Lehr- und Handbücher der Statistik

Herausgegeben von
Universitätsprofessor Dr. Rainer Schlittgen

Deskriptive und Explorative Datenanalyse

Von
Univ.-Prof. Dr. Siegfried Heiler
und
Dr. Paul Michels

R. Oldenbourg Verlag München Wien

Anschrift der Verfasser:

Prof. Dr. Siegfried Heiler
Fakultät für Wirtschaftswissenschaften und Statistik
Universität Konstanz, Postfach 5560, D-7750 Konstanz

Dr. Paul Michels
A. C. Nielsen Marketing Research GmbH,
Ludwig-Landmann-Str. 405, D-60486 Frankfurt/Main

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Heiler, Siegfried:
Deskriptive und Explorative Datenanalyse / von Siegfried
Heiler und Paul Michels. – München ; Wien : Oldenbourg, 1994
(Lehr- und Handbücher der Statistik)
ISBN 3-486-22786-6

NE: Michels, Paul:

Gesamtherstellung: R. Oldenbourg Graphische Betriebe GmbH, München

ISBN 3-486-22786-6

Vorwort

Es ist schon immer, seit praktische Statistik betrieben wird, üblich gewesen, statistische Daten durch geeignete graphische Darstellungen zu veranschaulichen, in Tabellen zu verdichten und schließlich durch einfache Kenngrößen zu charakterisieren. Dies gehört zu dem Aufgabenbereich der deskriptiven Statistik. Durch die Verdichtung des Datenmaterials wird ein Verlust an Einzelinformation in Kauf genommen zugunsten eines Gewinns an Aussagekraft.

Während für lange Zeit, ausgehend von der englischen Schule, die am Wahrscheinlichkeitsmodell orientierte mathematische Statistik das Bild unseres Faches geprägt hat, stellt man, etwa zu Beginn der 60er Jahre, eine Renaissance der datenorientierten Betrachtungs- und Vorgehensweise fest. Als richtungsweisend wird in diesem Zusammenhang gern ein Artikel von John Wilder Tukey aus dem Jahre 1962 genannt mit dem Titel "The Future of Data Analysis". Zum Aufgabengebiet der deskriptiven Statistik, nämlich der Zusammenfassung, der Konzentration und Präsentation von Daten, kommt das Ziel hinzu, in den Daten vorhandene (häufig versteckte) Strukturen und etwaige Auffälligkeiten, Anomalien, aufzudecken. Dazu wird das Bündel der zur Verfügung stehenden Methoden in vielfältiger Weise erweitert. Eine wesentliche Rolle spielt dabei die ständig wachsende Einsatzmöglichkeit immer leistungsfähiger werdender Computer. Sie erlaubt früher unmöglich gewesene Kalküle mit hohem Komplexitätsgrad durchzuführen und eröffnet, was vielleicht noch interessanter ist, der graphischen Analyse und Darstellung neue Wege.

Diese Renaissance der Datenanalyse wird geprägt durch John W. Tukey in den USA und J.-P. Benzécri in Frankreich. Das erste Buch auf diesem Gebiet war die 1971 erschienene erste vorläufige Version der "Exploratory Data Analysis" von J.W. Tukey. Die endgültige Fassung

ist 1977 erschienen. Das für den französischen Bereich grundlegende Werk "L'Analyse des Données" von J.-P. Benzécri et collaborateurs kam 1973 heraus. Es besteht aus den beiden Bänden "La Taxonomie Numérique" und "L'Analyse des Correspondences".

Die Feststellung, daß die Schwerpunkte der amerikanischen und der französischen Richtung verschieden sind, dürfte nicht überraschen. Zu den Schwerpunkten der Analyse des Données gehören die Clusteranalyse, die Hauptkomponenten- und die Faktorenanalyse, die Korrespondenzanalyse und die automatische Klassifikation. Diese Aspekte werden in dem vorliegenden Buch nicht behandelt mit Ausnahme der Hauptkomponentenanalyse, auf die im Zusammenhang mit der Diskussion der Streuung multivariater Daten eingegangen wird.

Die Einstellung der Amerikaner wird reflektiert in einer Umschreibung, die David Andrews in einer Enzyklopädie gegeben hat: Explorative Datenanalyse ist die Bearbeitung, Zusammenfassung und Darstellung von Daten, um sie dem menschlichen Verständnis zugänglicher zu machen. Auf diese Weise sollen vorhandene Strukturen in den Daten und bedeutsame Abweichungen von diesen Strukturen aufgedeckt werden.

Obwohl gewisse methodische Entwicklungen eine bedeutsame Rolle spielen, kann die EDA nicht als eine Methodensammlung verstanden werden. Es ist vielmehr eine Einstellung und eine Flexibilität im Umgang mit den Daten. So können bestimmte Verfahren sowohl in der explorativen als auch in der konfirmatorischen (d.h. klassich mathematisch-statistischen) Analyse Anwendung finden. Verschieden ist dann lediglich die Art der Interpretation der Ergebnisse.

Im Gegensatz zur schließenden Statistik (der konfirmatorischen Datenanalyse CDA), baut die EDA nicht auf einem vorher formulierten Wahrscheinlichkeitsmodell auf, in dem Annahmen gemacht und Hypothesen formuliert werden. Die EDA beginnt vielmehr mit einem Studium der Daten. Werden dabei nichttriviale Strukturen gefunden, so kann man versuchen, diese durch ein statistisches Modell zu beschreiben. Dabei sollten jedoch stark einschränkende Modellannahmen vermieden werden. Deshalb ist die Anwendung robuster (Ausreißer-resistenter) Schätzverfahren geboten.

Typisch für konfirmatorische Datenanalyse ist das Schema der testenden Statistik. Danach steht am Anfang eine Hypothese. Zu deren Überprüfung wird - etwa über einen geeigneten Versuchsplan - eine Zufallsstichprobe durchgeführt, und darauf wird dann ein Test angewandt. Das Ergebnis ist eine Aussage. In der Praxis ist die Situation meist nicht so einfach. "Hypothesen fallen nicht vom Himmel" (Victor). Häufig entstehen Fragen aus der Beschäftigung mit Daten. Und Fragestellungen, die aus wissenschaftlichen oder praktischen Problemen entstehen, sind im allgemeinen so vage, daß sie sich nicht direkt in ein präzises Wahrscheinlichkeitsmodell umsetzen lassen, wie es die Anwendung eines Tests erfordert.

"Finding the question is often more important than finding the answer", lautet einer der markanten Sätze von J.W. Tukey.
Explorative Verfahren geben durch die Suche nach Auffälligkeiten Anstöße zur Bildung von Hypothesen und Modellen und sie helfen bei der Präzisierung der Fragestellung im Sinne eines statistischen Tests.

Das Buch ist entstanden aus Unterlagen zu einer Lehrveranstaltung Statistik I, die in Konstanz für Studenten der Volkwirtschaftslehre und der Verwaltungswissenschaft regelmäßig angeboten wird. Ein wesentliches Anliegen bei diesen Veranstaltungen bestand darin, die Studenten auch mit einigen der Gedanken, Konzepten und Verfahren vertraut zu machen, die in der explorativen Datenanalyse entstanden sind und das Instrumentarium der deskriptiven Statistik beträchtlich erweitert haben. Der Stoff wurde erweitert um einige Aspekte, die wegen zeitlicher Restriktionen keinen Platz in den Lehrveranstaltungen finden konnten. Dazu kamen weiterführende Ergänzungen zu einigen der angesprochenen Analyseverfahren. Soweit das Verständnis solcher Ergänzungen an etwas tieferliegende mathematische Vorkenntnisse geknüpft ist, wurden die entsprechenden Abschnitte mit drei Sternen gekennzeichnet. Sie können bei einem ersten Durcharbeiten des Stoffes überschlagen werden. Allerdings gehen die für das Verständnis notwendigen Anforderungen an keiner Stelle über gewisse Kenntnisse der linearen Algebra, insbesondere des Vektor- und Matrizenkalküls, hinaus. Außerdem sollen die überall eingestreuten und durchgerechneten Beispiele aus verschiedenen Anwendungsbereichen das Verständnis des Stoffes erleichtern und zum Selbststudium anregen.

Die breite Streuung der Beispiele bringt zum Ausdruck, daß die vorgestellten Methoden in den unterschiedlichsten Anwendungsbereichen und Fachrichtungen einsetzbar sind. So werden Datensätze aus den Bereichen Demographie, Geschichte, Hydrologie, Kriminologie, Landwirtschaft, Marketing, Medizin, Meteorologie, Sozialwissenschaften, Technik, Wirtschaftswissenschaften und der Umweltforschung verwendet und analysiert.

Wir danken allen Mitarbeitern der Fachgruppe Statistik in Konstanz, die in unterschiedlicher Weise zur Fertigstellung dieses Buches beigetragen haben. Für die Erstellung wesentlicher Teile der Textvorlage in LaTeX danken wir Jeanette Blings, Cristina Fernández de Geiselhart, Katharina Kohn und insbesondere Sonja Schneider, die auch die Koordination und Zusammenstellung übernommen hat. Natürlich sind wir für alle enthaltenen Fehler selbst verantwortlich.

Siegfried Heiler
Paul Michels

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Was ist Statistik?

Dem Wort "*Statistik*" werden mehrere Bedeutungen zugeordnet. Zum einen versteht man darunter Tabellen, Graphiken, Zahlenkolonnen, Schaubilder oder daraus gewonnene Kenngrößen wie Mittelwerte oder relative Häufigkeiten. Manchmal verbindet man auch den (amtlichen) Apparat zur Erhebung und Dokumentation von Daten mit diesem Begriff. Zum anderen ist Statistik eine "wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung formaler Methoden zur Gewinnung, Beschreibung und Analyse sowie zur Beurteilung quantitativer Beobachtungen (Daten) ist" (Vogel, 1991). Als solche ist Statistik eine wissenschaftliche Methode zur Vorbereitung von Entscheidungen bei Unsicherheit. Unsicherheit birgt stets das Problem von Fehlentscheidungen in sich. Oft lassen Daten auch unterschiedliche Interpretationen zu. Beides hat dazu geführt, daß die Disziplin Statistik gelegentlich in Mißkredit geraten ist. Hier nur eine kleine Auswahl aus einer großen Fülle kritischer, häufig ironischer Bemerkungen:

> In der alten Zeit gab es keine Statistik, und daher mußten die Leute lügen. So ist denn die alte Literatur voll von gewaltigen Übertreibungen — es wimmelt nur so von Riesen und Wundern! Damals log man also, aber heute hat man die Statistik dafür; somit ist alles beim alten geblieben. (Stephen Leacock)

> Trau keiner Statistik, die du nicht selbst gefälscht hast.

