Walter Olbricht

Statistik zum Mitdenken

Ein Arbeits- und Übungsbuch

3., erweiterte und aktualisierte Auflage

Kohlhammer

Kohlhammer

Walter Olbricht

Statistik zum Mitdenken

Ein Arbeits- und Übungsbuch

3., erweiterte und aktualisierte Auflage

Verlag W. Kohlhammer

3., erweiterte und aktualisierte Auflage 2017

Gesamtherstellung: W. Kohlhammer GmbH, Stuttgart
Umschlagfoto: © tiero – Fotolia.com

Print:
ISBN 978-3-17-033289-8

E-Book-Formate:
pdf: ISBN 978-3-17-033290-4

Vor*sicht*,

Sie haben soeben ein altmodisches Lehrbuch zur Hand genommen. Es wird nämlich nicht dem modernen Anspruch gerecht, dass der Leser hierdurch wie durch einen „Nürnberger Trichter“ den Lernstoff schnell, spielerisch und vor allem anstrengungslos aufnimmt. Das funktioniert – jedenfalls nach der Erfahrung des Autors – sowieso nur, wenn lediglich einfache Informationen vermittelt werden sollen (und auch dann nur sehr begrenzt). Soll eine Kompetenz erworben werden, liegt die Hauptarbeit nicht beim Vermittler, sondern beim Erwerber, und sie kann ihm zwar erleichtert, aber nicht abgenommen werden. Nehmen wir einmal das Radfahren. Natürlich kann man durch Schaubilder oder sogar Videos hervorragend über die Gleichgewichtsproblematik informiert werden. Radfahren lernt man so indessen nicht. Dazu muss jeder – im Wortsinne – seine eigenen Er*fahr*ungen machen und u. U. unsanft mit der Gleichgewichtsproblematik Bekanntschaft schließen. Es gibt aber auch einen Lohn für die Mühe beim Kompetenzerwerb. Denn erstens ist „Radfahren können“ viel mehr als „von Radfahren wissen“ und zweitens: Radfahren verlernt man nicht.

Wenn man Radfahren lernt, können Stützräder und gute Anleitungen helfen – und in analoger Weise liegt hier der Beitrag dieses Buches. Wem dieser Ansatz einleuchtend erscheint, sollte einmal einen kurzen Blick in das erste Kapitel (Wegweiser) werfen, wo dieses Konzept noch detaillierter erläutert wird. Jedenfalls deuten schon diese wenigen Anmerkungen an, warum es sich bei diesem Werk nicht um ein Statistikbuch mit Aufgaben, sondern um Statistikaufgaben mit einem Buch(teil) handelt. Letzterer ist so gehalten, dass das Buch bei Bedarf auch als eigenständige Kursunterlage oder zum Selbststudium benutzt werden kann. Ideal dürfte eine Nutzung als Ergänzung und Vertiefung eines einführenden Statistikkurses sein. Spezielle Vorkenntnisse sind nicht erforderlich.

Das Buch ist aus einführenden Statistikveranstaltungen hervorgegangen, die ich in den vergangenen Jahren an der Universität Bayreuth für Hörer aller Fachbereiche gehalten habe. Es stützt sich im Kern auf die Hauptklausuren der letzten Jahre, die nahezu unverändert übernommen wurden. Insofern haben auch Freunde, Kollegen, Mitarbeiter und Studierende (– mit solchen übergreifenden Bezeichnungen sind hier und im ganzen Buch Männer und Frauen gleichermaßen gemeint –) in verschiedener Weise zu seiner endgültigen Form beigetragen, etwa durch Hinweise und nützliche Diskussionen, durch Proberechnen von Klausuren oder durch hartnäckiges Nachfragen. Im Einzelnen lässt sich das nicht mehr zurückverfolgen oder gar namentlich zuordnen,

