen Anwendungen die Flexibilität und Vielseitigkeit des Analyseinstruments der multiplen Regression. Ein Schwerpunkt liegt hier bei der Berücksichtigung von nominalskalierten erklärenden Variablen mit *Dummy-Variablen* als Prädiktoren.

In Kapitel 16 werden varianzanalytische Modelle vorgestellt. Die im Kontext der Analyse experimenteller Daten entwickelte Varianzanalyse kann als ein spezielles lineares Regressionsmodell verstanden werden. Varianzanalysen mit Messwiederholungen und mit zufälligen Effekten können zudem als Basis für die Modellierung von Paneldaten und Mehrebenenproblemen dienen und erleichtern so den Einstieg in diese speziellen Analysemodelle.

In *Teil F* (Kapitel 17–18) wird die Regressionsanalyse mit der Einführung *nichtlinearer Regressionsmodelle* auf nichtmetrische abhängige Variablen erweitert. Zunächst werden in Kapitel 17 mit der *binären logistischen Regression* und der *binären Probit-Regression* zwei nichtlineare Regressionsmodelle vorgestellt, mit denen die Effekte von metrischen und kategorialen erklärenden Variablen auf eine dichotome abhängige Variable untersucht werden können. In Kapitel 18 erfolgt die Verallgemeinerung auf die Analyse polytomer Variablen mit der *multinomialen logistischen Regression* und der *konditionalen logistischen Regression* für nominalskalierte abhängige Variablen. Mit der ordinalen Logit- und Probit-Regression können schließlich auch Effekte auf ordinale Variablen untersucht werden.

Auch wenn Regressionsanalysen und darauf aufbauende multivariate Analysemodelle das Standardinstrument der sozialwissenschaftlichen Datenanalyse bilden, gibt es eine Vielzahl weiterer Modelle für spezifische Datenkonstellationen. Als Abschluss dieser Einführung werden in *Teil G* (Kapitel 19) *nichtparametrische und verteilungsfreie Tests* am Beispiel von Gruppenvergleichen und voraussetzungsarme Signifikanztests vorgestellt. Anhand computerunterstützter Permutationstests und Boot-Straps

zeigen wir, wie Simulationsmethoden für inferenzstatistische Fragestellungen genutzt werden können.

1.5 Hinweise zum Arbeiten mit diesem Buch

Dieses Lehrbuch ist als Einführung konzipiert, die sich an BA- und MA-Studierende, aber auch an Doktoranden sowie Forscherinnen und Forscher in den Sozialwissenschaften richtet. Zunächst werden grundlegende Kenntnisse über die Prinzipien der Ziehung von Zufallsstichproben, über das damit verbundene Schätzen von Populationswerten und die Prüfung statistischer Hypothesen vermittelt. Ohne Kenntnis der entsprechenden Kapitel ist der nachfolgende Stoff nicht verständlich. Vor allem ist dann die korrekte Anwendung statistischer Analyseverfahren nicht möglich. Daher bilden neben den grundlegenden Kapiteln 2 und 3 die Inhalte der Kapitel 4 bis 8 die Basis der statistischen Datenanalyse. Diese Kapitel und die einfachen Grundlagen der statistischen Zusammenhangsanalyse (Kapitel 9 bis 11) gehören nach unserer Vorstellung zu einer Einführung in die Statistik im Rahmen eines sozialwissenschaftlichen BA-Studiums. Bei einem methodenorientierten BA-Studium ist zusätzlich die Einbeziehung der Kapitel 12 bis 14 sinnvoll.

Die Kapitel 15 bis 19 bieten Hilfestellungen für die Praxis der Datenanalyse. Sie eignen sich daher vor allem für die Vertiefung statistischer Kenntnisse in einem MA-Studiengang. Diese Kapitel sind allerdings auch für Promotionsstudierende geeignet, da insbesondere die Kapitel 15, 17 und 18 sehr anwendungsbezogen sind und Anregungen zur statistischen Auswertung eigener Forschungsergebnisse bieten. Die in diesen Kapiteln enthaltenen Hinweise auf mögliche Analysestrategien schließen auch Hinweise auf deren Grenzen ein, da neben der reinen Anwendungstechnik vor allem die Sorgfalt im Umgang mit den Daten und die angemessene Interpretation der erzielten Ergebnisse zu realistischen und tragfähigen Ergebnissen führt. Aus diesem Grund

werden alle statistischen Verfahren an Beispielen demonstriert, wobei gerade auf die zuletzt genannten Punkte geachtet wird.

Da die Darstellung der Regressionsanalyse und deren vielfältige Anwendungsmöglichkeiten breiten Raum einnimmt, können weitere Analyseverfahren nicht mehr vorgestellt werden. Da die Regressionsanalyse aber Ausgangspunkt für viele dieser hier nicht behandelten statistischen Analysestrategien ist, erscheint uns die Intensität, mit der die Regression behandelt wird, gerechtfertigt. Die hier bereitgestellten Grundlagen ermöglichen es den Studierenden, sich andere Analyseverfahren zu erschließen. Auch Spezialbereiche wie z. B. Schätzverfahren für korrekte Standardfehler bei komplexen Stichproben, Panelanalyse, spezielle Erhebungsdesigns und vieles andere mehr haben wir nicht behandelt, weil es den Rahmen einer Einführung sprengen würde. Weiterhin haben wir nicht die Verfahren behandelt, die vor der eigentlichen Datenanalyse liegen, wie z. B. Skalierungsverfahren und Faktorenanalysen.

Die im Buch enthaltenen Datenbeispiele zur Verdeutlichung der Logik und Vorgehensweise in der statistischen Datenanalyse beziehen sich auf unterschiedliche Stichproben in verschiedenen Jahrgängen der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (Allbus). Die Datensätze und Fragebögen sind beim GESIS-Datenarchiv in Köln (http://www.gesis.org) erhältlich. Die SPSS bzw. STATA-Syntax-Dateien zur Berechnung der Beispiele befinden sich auf der Homepage des Methodenzentrums Sozialwissenschaften der Universität Göttingen (http://www.uni-goettingen.de/de/87777.html). Dort finden sich auch aktuelle Informationen (Beispiele, ergänzende Hinweise, Errata) zu diesem Buch.

Neben dem Nachvollziehen der Beispiele ist vor allem das Einüben der Logik der vorgestellten statistischen Verfahrensweisen wichtig, um diese auf eigene Forschungsfragen anwenden zu können. Zu diesem Zweck haben wir ein Übungsbuch, die «Aufgabensammlung», zu diesem Lehrbuch verfasst, dessen Struktur

dem Aufbau des Lehrbuchs entspricht. In diese Aufgabensammlung mit kommentierten Lösungen haben wir sog. Berechnungsformeln aufgenommen, die im Lehrbuch aus Platzgründen nicht enthalten sind, die aber das Berechnen von Statistiken mit dem Taschenrechner erleichtern. Weiterhin sind Rechenregeln zum Umgang mit dem Summenzeichen sowie zum Logarithmieren und Potenzieren in die Aufgabensammlung aufgenommen worden.

In einigen Kapiteln dieses Buches gibt es Abschnitte, die als *Hinweise* gekennzeichnet sind. Diese Abschnitte enthalten statistische Anwendungen (nicht Beispiele), Herleitungen, Vertiefungen und Diskussionsanregungen.

Wenn möglich, werden bei der Einführung statistischer Tests die biografischen Daten ihrer Urheber genannt, die wir zumeist Wikipedia – Die freie Enzyklopädie (de.wikipedia.org) entnommen haben.