Es gibt drei Arten von Lügen: Notlügen, gemeine Lügen — und dann noch die Statistik. (Disraeli)

Statistik kann nicht das auf Unsicherheit beruhende Risiko von Fehlentscheidungen ausschalten. Aber sie macht es kalkulierbar. Die Statistik befaßt sich nicht mit *Einzelphänomenen* oder Individuen. Sie trifft deshalb auch keine Aussagen für einzelne Sachverhalte, sondern hat stets ein *Gesamtbild* im Auge. In dieses Gesamtbild gehen von den Individuen nur einige wenige Charakteristika (ihre statistischen "Schatten") ein. Häufig wurde die Statistik auch als Lehre von der Analyse von *Massenerscheinungen* bezeichnet. Der mit dem Wort Masse suggerierte Eindruck ist jedoch falsch. Wir sind es heute durchaus gewohnt, aus Stichproben relativ kleinen Umfangs allgemeine, weitreichende Schlüsse zu ziehen.

Die wissenschaftliche Disziplin Statistik wird häufig in die zwei Teilgebiete "*Deskriptive Statistik*" und "*Schließende Statistik*" unterteilt, welche an Fachbereichen deutscher Universitäten in getrennten Vorlesungen behandelt werden. Aufgabe der deskriptiven Statistik ist die Aufbereitung mitunter umfangreichen Datenmaterials. Dazu zählt vor allem die graphische Darstellung und die Charakterisierung der Daten durch einige wenige Kenngrößen. Ihre Aussagen beschränken sich auf den untersuchten Datensatz selbst; Schlüsse über die vorliegenden Daten hinaus sind nicht beabsichtigt. Die schließende Statistik bedient sich hingegen formaler (mathematischer) Modellvorstellungen. Das Vorliegen unkontrollierbarer Einflüsse führt dann zu sogenannten stochastischen (d.h. auf der Wahrscheinlichkeitstheorie aufbauenden) Modellen. Innerhalb dieser ist es möglich, Rückschlüsse von einer zufällig ausgewählten Teilmenge auf die Gesamtheit, aus der diese entnommen wurde, zu ziehen. Nachdem in Anwendung und Wissenschaft lange Zeit ein vorrangiges Interesse an der schließenden Statistik mit Hilfe von wahrscheinlichkeitstheoretischen Modellen bestand, rückten Ende der siebziger Jahre vor allem die Arbeiten von John W. Tukey den datenanalytischen Aspekt der Statistik ins Bewußtsein. Tukeys "*Explorative Datenanalyse (EDA)*" stellt ein modernes, vielfältiges Instrumentarium zur Erkennung und Darstellung von Strukuren in den Daten dar. Neben der Behandlung der klassischen Methoden der deskriptiven Statistik wird daher in diesem Buch besonderer Wert auf diesen datenanalytischen Aspekt gelegt.

1.2 Zur Geschichte der Statistik

Die Statistik in der heutigen Form ist aus vier historischen Wurzeln entstanden.

I. Amtliche Erhebungen

Schon im frühen Altertum wurden amtliche Erhebungen als Volkszählungen und andere Registrierungen durchgeführt. Sie dienten vorwiegend militärischen und steuerlichen Zwecken.

Als eine erste Quelle wird gern ein ägyptischer Gedenkstein erwähnt, dessen Entstehung auf den Zeitraum um 2600 v. Chr. datiert. Er enthält Bevölkerungszahlen und ist so Zeugnis einer Erhebung, die, so vermutet man, der Organisation des Pyramidenbaus diente.
Kung-Fu-Tse (Konfuzius), 552 – 479 v. Chr., der ältere schriftliche Überlieferungen sammelte, berichtet, daß in der Yang-Schao-Kultur am Hoang-Ho neben dem Volk auch Grund und Boden sowie Gewerbe und Handel erfaßt wurden.
Im 2. Buch Samuel, 34, ist die Zählung der wehrfähigen Männer Israels durch König David erwähnt. ("Schon auf dieser Zählung lastete ein Fluch, und David wurde dafür von Gott bestraft.") Der römische Kaiser Servius Tullius verfügte um 550 v. Chr. einen Zensus (Volkszählung), der ab 433 periodisch, erst in 5-, dann in 10- und schließlich in 15-jährigem Abstand durchgeführt wurde, insgesamt 69 mal. Dokumentationen dazu sind im "Brevarium Augusti" von Kaiser Augustus (63 v. Chr. - 14 n. Chr.) angelegt und später fortgeführt worden. Die Weihnachtsgeschichte in der Bibel berichtet bekanntlich von so einer Volkszählung. Karl der Große (768 - 814) ordnete eine Bestandsaufnahme all seiner Güter, Domänen, etc. an, in der vielen bis ins kleinste Detail dokumentiert wurde. Wilhelm der Eroberer (1027 - 1087) ließ das *Domesday Book* anlegen, in dem die Ergebnisse einer 1086 durchgeführten Erhebung von Land und Bevölkerung aufgezeichnet sind. Dieses "Buch des jüngsten Gerichts" wird auch als Grundkataster des Okzidents bezeichnet. Gegen Ende 19. Jahrhunderts erfolgte die Gründung der ersten nationalen statistischen Ämter. In Deutschland entstanden:

1805 das Statistische Bureau in Preussen

1834 das Statistische Centralbureau des Deutschen Zollvereins und

1871 das Kaiserliche Statistische Reichsamt (in diesem Jahr wurde auch bereits die erste deutsche Volkszählung durchgeführt)

II. Die deutsche Universitäts- oder Kathederstatistik

Als älteste Wurzel der Statistik als Wissenschaft kann die *(deutsche) Universitäts- oder Kathederstatistik* im 17. und 18. Jahrhundert angesehen werden. Statistik wurde als Lehre von den Staatsmerkwürdigkeiten ("merkwürdig" im Sinne von "bemerkenswert") verstanden. Gesammelte Informationen zur Beschreibung der geographischen Gegebenheiten, der Bevölkerung, der Verfassung, der Wirtschaft, der Verwaltung und des Militärs wurden für einen einzelnen Staat oder für den Vergleich mehrerer Staaten zusammengefaßt. Vorgänger dieser Richtung sind

- die *Politiken des Aristoteles* (384-332 v. Chr.),
- das 1562 in Venedig erschienene Buch "Del governo et administratione di diversi regai" von *Francesco Sansovino* (1521 - 1586), das auf Reisebeschreibungen von Privatleuten und Gesandtenberichten beruht,
- Arbeiten von *Giovanni Botero* (1540 - 1617), der die "vergleichende Methode" verwendet, sowie
- eine Schriftenreihe mit 36 Bänden des Direktors der holländisch-westindischen Kompanie, *Jan de Laet* (1583 - 1649).

Die deutsche Universitäts- und Katheterstatistik wurde vom 1656 erschienenen Buch "Teutscher Fürstenstaat" von dem seinerzeit berühmten Historiker und Staatsmann *Veit Ludwig von Seckendorff* (1626 - 1692) geprägt. Daraus entwickelte sich die *Lehre von den Staatsmerkwürdigkeiten.* Erster Lehrer dieses Faches war *Hermann Conring* (1606 - 1681), Professor für Philosophie, Medizin und Politik in Helmstedt. Seine säamtlichen Schriften sind in Lateinisch und rein verbal gehalten. Die Vorlesung "*Collegium politico-statisticum*" von *Martin Schmeitzel* (1679 - 1747), Professor in Jena und Halle, hat den Namen *Statistik* geprägt und *Gottfried Achenwall* (1719 - 1772), seinen Schüler, zur ersten in deutsch gehaltenen Veranstaltung über Staatenkunde mit dem Titel "*Statistik*" an der Universität Göttingen (1747) angeregt. In seinem 1749 erschienenen Werk "*Abriß der Staatswissenschaft der Europäischen Reiche*" wird alles "Bemerkenswerte" in den gegenwärtigen Zuständen des Territoriums, der Bevölkerung und der Verwaltung eines Landes beschrieben. Sein Nachfolger *August Ludwig von Schlözer* (1735 - 1809) gilt als Vater der deutschen Publizistik.

Mit der Ausdehnung der amtlichen Statistik wurden zur Darstellung immer mehr Tabellen herangezogen. Die Vertreter dieser, von der aus England kommenden *Politischen Arithmetik* beeinflußten Richtung wurden von den Anhängern Achenwalls als "Tabellenknechte" beschimpft, die nur "gemeine" statt "höhere" Statistik betrieben und die idealen Faktoren in einem Staatswesen übersahen.

> "Die ganze Wissenschaft der Statistik, eine der edelsten, ist durch die politischen Arithmetiker um alles Leben, um allen Geist gebracht und zu einem Skelett, zu einem wahren Kadaver herabgewürdigt, auf das man nicht ohne Widerwillen blicken kann." (Göttingische gelehrte Anzeigen 1806, S. 84)

Carl Gustav Adolph Knies (1821 - 1898) zieht 1850 einen Schlußstrich unter die deutsche Universitätsstatistik, indem er jener Richtung den Namen Statistik zuerkennt, die auf Beobachtung beruhende Zahlenangaben zu mathematisch fundierten Schlüssen verwendet.

III. Die politische Arithmetik

Während die deutschen Universitätsstatistiker gesammelte Daten nur zur Beschreibung benutzten, dienten sie den politischen Arithmetikern als Grundlage für Analysen. Sie suchten nach Gesetzmäßigkeiten in den bevölkerungs- und sozialstatistischen Daten. Die parallel zur deutschen Universitätsstatistik verlaufende Entwicklung der *politischen Arithmetik*, die ihren Ursprung in England hatte, war durch die folgende Arbeiten gekennzeichnet:

1662 ging der Royal Society of London die Schrift "*Natural and political observations upon the bills of mortality, chiefly with reference to the government, religion, trade, growth, air, deseases etc. of the City of London*" von Captain *John Graunt* zu, deren Aussagen auf den Geburts- und Sterbelisten Londons seit 1603 beruhten. Hier wurden zum ersten Mal aus Daten allgemeine Schlüsse gezogen.

1693 erschien ein Werk des Astronomen *Edmund Halley* (1656 - 1742), das die erste vollständige Sterbetafel — ermittelt anhand der Kirchenbücher der Stadt Breslau — und daraus resultierende Prämienberechnungen für Lebensversicherungen enthielt. Die Daten stamm-

ten von dem Breslauer Pastor *Kaspar Neumann*, der Aufzeichnungen aus Kirchenbüchern sammelte.