aber ihnen allen sei an dieser Stelle herzlich gedankt. Herr Dr. U. Fliegauf (Verlagsleitung des Verlages Kohlhammer) hat das Buch in jeder Hinsicht in vorzüglicher Weise betreut; Herr Dr. D. Kirn (Lektorat des Verlages Kohlhammer) half mit Rat und Tat bei der Bewältigung von Schwierigkeiten des Zweifarbendrucks. Beiden Herren und dem Verlag W. Kohlhammer GmbH danke ich für die stets reibungslose und angenehme Kooperation. Meine Frau Katrin hat mir (nicht nur) bei der Abfassung dieses Buches den nötigen Rückhalt gegeben. Ihr ist dieses Buch gewidmet.

Bayreuth, im September 2011 Walter Olbricht

Vorwort zur dritten Auflage

Die dritte Auflage dieses Buches wurde genutzt, um einige besonders hartnäckige Druckfehler zu beseitigen, die sich in den ersten beiden Auflagen geschickt verborgen hielten, und ich möchte allen herzlich danken, die mich auf solche Unzulänglichkeiten hingewiesen haben. Um dem Druckfehlerteufel aber auch dieses Mal wieder eine faire Chance zu geben, wurde die vorliegende Auflage um gleich drei neue Klausuren erweitert. Das Leitmotiv der vorletzten Klausur (in Abschnitt 3.15) erscheint mir dabei besonders gelungen, seit mir eine Teilnehmerin mitteilte, dass sie seither jedes Mal an Statistik denken muss, wenn sie Schokolade isst. Nachhaltiger kann Lehre kaum sein...

Auch bei dieser Auflage stand mir mit Herrn Dr. U. Fliegauf (Verlagsleiter GW des Verlages Kohlhammer) und Herrn Dr. D. Kuhn (Lektorat Geschichte des Verlages Kohlhammer) das bewährte Team bei allen Schwierigkeiten immer hilfsbereit und hilfreich zur Seite. Beiden Herren und dem Verlag W. Kohlhammer GmbH möchte ich für die stets angenehme, nunmehr schon sehr langjährige Zusammenarbeit herzlich danken.

Bayreuth, im September 2017 Walter Olbricht

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Wegweiser

1.1 Ausgangspunkt

Statistikklausuren sind ein Ärgernis – das finden in überraschender Einmütigkeit viele Studierende und Dozenten von Statistikkursen für Hörer verschiedenster Fächer. Allerdings mit unterschiedlicher Begründung: Die Studierenden monieren, dass sie leere Formalismen pauken müssen, deren Sinn ihnen verschlossen bleibt und die sie nach bestandener Klausur baldmöglichst vergessen. Die Dozenten bedauern, dass die Studierenden auf die Klausuren fixiert und an dem „eigentlichen" Stoff gar nicht interessiert sind. Diese doppelt beklagenswerte Situation enthält freilich auch eine Chance und trägt gewissermaßen schon einen Schlüssel zu ihrer Auflösung in sich. Wenn nämlich die Klausuraufgaben gerade den „eigentlichen" Stoff einfordern, also das widerspiegeln, was vermittelt werden soll, ist die Motivation der Studierenden von alleine in die richtige Richtung gelenkt. Wenn zudem dieser „eigentliche" Stoff auch einsehbar praxisrelevant ist, werden die Studierenden ihn auch nicht als leeren Formalismus empfinden.

Erfolgreiche Lehre für Anwender muss demnach bei den Klausuraufgaben ansetzen. Das ist der Ausgangspunkt dieses Buches, das auf 16 Klausuren basiert, die der Autor mit eben dieser Ausrichtung in den vergangenen Jahren gestellt hat. Die Klausuraufgaben sind meist als kleine Fallstudien konzipiert und manchmal sogar direkt aus der Tagespresse entlehnt. Ihre Relevanz ist ziemlich leicht erkennbar. In der Tat benötigt man statistische Grundkenntnisse in nahezu allen Lebensbereichen. Man betrachte dazu einmal die Titelseite einer beliebigen (renommierten) Tageszeitung und ermittele – schon wieder eine statistische Information! – den Prozentsatz, den statistische (also im weitesten Sinne quantitative) Information dort einnimmt. Schon um diese Information sinnvoll nutzen und bewerten zu können, benötigt man statistische Kenntnisse – man benötigt sie allerdings im aktiven Wissensschatz, nicht als halbvergessene tote Formeln von „damals".