Den Namen erhielt die neue Richtung von *William Petty* (1623 - 1687), einem Freund Graunts, der wirtschaftliche Aspekte in seinen "*Essays in Political Arithmetic*" einbezog. Den Höhepunkt erreichte sie durch *Thomas Robert Malthus* (1766 - 1834), dessen "Gesetze" über die Zunahme der Bevölkerung und der Nahrungsmittelproduktion berühmt geworden sind.

Inzwischen hatte sich die politische Arithmetik über Belgien, Holland, Frankreich, Schweden verbreitet und erst relativ spät, wegen der dort herrschenden Universitätsstatistik, Deutschland erreicht. Erster deutscher Vertreter war der Pastor und Dozent *Kaspar Neumann* (1648 - 1715), der auf Vorschlag von Leibniz 1706 als einer der ersten zum Mitglied der neugegründeten Akademie der Wissenschaften in Berlin gewählt wurde. Wichtigster deutscher Vertreter wurde der preussische Feldprediger *Johann Peter Süßmilch* (1707 - 1767) mit seinem Werk "*Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts, aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen*" (1741). Der Titel gibt am besten Auskunft über die Intention dieser Richtung. Ihre stärkste Ausprägung, aber auch ihre größte Übertreibung erfuhr die Politische Arithmetik durch den Belgier *Lambert Adolphe Jacob Quetelet* (1796 - 1874), der die erste Volkszählung nach heutigem Muster 1846 in Belgien organisierte. Auf seiner Suche nach Gesetzmäßigkeiten in sozialstatistischen Daten glaubte er, der Physik ähnliche Naturgesetze gefunden zu haben und entwickelte in seinem Werk "Sur l'homme et ledéveloppement de ses facultés ou essai de physique sociale" (erstmals 1835 in Paris erschienen) die Lehre vom *homme moyen* als Ideal, das sämtliche Merkmale in mittlerer Ausprägung aufweist.

IV. Die Wahrscheinlichkeitstheorie

Von größter Bedeutung für den Aufschwung der Statistik im 19. und 20. Jahrhundert war die Entwicklung der *Wahrscheinlichkeitstheorie*, die aus der Chancenberechnung für Glückspiele im 16. Jahrhundert entstanden ist. Als erste Ausführungen in dieser Richtung sind die Schriften von *Gerolamo Cardano* (1501 - 1576) (beispielsweise das um 1526 erschienene Buch "*Liber de ludo aleae*") sowie die Abhandlung "*Sopra le scorpeste dei Dadi*" (über die Wahrscheinlichkeit beim Spiel mit drei Würfeln) von *Galileo Galilei* (1564 - 1642) zu nennen.

Berühmt geworden sind die Anfragen des Spielers *Antoine Chevalier de Méré* (1610 - 1684) bei dem französischen Mathematiker *Blaise Pascal* (1623 - 1662), die zu einem Briefwechsel zwischen Pascal und dem Mathematiker *Pierre de Fermat* (1602 - 1665) in den Jahren 1651 - 1655 führten, in dem wahrscheinlichkeitstheoretische (kombinatorische) Fragen behandelt wurden. Diese Überlegungen gingen in das erste Buch über Wahrscheinlichkeitstheorie "De Ratiociniis in Ludo Aleae" (1657) des holländischen Physikers *Christian Huygens* (1629 - 1695) ein. Zur Veranschaulichung seiner Ergebnisse verwendete Huygens das *Urnenmodell* (Ziehen von Kugeln unterschiedlicher Farben aus einer gut durchgemischten Urne).

Die weitere Geschichte der Wahrscheinlichkeitstheorie ist eng verknüpft mit dem Namen *Bernoulli* (vgl. etwa Begriffe wie *Bernoulli-Verteilung* und das *Bernoulli-Theorem*). Die Baseler Mathematikerfamilie stellte in vier aufeinander folgenden Generationen acht Professoren. *Jakob Bernoulli* (1654 - 1705) schrieb das Buch "Ars Conjectandi" (Kunst des Vermutens), das 1713 von seinem Neffen Nikolaus Bernoulli (1695 - 1726) veröffentlicht wurde. Es bestand aus 4 Teilen. Teil 1 war von dem Physiker Huygens geschrieben und von Jakob Bernoulli ergänzt, Teil 2 behandelt die *Kombinatorik*, Teil 3 die Anwendung auf Glücksspiele, Teil 4 die Anwendung auf wirtschaftliche Probleme. Die *Binomialverteilung* und die sogenannten Gesetze der großen Zahlen werden dort erstmals erwähnt.

Der Zusammenhang zwischen der Normalverteilung und der Binomialverteilung wird von *Abraham de Moivre* (1667 - 1754) erkannt und in seinem Werk "*The Doctrine of Chances*" 1718 veröffentlicht. Der englische Geistliche *Thomas Bayes* (1702 - 1761) schrieb "*An Essay Towards Solving a Problem in the Doctrine of Chances*", in dem bedingte Wahrscheinlichkeiten und Rückschlüsse auf Wahrscheinlichkeiten bei Kenntnis von Spielausgängen behandelt werden. Das Werk "*Théorie analytique des probabilités*" (1812) von *Pierre Simon de Laplace* (1749 - 1827) enthält einen Überblick über den Stand der Wahrscheinlichkeitsrechnung. In einem späteren Werk werden der nach ihm benannte (auf Bernoulli zurückgehende) *Laplacesche Wahrscheinlichkeitsbegriff* eingeführt und Anwendungen auf Zeugenaussagen und Gerichtsurteile diskutiert. Die *Methode der kleinsten Quadrate* geht auf *Adrien Marie Legendre* (1752 - 1833) zurück. *Carl Friedrich Gauß* (1777 - 1855), dessen Portrait den 10-DM-Schein schmückt, trug wesentliche Arbeiten zur *Normalverteilung ("Gauß-"Verteilung)*, zur *Methode der kleinsten Quadrate* und zur *Theorie der Beobachtungsfehler* bei. *Siméon Denis Poisson* (1781 - 1840) erweiterte in seiner Schrift "Le lois des grands nombres" (*Gesetz der großen Zahlen*) das Bernoullische Theorem.

In Deutschland wurden diese mathematische Methoden der Statistik lange Zeit nicht erkannt oder sogar abgelehnt. Vor allem *Gustav Rümelin* (1815 - 1889) und der als "Altmeister der deutschen Statistik" bezeichnete bayrische Landesstatistiker *Georg von Mayr* (1841 - 1925) taten sich in der Kritisierung dieser Entwicklung hervor und trugen so zur Verkundung ihrer Ausbreitung in Deutschland bei. Auch an der mathematischen Fakultäten deutscher Universitäten wurde die Bedeutung dieser Fachrichtung lange Zeit nicht erkannt.

So entstanden die Grundlagen der *modernen mathematischen Statistik* vor allen im angelsächsischen Raum. Als Vertreter der *angelsächsischen Schule* sind insbesondere die folgenden Persönlichkeiten zu nennen: der Biologe *Francis Galton* (1822 - 1911, Entwicklung des *Korrelationskoeffizienten*), *Francis Ysidro Edgeworth* (1845 - 1926), *Karl Pearson* (1857 - 1936, Wiederentdeckung der χ^2-*Verteilung* im Jahre 1900, χ^2-*Test* zur Überprüfung der Anpassungsgüte eines Modells an die Daten, *Korrelation* und *Regression*, β-*Funktion*, β- *und* Γ-*Verteilung*, Gründung der *Biometrika* im Jahre 1901), *G. U. Yule* (1871 - 1951, *Assoziationskoeffizient*), *William Sealy Gosset* (1876 - 1937, "*Student*" *'sche t-Verteilung* als Verteilung des Stichprobenmittels) und *Sir Ronald Aylmer Fisher* (1890 - 1962), in dessen Buch "*The Design of Experiments*" (1935) grundlegende Methoden zur *Versuchsplanung* und *Varianzanalyse* behandelt werden. Zu R.A. Fishers Ehren wurde die von *Snedecor* abgeleitete *F-Verteilung* benannt. Der Entwicklung der statistischen Test- und Entscheidungstheorie seit dem zweiten Weltkrieg ist vor allem durch die Arbeiten von *Jerzy Neyman* und *Egon S. Pearson* sowie von *Abraham Wald* maßgeblich geprägt worden.

Es kann zusammenfassend festgestellt werden, daß die Grundlagen einer modernen mathematischen Statistik von der sogenannten "englischen Schule" gelegt wurden. Trotz der erwähnten Probleme im deutschsprachigen Raum kann jedoch auch auf wichtige Entwicklungen der "kontinentalen Schule" hingewiesen werden. Als Vorläufer dieser kontinentalen Schule sind der Phychologe G. Th. Fechner (1801-1887), Untersuchungen über schiefe Verteilungen, Fechnersche Lageregel) sowie der Geodät F. Helmert zu nennen. Als Hauptvertreter der kontinentalen Schule gelten die Nationalökonomen W. Lexis (1837-1914), Lexische Dispersionstheorie) und L. von Bortkiewicz (1868-1931), Indextheorie, Poissonsapproximation als "Gesetz der kleinen Zahlen") sowie A. A. Tschuprov (1874-1926), Untersuchung der empirischen Verteilung), A. A. Markov (1857-1922), Gesetze der großen Zahlen, Markov-Ketten) und O. Anderson (1887-1960). L. von Bortkiewicz regte mit Arbeiten zur Variationsbreite einer Stichprobe R. von Mises (1883-1953) sowie E. J. Gumbel (1891-1966) zu ihren Arbeiten

auf dem Gebiet der Extremwerttheorie an.

Die Entwicklung der Wahrscheinlichkeitstheorie wurde besonders in Rußland weiter vorangetrieben. Neben A. A. Markov sind hier insbesondere zu nennen P. L. Tschebyschev (1821-1914), A. M. Liapunov (1857-1918), zentraler Grenzwertsatz) und A. N. Kolmogorov (1903-1988), axiomatische Begründung der Wahrscheinlichkeitstheorie).

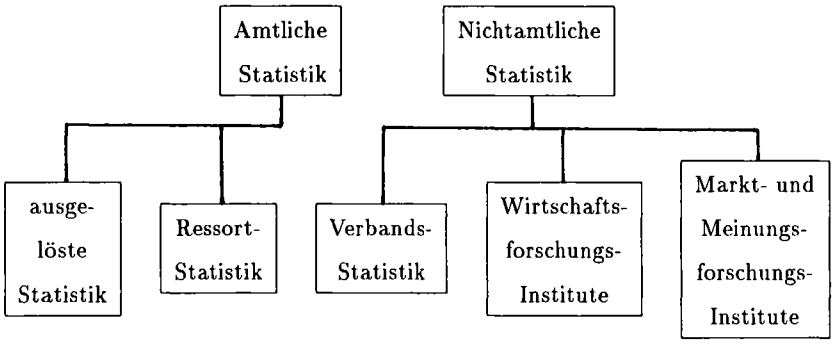
V. Neuere Entwicklungen

Viele, inzwischen auch schon als klassisch zu bezeichnenden Methoden der mathematischen Statistik basieren auf der Annahme, daß die Meßwerte einer sogenannte Normalverteilung folgen. Untersuchungen haben gezeigt, daß dieser Verteilungstyp in der Praxis weit weniger oft anzutreffen ist als lange Zeit vermutet. Weiter wurde festgestellt, daß die unter Normalverteilung optimalen Verfahren häufig schon bei kleineren Abweichungen von diesem Modelltyp sehr unbefriedigende Ergebnisse liefern. Dies hat in den sechsiger Jahren zur Entwicklung robuster Verfahren geführt, die so konzipiert sind, daß sie auch bei Abweichungen von der Normalverteilung zuverläßige Schätz- und Testergebnisse liefern und sogenannte Ausreißer "verkraften" können.

Etwa zur selben Zeit stellt man eine Renaissance der datenorientierten Betrachtungs- und Vorgehensweise fest. Das Ziel dabei ist, in der Daten vorhandene (häufig versteckte) Strukturen und etwaige Auffälligkeiten oder Anomalien aufzudecken, um so erst Hypothesen und Modelle zu finden, die in einer späteren Phase mit mathematisch-statistischen Verfahren überprüft werden können. Zu diesem Zweck wird das Bündel der in der deskriptiven Statistik üblichen Methoden in vielfältiger Weise erweitert. Dabei spielt, wie in der robusten Statistik, die Resistenz gegenüber Ausreißern eine wichtige Rolle - um diese in den Ergebnissen, als möglicherweise interessante Abweichungen, auch klar erkennen zu können und um zu verhindern, daß Ausreißer die für die Mehrzahl der Daten gültigen Aussagen verfälschen. Diesem Aspekt wird in dem vorliegenden Buch besondere Aufmerksamkeit gewidmet.

Die obige Ausführungen zur Geschichte der Statistik sind sehr knapp gehalten und setzen teilweise auch schon ein gewisses überblickartiges Wissen voraus, das an dieser Stelle von vielen Lesern noch gar nicht erwartet werden kann. Weiter Interessierte seien auf die Bücher von T. M. Porter (1986), S. Stigler (1986), I. Hacking (1990) und I. Schneider (1988) hingewi-

Abbildung 1.1: Statistische Institutionen in der Bundesrepublik Deutschland



sen. Ein Überblick über die Entwicklung der mathematischen Statistik, der besonders auch den deutschsprachigen Raum einbezieht, ist zu finden in dem Beitrag von H. Witting zur Festschrift zum Jubiläum der Deutschen Mathematiker-Vereinigung (1990). Der Beitrag von U. Krengel in derselben Festschrift schildert die Entwicklung der Wahrscheinlichkeitstheorie im deutschen Sprachraum.

1.3 Statistische Institutionen in Deutschland

Die statistischen Institutionen in der Bundesrepublik können in die Bereiche *amtliche* und *nicht amtliche Statistik* untergliedert werden (vgl. Abbildung 1.1). Die amtliche Statistik umfaßt die *ausgelösten* Behörden, die primär für statistische Aufgaben zuständig sind, und die *Ressorts*, die sich sekundär mit Statistik beschäftigen.

Träger der ausgelösten Statistik sind

- das *Statistische Bundesamt*, Wiesbaden,
- die *Statistischen Landesämter* und
- *Kommunalstatistische Ämter.*

Die Aufgabe dieser statistischen Ämter besteht darin, Daten zu erheben, zu sammeln, aufzubereiten und darzustellen. Diese Ergebnisse werden einerseits vom Gesetzgeber, den Regie-

rungen und den Verwaltungen als Planungsgrundlage und Erfolgskontrolle staatlicher Maßnahmen benötigt; andererseits bilden diese Daten die Basis für Analysen und Prognosen der gesellschaftlichen und wirtschaftlichen Situation und deren Entwicklung. Die Resultate der Bundesstatistik sind daher von öffentlichen Interesse und für jedermann zugänglich. Nicht alles, was an Datenmaterial anfällt, kommt zur Veröffentlichung. Über den *Auskunftsdienst des Statistischen Bundesamtes* und die Nutzung der externen Datenbank *STATIS-BUND* ist es möglich, weiteres gegebenfalls auch entsprechend aufgearbeitetes Material zu beziehen. Der Charakter der ausgelösten Statistik ist von drei wesentlichen Prinzipien geprägt:

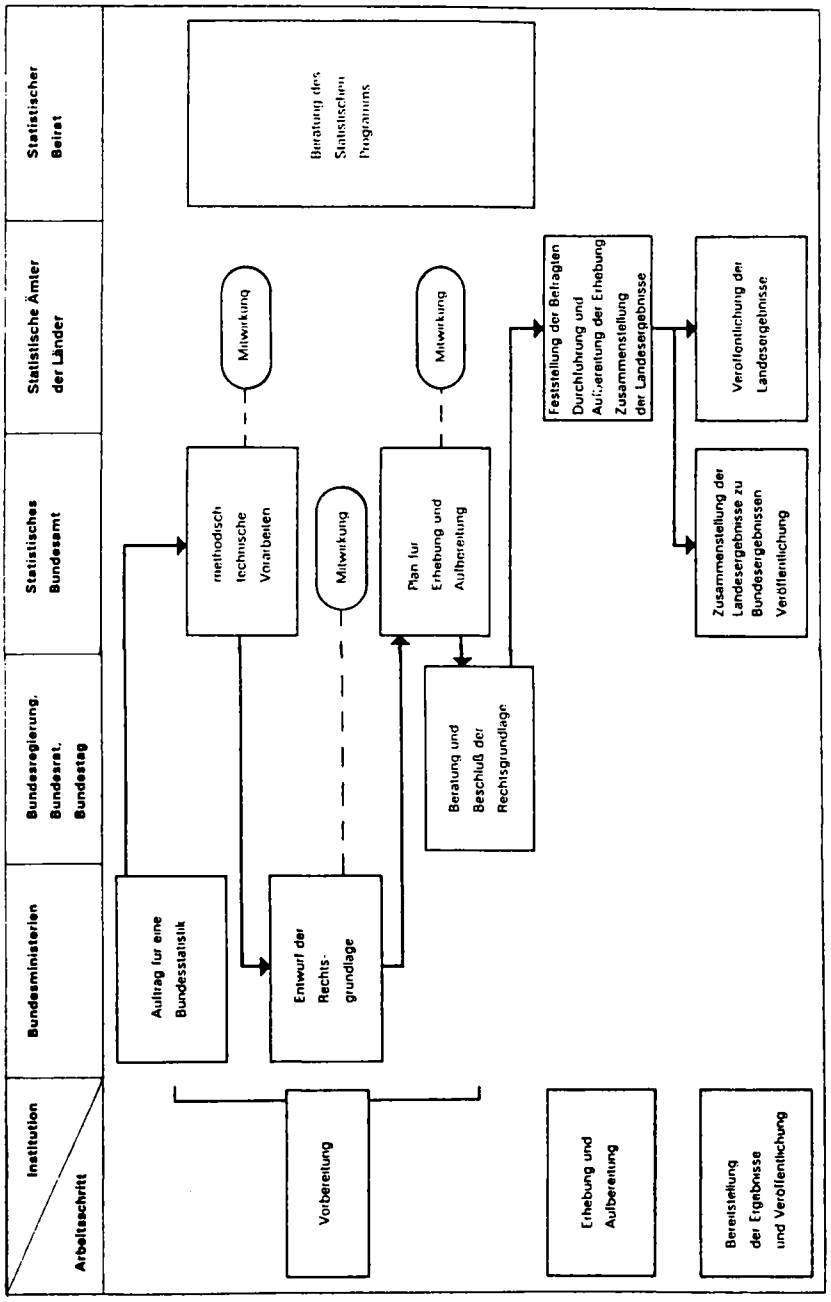
- der *fachlichen Konzentration* aller statistischen Arbeiten in statistischen Ämtern als eigens dafür eingerichtete Fachbehörden,
- der *regionalen Dezentralisierung* entsprechend dem föderalistischen Staats- und Verwaltungsaufbau und
- der *Legalisierung* der Arbeit der amtlichen Statistik auf der Grundlage von Gesetzen und Rechtsverordnungen.

Der Ablauf von Bundesstatistiken ist in der Abbildung 1.2 veranschaulicht.

Im umfangreichen *Veröffentlichungsprogramm des Statistischen Bundesamtes* sind zusammenfassende Veröffentlichungen, wie etwa das *Statistische Jahrbuch*, die Monatszeitschrift *Wirtschaft und Statistik* und der *Statistische Wochendienst*, sowie speziellere *Fachserien* enthalten. Derzeit werden vom Statistischen Bundesamt 19 Fachserien herausgegeben. Einen Überblick über die Veröffentlichungen des Statistischen Bundesamtes enthält die Abbildung 1.3.

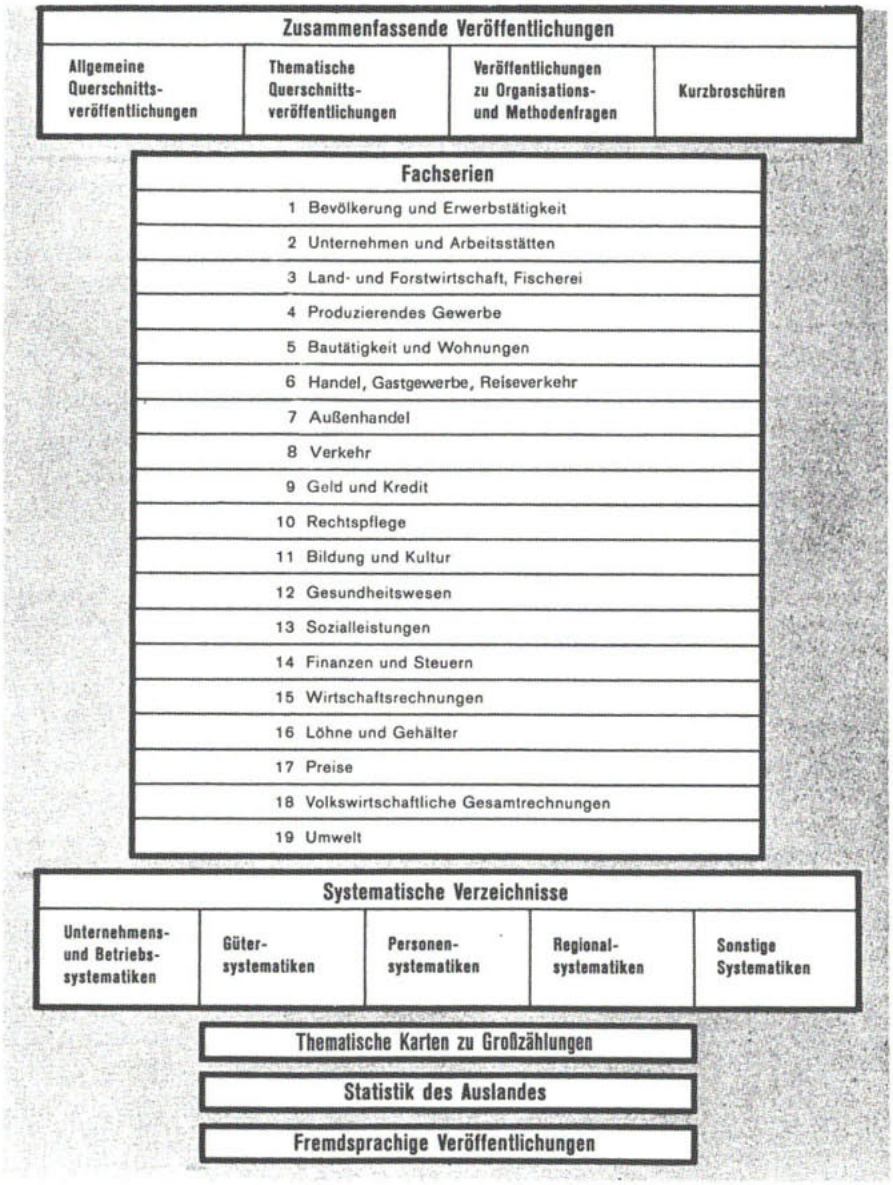
Das Schaubild 1.4 stellt die Einbindung des Statistischen Bundesamtes in ein System internationaler Organisationen dar. Detaillierte Informationen über die Arbeit der ausgelösten Statistik (inbesondere auch über Daten zu diversen Themenschwerpunkten) können vom Statistischen Bundesamt, den Statistischen Landesämtern und den Kommunalstatistischen Ämtern direkt bezogen werden. Einen umfassenden Überblick über die Bundesstatistk bietet die vom Bundesamt in mehrjährigen Abständen herausgegebene Schrift "Bundesstatistik, das Arbeitsgebiet derDas Arbeitgebiet der Bundesstatistik". Vom Bundesamt direkt kann das *Veröffentlichungsverzeichnis des Statistischen Bundesamtes*, ein Katalog der lieferbaren Pu-

Abbildung 1.2: Der Ablauf von Bundesstatistiken



Quelle: Bundesstatistik - für wen und wofür, 3.Auflage 1989, Statistisches Bundesamt, Wiesbaden

Abbildung 1.3: Veröffentlichungssystem des Statistischen Bundesamtes



Quelle: Verzeichnis der Veröffentlichungen 1992/93, Statistisches Bundesamt, Wiesbaden

blikationen (mit Bezugspreisen), bezogen werden. Hinzu kommen noch die Publikationen der Statistischen Landesämter und der Kommunalstatistischen Ämter, die eine starke regionale Differenzierung der Ergebnisse beinhalten.

Im Gegensatz zur ausgelösten Statistik, die sich zur Dienstleistungseinrichtung mit dem Ziele der Datenlieferung für vielfältigee Nutzungen entwickelt hat, wird die *Ressortstatistik* in Institutionen (Ressorts) betrieben, deren Hauptaufgaben auf einem anderen Gebiet liegen. Als Träger dieses Bereiches der amtlichen Statistik sind beispielsweise

- die Bundesanstalt für Arbeit (Erwerbstätigkeit),
- die Deutsche Bundesbank (Geld und Kredit),
- die Bundesministerien,
- das Bundesaufsichtsamt für das Versicherungs- und Bausparwesen,
- das Kraftfahrt-Bundesamt oder
- das Umwelt-Bundesamt

zu nennen, die zum Teil eigene statistische Veröffentlichungen herausgeben. Beispiele hierfür sind die *Monatsberichte der Deutschen Bundesbank* und die *Statistik der Arbeitslosigkeit.* In den Ressorts werden Daten gesammelt und analysiert, die für das jeweilige Fachgebiet von besonderem Interesse sind.

Die *nichtamtliche Statistik* in der Bundesrepublik kann in die Bereiche *Verbandsstatistik*, *Statistik der Wirtschaftsforschungsinstitute* und *Statistik der Markt- und Meinungsforschungsinstitute* untergliedert werden.

Träger der *Verbandsstatistik* sind (auszugsweise)

- Industrie- und Handelskammern,
- Handwerkskammern,
- Landwirtschaftskammern,
- Fachorganisationen des Handwerks,
- Mitgliederverbände
    - des Bundesverbandes der Deutschen Industrie,

  - des Deutschen Bauernverband,
  - des Bundesverbandes der Freien Berufe,
  - der Bundesvereinigung der Deutschen Arbeitgeberverbände und

- Gewerkschaften.

Die Untersuchungen und die statistischen Auswertungen dienen diesen Verbänden als Informationsquelle für ihre Mitglieder und Argumentationsgrundlage zur Durchsetzung der verbandsorientierten Ziele. Man sollte daher beachten, daß Erhebungs- und Analysemethoden sowie Veröffentlichungsstrategien in enger Verbindung zur Zielsetzung des Verbandes stehen können.

Wirtschaftsforschungsinstitute (wie auch Verbände) greifen häufig auf das von der amtlichen Statistik bereitgestellte Datenmaterial und auf eigene Untersuchungen zurück und verwenden die Daten zur Beschreibung und Prognose der wirtschaftlichen Entwicklung. Der Arbeitsgemeinschaft deutscher wirtschaftswissenschaftlicher Forschungsinstitute gehören

- das Deutsche Institut für Wirtschaftsforschung e.V. Berlin (DIW),
- das IFO-Institut für Wirtschaftsforschung e.V. München,
- das Rheinisch-Westfälische Institut für Wirtschaftsforschung e.V. Essen (RWI),
- das Institut für Weltwirtschaft Kiel und
- das Hamburgische Weltwirtschaftsarchiv (HWWA)

an. Als weitere Einrichtungen seien hier

- das Wirtschaftswissenschaftliche Institut der Gewerkschaften GmbH, Köln, und
- das Deutsche Industrieinstitut e.V., Köln,

genannt.

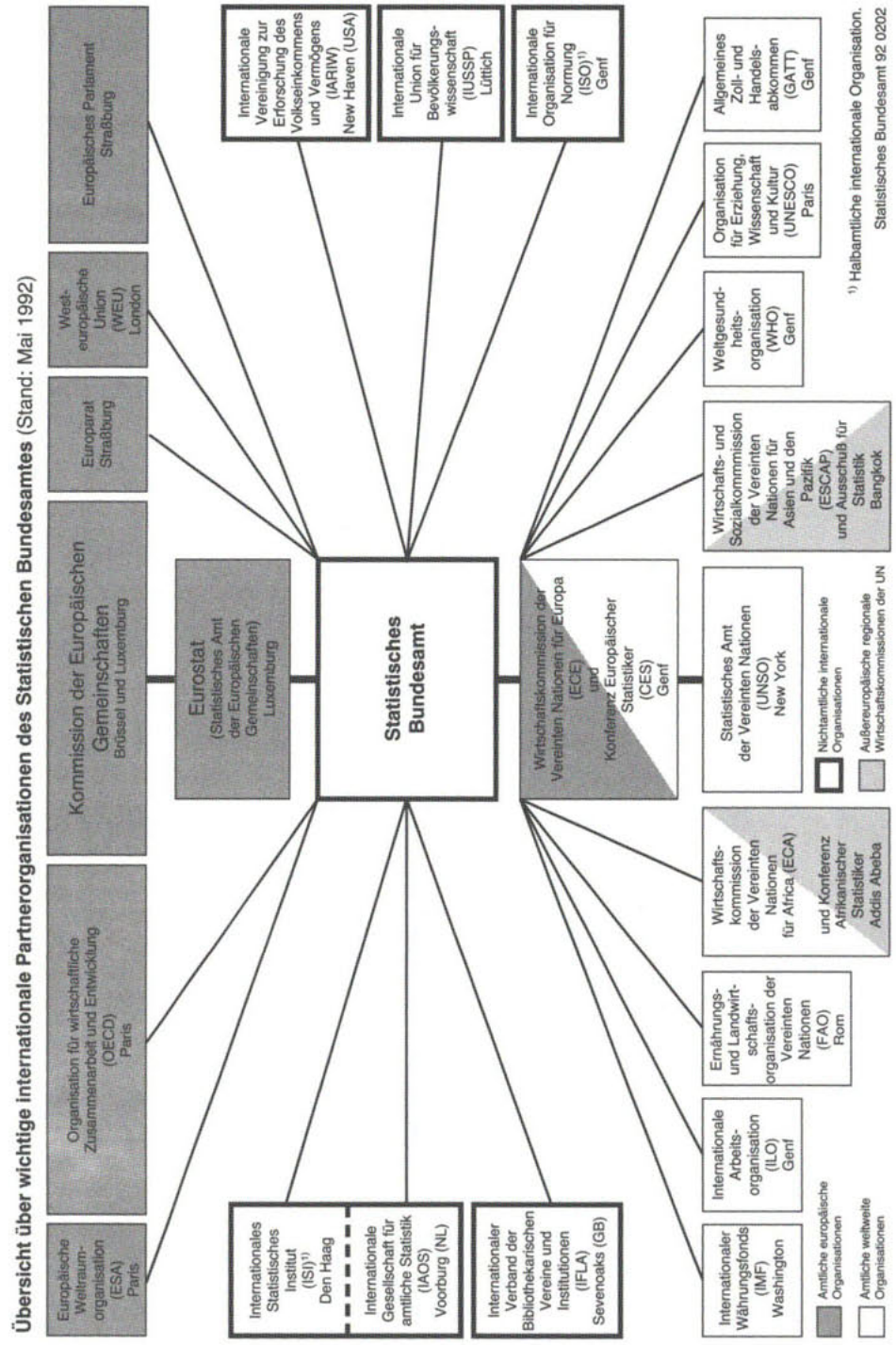
Markt- und Meinungsforschungsinstitute führen in der Regel eigene Erhebungen durch, um Informationen etwa über das Kaufverhalten von Kunden, die Einstellung der Bevölkerung zu aktuellen Problemen, die Absatzmöglichkeiten für bestimmte Produkte oder die Wahlchancen von politischen Parteien zu gewinnen. Ihre Auftraggeber sind vor allem Unternehmen,

denen die Untersuchungsergebnisse als Entscheidungsgrundlage dienen. Aber auch Parteien und öffentliche Institutionen geben Studien in Auftrag. In der Bundesrepublik gibt es eine zunehmend größer werdende Anzahl solcher Institute, von denen hier nur einige genannt werden können:

- Infratest, München,
- Gesellschaft für Konsumforschung (GfK), Nürnberg,
- A.C. Nielsen GmbH, Frankfurt,
- IMS, Frankfurt,
- GFM-Getas, Hamburg,
- IVE, Hamburg,
- Kehrmann, Hamburg,
- Infas, Bad Godesberg,
- Sample, Mölln,
- Marplan, Offenbach,
- Emnid, Bielefeld,
- Burke, Frankfurt,
- Contest Census, Frankfurt,
- Institut für Demoskopie, Allensbach,

Viele Marktforschungsintitute haben sich dem Arbeitkreis Deutscher Marktforschungsinstitute (ADM) angeschlossen, bei dem man auch eine detaillierte Beschreibung der Arbeitsgebiete seiner Mitgliedsintitute anfordern kann (Anschrift: Papenkamp 2-6, 2410 Mölln).

Abbildung 1.4: Internationale Einbindung der Bundesstatistik



Quelle: Statistisches Bundesamt 920202

Kapitel 2

Durchführung einer Untersuchung

Am Anfang einer Untersuchung steht stets ein *Sachproblem.* Eine Problemstudie legt mögliche Ursachen, Voraussetzungen und schließlich das Forschungsziel fest. Schon in dieser Planungsphase sollten Empiriker bzw. Sachwissenschaftler und Statistiker zusammenarbeiten, denn auch noch so ausgeklügelte Instrumente der Datengewinnung, -aufbereitung, -darstellung und -auswertung können die in dieser Phase gemachten Fehler nicht mehr ausgleichen. Nach der Festlegung möglicher Einfluß- und Zielgrößen werden *Hypothesen*, d.h. Vermutungen über deren mögliches Zusammenwirken aufgestellt. Eine statistische Untersuchung kann solche Hypothesen nicht "beweisen" oder falsifizieren", sondern sie führt zu deren Verwerfung oder Akzeptierung (mit der Möglichkeit von Fehlentscheidungen).

2.1 Statistische Grundbegriffe

Zunächst werden die verwendeten *Begriffe präzisiert* (d. h. operational gemacht). Die Phänomene, Objekte, Personen etc., auf die sich die Untersuchung bezieht, heißen *Untersuchungseinheiten* und werden im folgenden durch das Symbol ω dargestellt. Die Gesamtheit aller Untersuchungseinheiten heißt *Grundgesamtheit* (*ensemble, statistische Gesamtheit, statistische Masse*) und wird durch das Symbol Ω repräsentiert. Von jedem Objekt muß eindeutig feststellbar sein, ob es unter den jeweiligen Begriff fällt. Dazu bedarf es einer *räumlichen*, *zeitlichen* und *sachlichen* Abgrenzung. Letztere ist häufig kompliziert. Erfüllt eine Untersuchungseinheit die örtliche, zeitliche und sachliche Begriffsabgrenzung, so soll dies durch die

Relation

$$w \in \Omega$$

ausgedrückt werden.

Beispiel 2.1 Eine Befragung der Studenten der Universität Konstanz soll Aufschluß geben über die von ihnen benutzten Beförderungsmittel und -zeiten zur Universität. Die Grundgesamtheit könnte dann etwa aus allen zu Beginn des Wintersemesters 1991/92 an der Universität Konstanz eingeschriebenen Studenten bestehen. Die räumliche und zeitliche Abgrenzung ist somit klar. Die sachliche Abgrenzung läßt jedoch noch Fragen offen:

- Werden Gasthörer einbezogen?
- Wie werden "Karteileichen" behandelt?
- Sollen auch beurlaubte Studenten erfaßt werden?
- Wie werden Aufbaustudiengänge oder eingeschriebene Doktoranden behandelt?

Über diese Fragen sollte im Hinblick auf das Untersuchungsziel entschieden werden. Schließlich muß von jeder Person ω ausgesagt werden können, ob sie zu der Grundgesamtheit

$$\Omega = \{\omega | \omega \text{ ist zu Beginn des Wintersemesters 91/92 Student der Universität Konstanz}\}$$

gehört oder nicht. ⋈

Für die spätere statistische Analyse ist es relevant, ob Ω eine sogenannte *Bestandsmasse* (englisch: stock) ist, deren Elemente ω eine gewisse zeitliche Verweildauer aufweisen, oder eine *Bewegungsmasse* (flow) vorliegt, bei der die Untersuchungseinheiten ω als Ereignisse gewissen Zeitpunkten zugeordnet werden. Bestandsmassen — wie die Gesamtheit der Studenten der Universität Konstanz — werden zu festgelegten Zeitpunkten gezählt oder gemessen, wohingegen Bewegungsmassen stets auf Zeiträume bezogen angegeben werden. Die Absolventen der Universität Konstanz im Jahre 1991 sind ein Beispiel für eine Bewegungsmasse. Weitere Beispiele für Bestandsmassen sind der Lagerbestand, die Bevölkerung oder die Belegschaft einer Firma; zugörige Ereignismassen sind Lagerzu- und abgänge, Geburts- und Todesfälle, Zu- und Fortzüge oder Einstellungen, Entlassungen und Pensionierungen.

Bei den Untersuchungseinheiten interessieren gewisse, für das Untersuchungsziel relevante *Merkmale* oder *Statistische Variablen*, die man durch die (im allgemeinen vektorielle) Zuordnungsvorschrift

$$\mathbf{X} : \Omega \to S$$

zum Ausdruck bringt. Der *Merkmalsvektor* $\mathbf{X}$ ordnet also jeder Untersuchungseinheit ω die *Merkmalsausprägungen*, *Realisationen* oder *Beobachtungen* $\mathbf{x}$ (ihren "statistischen Schatten") zu. Man schreibt

$$\mathbf{X}(\omega) = \mathbf{x}.$$

Die Merkmalsausprägungen $\mathbf{x}$ liegen im sogenannten *Merkmalsraum* S (der Menge möglicher Merkmalsausprägungen und deren Kombinationen.)

Beispiel 2.2 In Beispiel 2.1 können etwa die Merkmale

(a) das vornehmlich gewählte Verkehrsmittel,

(b) die Anfahrtzeit mit öffentlichen Verkehrsmitteln,

(c) die Zufriedenheit mit den öffentlichen Verkehrsmitteln,

(d) die Semesterzahl und

(e) das Studienfach

von Interesse sein. Ausprägungen dieser Merkmale wären beispielsweise

(a) Bus, Fahrrad, privates Auto, zu Fuß;

(b) 5, 13, 28, 54 Minuten;

(c) sehr, einigermaßen, weniger, überhaupt nicht zufrieden;

(d) 1., 4., 7., 15. Semester;

(e) Mathematik, Physik, Chemie, Volkswirtschaftslehre, Verwaltungswissenschaften, Literaturwissenschaften, Psychologie.

⋈

Nach Art der möglichen Ausprägungen werden *qualitative (artmäßige) Merkmale* mit verschiedenartigen Ausprägungen und *quantitative (zahlenmäßige)* Merkmale, die durch Zahlen erfaßt werden, unterschieden. Für das weitere Vorgehen ist es wichtig, auf welcher *Skala* die Merkmalsausprägungen gemessen werden. Der Anwender statistischer Methoden sollte sich zu Beginn seiner Analysen stets die Frage nach der *Skalierung* der Merkmalsausprägungen stellen, denn diese ist entscheidend für die Transformationsmöglichkeiten der Daten und somit schließlich für die Auswahl eines geeigneten Analyseverfahrens. Nach der Art der möglichen Ausprägungen unterscheiden wir die folgenden Skalenniveaus:

- *Nominalskala:* Die Merkmalsausprägungen sind nicht vergleichbar, und ihre Anordnung drückt keine Rangordnung aus.
- *Ordinal- oder Rangskala:* Die Realisationen sind intensitätsmäßig unterscheidbar und lassen sich je nach Stärke der Intensität ordnen. Man kann also die Rangordnung (nicht aber die Differenzen) der Merkmalsausprägungen interpretieren.
- *Metrische Skala:* Zwischen den Merkmalsausprägungen können zusätzlich Abstände gemessen und interpretiert werden. (auch: *Kardinalskala*). Die metrische Skala kann weiter untergliedert werden in die
  - *Intervallskala:* (Nur) Abstände können verglichen werden.
  - *Verhältnisskala:* Zur Intervallskala kommt ein natürlicher Nullpunkt hinzu.
  - *Absolutskala:* Zur Verhältnisskala kommt eine natürliche Einheit hinzu.

In der Psychologie, der Markt- und Meinungsforschung oder bei Testauswertungen versucht man oft, eine metrische Skala erst aus den Daten der Erhebung zu kontruieren. Komplexe Skalierungsverfahren, wie sie etwa bei der Messung von Einstellungen und Motiven in der empirischen Sozialforschung Verwendung finden, werden in diesem Buch ausgeklammert. Der Hinweis auf Atteslander (1974) möge hier genügen. Die Benutzung natürlicher (oder sogar reeller) Zahlen zur Festlegung von Merkmalsausprägungen impliziert nicht automatisch das Vorliegen einer metrischen Skala. Als Beispiel hierfür sind etwa Schulnoten anzuführen, die zwar eine Ordnung zum Ausdruck bringen, deren Abstände jedoch im allgemeinen nicht zu interpretieren sind. Die Aussage " 'befriedigend' ist im Vergleich zu 'ausreichend' um genau soviel besser wie 'sehr gut' gegenüber 'gut'" ist i.a. nicht zulässig.

Ein Merkmal (eine Variable) $\mathbf{X} : \Omega \to S$ heißt *diskret*, falls S höchstens abzählbar ist (d.h. es gibt endlich oder abzählbar unendlich viele Ausprägugen) und *stetig*, falls S eine (kompakte) Teilmenge des $\mathbb{R}(\mathbb{R}^p)$ ist (d.h. es gibt, zumindest theoretisch, überabzählbar viele mögliche Ausprägungen).

Natürlich ist in der Realität genau genommen wegen der endlichen Meßgenauigkeit jedes Merkmal diskret. Aber selbst bei einer endlichen Anzahl von Elementen des Zustandsraumes ($|S| < \infty$) bietet sich das Modell der stetigen Variablen für die statistische Behandlung eher an, wenn $|S|$ nur genügend groß ist. Man spricht in diesem Fall auch von *quasistetigen Merkmalen*. Beispiele hierfür sind Einkommen und Preise.

Ein metrisches Merkmal heißt *extensiv*, falls sich die Merkmalsausprägungen sinnvoll summieren lassen (z.B. Einkommen) und *intensiv*, falls nur Mittelwerte, nicht aber Summen sinnvoll interpretierbar sind (z.B. Preise).

Beispiel 2.3 In Beispiel 2.2 sind die Merkmale Verkehrsmittel (a) und Studienfach (e) nominal, das Merkmal Zufriedenheit (c) ordinal und die Merkmale Anfahrtszeit (b) und Semesterzahl (d) metrisch skaliert. Die Ausprägungen des Merkmals Semesterzahl sind Anzahlen und besitzen als solche eine natürliche Einheit; sie werden daher auf der Absolutskala gemessen. Die Ausprägungen des Merkmals Anfahrtzeit besitzen einen natürlichen Nullpunkt aber keine natürliche Einheit, da sie etwa in Sekunden, Minuten oder Stunden angegeben werden können. Ihnen liegt eine Verhältnisskala zu grunde, und mithin können Verhältnisse interpretiert werden. Somit ist die Aussage "Student A benötigt die doppelte Zeit für die Anfahrt zur Universität wie Student B" sinnvoll. Als Beispiel für ein Merkmal, dessen Ausprägungen auf der Intervallskala gemessen werden, seien hier in Celsius gemessene Temperaturen angeführt. Zwar können Temperaturdifferenzen, nicht aber Temperaturverhältnisse interpretiert werden. Die Aussage "heute ist es doppelt so kalt wie gestern", wenn gestern $10°C$ und heute $5°C$ gemessen wurden, macht keinen Sinn. Verkehrsmittel (a), Zufriedenheit (c) oder Studienfach (e) sind qualitativer Merkmale, Semesterzahl (d) oder Anfahrtzeit (b) sind quantitative Merkmale. Die Anfahrtzeit (b) ist ein stetiges Merkmal, die restlichen Merkmale aus Beispiel 2.2 sind diskret. ⋈

2.2 Datenerhebung

Das Gewinnen von Daten bezeichnet man als *Erhebung.* Für eine Erhebung ist eine gründliche Erhebungsplanung erforderlich. Erhebungen werden bei den sogenannten *Erhebungseinheiten* e durchgeführt, wobei wiederum eine genaue örtliche, zeitliche und sachliche Begriffsabgrenzung vorzunehmen ist. Die Erhebungseinheiten brauchen nicht mit den Untersuchungseinheiten übereinzustimmen. Die Menge der Erbebungseinheiten bildet die *Erhebungsgesamtheit* E. Häufig gilt $E \subset \Omega$, es kann aber auch $E \cap \Omega = \phi$ gelten. So wird man bei einer Untersuchung über Kinderkrankheiten nicht die Kinder selbst sondern deren Eltern befragen.

Unter Kosten- und zeitlichen Gesichtspunkten ist zu überlegen, ob für die Zwecke der Untersuchung nicht auf Daten aus anderen Quellen zurückgegriffen werden kann. Wird für die vorliegende Untersuchung keine gesonderte Erhebung durchgeführt, so spricht man von einer *Sekundärstatistik.* Beispiele hierfür sind Buchhaltungsunterlagen, Unfallprotokolle oder Krankenblätter. Ferner sollte überprüft werden, ob Daten aus anderweitig durchgeführten Erhebungen — etwa aus den Bereichen der *amtlichen* und *nichtamtlichen Statistik* (vgl. Abschnitt 1.3) — die gewünschten Informationen beinhalten. Eine Sekundärstatistik ist i.a. schneller und billiger zu erstellen, hat aber häufig den Nachteil, daß die den Daten zugrunde liegenden Bregriffsabgrenzungen nicht genau mit denen des Untersuchungsziels übereinstimmen (*Adäquationsproblem*). Entscheidet man sich für eine eigene Erhebung, so ist zu überlegen, ob diese als *Vollerhebung* (Totalerhebung) oder als *Stichprobe* (Teilerhebung) durchgeführt werden soll. Im ersten Fall wird ganz Ω, im zweiten Fall nur eine Teilmenge von Ω erfaßt. Mit der Auswahl geeigneter Teilmengen beschäftigt sich die *Stichprobentheorie.* Eine Stichprobe ist im allgemeinen schneller und billiger, und sie kann sogar genauer sein als eine Vollerhebung, wenn auf die (geringere Anzahl von) Einzelangaben mehr Zeit und Sorgfalt verwendet wird. Dennoch kann man manchmal auf Vollerhebungen — in größeren Zeitabständen — nicht verzichten, schon, um eine Auswahlgrundlage für Stichproben zu schaffen.

Als Arten der Datengewinnung werden

- die Befragung,
- die Beobachtung und

- das Experiment

unterschieden. Während die Beobachtung und das Experiment in naturwissenschaftlichen Anwendungen überwiegen, ist in den Wirtschafts- und Sozialwissenschaften die Befragung die häufigste Erhebungsmethode.

Die Befragung

Befragungen können

a) *schriftlich* (durch Fragebogen)

b) *mündlich* (durch persönliche oder telefonische Interviews)

c) *kombiniert* schriftlich / mündlich

durchgeführt werden. Bei der *Fragebogenerstellung* sind erhebungstechnische und psychologische Gesichtspunkte wie etwa die Reihenfolge der Fragen oder die Frageformulierung wesentlich. Von großer Wichtigkeit ist dabei die Einfachheit und Eindeutigkeit der Fragestellung. Eventuell empfiehlt es sich, zum "Testen" des Fragebogens eine *Probeerhebung* durchzuführen. Bei der schriftlichen Form der Befragung kann es durch niedrige Antwortquoten zu einer Verzerrung der Ergebnisse kommen, wenn sich die Gruppen der Beantworter und Antwortverweigerer bezüglich der interessiernden Merkmale systematisch unterscheiden.

Das *Interview* ist teurer, kann jedoch kompliziertere Fragestellungen behandeln und hat eine niedrigere Verweigerungsquote. Hier kann jedoch der Interviewer einen nicht zu unterschätzenden Einfluß auf das Antwortverhalten der befragten Personen haben *(Interviewer-Verzerrung)*. Eine billigere und schnellere Alternative zu persönlichen Interviews sind *Telefoninterviews*, bei welchen auch eine Reduktion des Interviewereinflusses zu erwarten ist. Wesentliche Faktoren für eine Verzerrung bilden hier die Unterschiede in der Erreichbarkeit für bestimmte Gruppen von Erhebungseinheiten (z.B. Hausfrauen und Berufstätige).

Bei der Durchführung des Interviews können die folgenden Formen unterschieden werden: Beim *standardisierten Interview* ist sowohl der Wortlaut der Fragen als auch deren Reihenfolge festgelegt. Dieses verringert den Interviewereinfluß und macht die Ergebnisse besser

vergleichbar. Das *nichtstandardisierte Interview* mit freier Frageformulierung ermöglicht ein individuelles Eingehen des Interviewes auf spezifische Besonderheiten des Befragten und kann weitergehende Informationen etwa über das Persönlichkeitprofil der Befagten offen legen. Eine geeignete statistische Analyse auf Grund solcher Ergebnisse ist dann aber sehr schwierig. Ferner unterscheidet man *weiche* und *harte Interviews* je nach dem, ob der Interviewer passiv und geduldig die Antworten abwartet oder die Fragen Schlag auf Schlag folgen, um spontane Antworten zu provozieren.

Damit die Ergebnisse einer Befragung überhaupt sinnvolle Aufschlüsse geben können ist es wesentlich, daß die Interviewer gut informiert und geschult werden, denn von ihnen hängen Erfolg und Mißerfolg einer Studie ganz wesentlich ab.

> "Man glaubt gar nicht, welcher Unsinn entstehen kann, wenn ohne Beteiligung geschulter Fachleute Erhebungen durchgeführt werden." (Wagemann, 1935)

Die Beobachtung

Neben Anwendungen in den Ingenieur- und Naturwissenschaften spielt die Beobachtung etwa auch in der Verkehrserfassung, bei innerbetrieblichen Studien, in der Psychologie, in der Marktforschung und der Medizin eine wesentliche Rolle. In den letztgenannten Disziplinen werden die Informationen unabhängig von der Auskunftsbereitschaft des Beobachteten gewonnen, so daß das Problem der Antwortverweigerung entfällt. Nur die "statistischen Schatten" der beobachteten Phänomene werden registriert. In der Psychologie unterscheidet man die *offene* und die *verdeckte* Beobachtung je nach dem, ob die beobachtete Person von der Beobachtung in Kenntnis gesetzt worden ist oder nicht. Verdeckte Beobachtungen haben den Vorteil, daß die Information für das Beobachtungsobjekt unbewußt ermittelt wird und somit dessen natürliches Verhalten widerspiegelt.

Das Experiment

Experimente überwiegen in den Naturwissenschaften wie etwa in der Landwirtschaft, der Biologie, der Medizin oder der Pharmazie, finden sich aber auch in der Produktplanung. Sie zeichnen sich dadurch aus, daß sich die interessierenden Einflußgrößen unter der Kontrolle

des Forschers befinden. Mit der Durchführung beschäftigt sich eine eigenständige Disziplin, die *statistische Versuchsplanung.* In einem *Versuchsplan* werden

- die Behandlungsarten,
- die Anzahl der Objekte,
- die Zuordnung von Behandlungsarten zu Objekten sowie
- die Meßvorschrift, nach der die Ergebnisse gewonnen und schließlich ausgewertet werden,

festgelegt.

2.3 Aufbereitung des erhobenen Datenmaterials

Nach Durchführung der Erhebung erfolgt die Zusammenfassung der Daten in *Tabellen* und (eventuell) ihre Veranschaulichung durch *Abbildungen.*

Da je Untersuchungseinheit im allgemeinen eine größere Anzahl von Merkmalen erfaßt wird, muß im *Aufbereitungsprogramm* festgelegt werden, welche Variablenpaare, Tripel, usw. aus dem Merkmalsvektor gemeinsam tabellarisch aufbereitet werden. Diese Auswahl richtet sich natürlich nach der Zielsetzung der Untersuchung. Dabei sollte stets beachtet werden, daß nach Vernichtung des sogenannten *Urmaterials* (Fragebogen, Interviewprotokolle usw.) unberücksichtigt gebliebene Merkmalskombinationen später nicht mehr untersucht werden können. Wenn möglich, sollten die Daten in einer Datenmatrix $\mathcal{X}$ der Form

$$\mathcal{X} = (x_{jk})_{j=1,\ldots,n;\ k=1,\ldots,p} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \tag{2.1}$$

vollständig abgespeichert werden. Diese Form der Abspeicherung wird von den meisten statistischen *Softwarepaketen* unterstützt. Der Eintrag x_{jk} in der Datenmatrix bedeutet die Ausprägung des k-ten Merkmals, die an der j-ten Untersuchungseinheit gemessen wurde. Die j-te Zeile $(x_{j1}, x_{j2}, \ldots, x_{jp})$ gibt alle für die j-te Untersuchungseinheit erhobenen Daten

an, und die k-te Spalte $(x_{1k}, x_{2k}, \ldots, x_{nk})'$ enthält die Ausprägungen des k-ten Merkmals für alle Untersuchungseinheiten. Qualitative Merkmale können numerisch *kodiert* werden. Die Datenaufbereitung sollte auch mit einer *Plausibilitätskontrolle* verbunden sein.

Beispiel 2.4 Für die Merkmale Verkehrsmittel (X_1), Anfahrtzeit (X_2), Zufriedenheit (X_3), Semesterzahl (X_4) und Studienfach (X_5) in Beispiel 2.2 wäre etwa die folgende Datenmatrix denkbar:

$$\mathcal{X} = (x_{j,k})_{j=1,\ldots,10;\ k=1,\ldots,5} = \begin{pmatrix} 1 & 17 & 2 & 5 & 5 \\ 0 & 25 & 3 & 7 & 1 \\ 2 & 31 & 4 & 3 & 9 \\ 3 & 13 & 2 & 8 & 16 \\ 1 & 21 & 1 & 1 & 11 \\ 2 & 12 & 3 & 2 & 15 \\ 0 & 12 & 2 & 11 & 3 \\ 1 & 10 & 2 & 9 & 2 \\ 1 & 21 & 1 & 3 & 10 \\ 3 & 12 & 1 & 7 & 2 \end{pmatrix} \tag{2.2}$$

Sie enthält die Ausprägungen für 10 befragte Studenten, wobei die in Tabelle 2.1 angegebenen Kodierungen für qualitative Variablen verwendet wurden. ⋈

Bei der Aufbereitung von Daten in Tabellen sind bei stetigen Merkmalen immer, bei diskreten Merkmalen oft *Gruppen* (*Klassen*) zu bilden. Dabei ist jedoch zu beachten, daß mit jeder Klassenbildung ein gewisser Informationsverlust einhergeht. In wissenschaftlichen Arbeiten unterscheidet man *Quellentabellen* und *Aussagetabellen*. Erstere gehören in den Anhang. Letztere bilden Ausschnitte oder Zusammenfassungen und erscheinen im Text, sollten aber möglichst für sich selbst lesbar und verständlich sein. Die Überschrift sollte zumindest den Untersuchungsgegenstand sachlich, zeitlich und räumlich abgrenzen. Zum Schema einer Tabelle siehe die Abbildung 2.1.

Für Tabellen der *amtlichen Statistik* gelten die folgenden Vereinbarungen. Es bedeutet:

Tabelle 2.1: Kodierung der qualitativen Variablen Verkehrsmittel, Zufriedenheit und Studienfach

Merkmal	Ausprägung	Kodierung
Verkerhrsmittel	zu Fuß	0
	Bus	1
	Auto	2
	Fahrrad	3
	sonstige	4
Zufriedenheit	sehr zufrieden	1
	einigermaßen zufrieden	2
	weniger zufrieden	3
	überhaupt nicht zufrieden	4
Studienfach	Mathematik	1
	Physik	2
	Chemie	3
	Biologie	4
	Psychologie	5
	Soziologie	6
	Verwaltungswissenschaft	7
	Politikwissenschaft	8
	Erziehungswissenschaft	9
	Sportwissenschaft	10
	Wirtschaftswissenschaften	11
	Jura	12
	Phylosophie	13
	Geschichte	14
	Literaturwissenschaft	15
	Sprachwissenschaft	16

Die Anfahrtzeit ist in Minuten angegeben.

Abbildung 2.1: Schema einer Tabelle:

—Überschrift—

Variablen- Vari- angabe ablenangabe			Tabellenkopf		ins- ge- samt
Vorspalte	Zahlenteil				
insgesamt					

Fußnoten

Quellenangabe

- genau Null

x Eintragung aus sachlichen Gründen nicht möglich

0 von Null verschieden, aber weniger als die Hälfte der kleinsten ausgewiesenen Einheit

. unbekannt oder geheim

... Daten noch nicht verfügbar

p vorläufige Daten

r berichtigte Daten

s geschätzte Daten

bei Stichprobenerhebungen:

/ keine Angabe möglich, da das Ergebnis nicht ausreichend gesichert ist

( ) Angabe unter dem Vorbehalt, daß das Ergebnis erhebliche Fehler aufweisen kann.

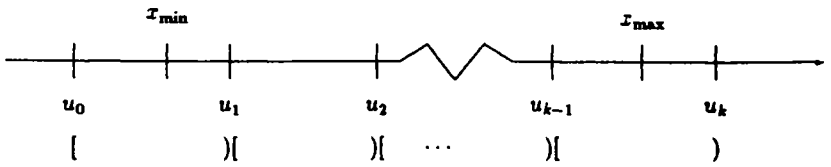
Kapitel 3

Graphische Darstellungen univariater Daten

Zunächst werden Methoden für die Behandlung eines einzigen Merkmals (*univariate Daten*) vorgestellt. Obwohl in aller Regel bei einer Untersuchung mehrere Merkmale (*multivariate Daten*) erhoben werden, können auch hier für jedes einzelne Merkmal (jede einzelne Spalte der Datenmatrix $\mathcal{X}$) Verfahren für univariate Daten angewandt werden. Dies liefert des öfterern schon wesentliche Aufschlüsse über die Struktur der Daten und weist gegebenenfalls auf offenkundige Unstimmigkeiten hin. Zur expliziten Angabe der Verfahren bedarf es einiger Notationen, die nun eingeführt werden sollen. Mit n wird der *Erhebungsumfang (Stichprobenumfang)* — also die Anzahl der befragten Personen oder untersuchten Objekte — bezeichnet. $x_j = X(\omega_j)$ ist die Merkmalsausprägung bei der j-ten Untersuchungseinheit, $j = 1, \ldots, n$. Gibt es eine verhältnismäßig kleine Anzahl k von Merkmalsausprägungen, wie es bei nichtmetrischen oder diskreten Merkmalen häufig der Fall ist, so weisen im allgemeinen mehrere Untersuchungseinheiten ein und dieselbe Ausprägung auf. Ist $X(\Omega) = \{a_1, a_2, \ldots, a_k\}$ die Menge aller vorkommenden Merkmalsausprägungen, so ist es günstiger, anstelle der n Messungen x_j, $j = 1, \ldots, n$, lediglich die k Ausprägungen a_i, $i = 1, \ldots, k$, und die Anzahl der Fälle n_i zu notieren, in denen a_i beobachtet wurde. In Termen der sogenannten Indikatorfunktion der Menge A, $\mathbf{1}_A(x) = 1$, falls $x \in A$, $\mathbf{1}_A(x) = 0$ sonst, ist

$$n_i = \sum_{j=1}^{n} \mathbf{1}_{\{a_i\}}(x_j), \quad i = 1, \ldots, k \quad . \tag{3.1}$$

Abbildung 3.1: Intervallaufteilung bei Klassenbildung



Die Anzahlen n_i, $i = 1, \ldots, k$, heißen *absolute Häufigkeiten.* Zum Vergleich von Datensätzen mit unteschiedlichen Erhebungsumfängen sollte man anstelle der absoluten die *relativen Häufigkeiten*

$$r_i = \frac{n_i}{n}, \quad i = 1, \ldots, k,$$

verwenden, welche den Anteilen der Fälle entsprechen, die auf die Ausprägungen a_i entfallen.

Zur Datenaufbereitung für *stetige oder quasistetige Merkmale* bedarf es im allgemeinen einer *Klassen- (Gruppen-) bildung*: Sei dazu x_{min} die kleinste und x_{max} die größte vorkommende Merkmalsausprägung. Zunächst legt man eine geeignete Anzahl k der Klassen (vgl. dazu Abschnitt 3.4) sowie geeignete Klassengrenzen $u_0 < u_1 < \ldots < u_{k-1} < u_k$ mit $u_0 \leq x_{min}$ und $x_{max} < u_k$ fest (vgl. Abbildung 3.1). Die $i-te$ Klasse ist dann durch das Intervall $[u_{i-1}, u_i)$ (lies: von u_{i-1} bis unter u_i) gegeben und besitzt die *Klassenbreite (Gruppenbreite)*

$$b_i = u_i - u_{i-1}, \qquad i = 1, \ldots, k.$$

$$n_i = \sum_{j=1}^{n} \mathbf{1}_{[u_{i-1}, u_i)}(x_j)\,, \qquad i = 1, \ldots, k,$$

ist dann die Anzahl der Beobachtungen (Messungen) in der Klasse i. Entsprechend ist

$$r_i = n_i/n$$

die relative Häufigkeit der Fälle in dieser Klasse. Eine Klassenbildung ist häufig auch bei diskreten Merkmalen, die viele unterschiedliche Ausprägungen besitzen, sinnvoll. Ein Beispiel hierfür wäre etwa die Beschäftigtenzahlen von Betrieben.

3.1 Klassische Methoden graphischer Darstellung

Mit Hilfe des *Stabdiagramms, (Säulendiagramms)* können Häufigkeitsverteilungen diskreter Merkmale veranschaulicht werden. Die (absoluten oder relativen) Häufigkeiten werden über