Natürlich klingt das fast zu schön, um wahr zu sein. Und in der Tat wird man schon wegen der Prüfungssituation einige Kompromisse machen müssen. So lassen sich ganz einfach nicht alle für die statistische Praxis relevanten Fertigkeiten in Aufgaben fassen. Auch müssen Klausuraufgaben in vertretbarer Zeit korrigierbar sein und sollten – schon aus Gründen der rechtlichen Überprüfbarkeit – scharf umrissene Lösungen zulassen. Man wird das Ideal also sicher nicht ganz erreichen. Aber das ist kein Grund, sich nicht wenigstens in die richtige Richtung zu bewegen.

1.2 „Statistik-Fahrschule“: 16 Doppelstunden und etwas Theorie

Das Konzept dieses Buches lässt sich gut mit der Analogie zur Fahrschule beschreiben. Dort gibt es im Wesentlichen drei Komponenten: die Fahrstunden, die theoretische Ausbildung und den Fahrlehrer. Gehen wir sie einmal der Reihe nach durch.

1.2.1 Die Fahrstunden

Wohl für jeden Fahrschüler sind diese das Herzstück seiner Ausbildung. Man stelle sich einmal vor, die Fahrausbildung bestünde nur in theoretischer Unterweisung oder nur im praktischen Drill gewisser Standardtechniken (etwa Einparken oder Tachometer ablesen)! (Die Analogie zu gewissen Statistikkursen und vor allem Statistikklausuren kann jeder Leser selbst bilden.) Das Besondere an den Fahrstunden ist gerade, dass sie den Fahrschüler dem realen Verkehr mit seinen ständig neuen Situationen aussetzen und ihn zwingen, viele verschiedene Standardtechniken und Standardwissen zu koordinieren und situationsgemäß einzusetzen. Nicht zuletzt deswegen ist man in den ersten Stunden komplett überfordert. Später hat man den Durchblick erworben und das Zusammenspiel gelernt. Dann geht es wie von selbst, und man kann nebenher Radio hören.

Die folgenden Übungsklausuren sind ähnlich gehalten. Die Teilnehmer werden darin, soweit möglich, den Bedingungen der Praxis ausgesetzt. Es geht nicht darum, nur schematisch etwas auszurechnen. Gute Statistik besteht ja (wie gutes Autofahren) gerade darin, sich in unübersichtlichen Situationen zurechtzufinden und die Aufmerksamkeit auf das jeweils Wichtige zu lenken. Wie die erste Fahrstunde mag der Leser auch die erste Klausur als unübersichtlich und als Überforderung empfinden, weil sie nicht schematisch ist. Aber genau auf die Fähigkeit, damit zurande zu kommen, kommt es an. Weil eben auch die Realität nicht schematisch ist.

Aus diesem Grunde sind die Klausuraufgaben auch als ganze Klausuren belassen und nicht etwa nach Themengebieten sortiert. Denn wenn man weiß, welches Themengebiet gerade behandelt wird, ahnt man meist schon die Lösung. Diese scheinbare Unübersichtlichkeit ist also ganz bewusst gewählt. Der Leser – besser: Löser – muss selbst herausfinden, welche Technik jeweils angebracht ist – ganz wie im richtigen Leben.

1.2.2 Die theoretische Ausbildung

Diese ist idealerweise bereits durch einen Statistikkurs erfolgt oder erfolgt parallel. Der Autor würde dazu das hervorragende Lehrbuch [6] von Freed-

man/Pisani/Purves empfehlen, aber auch für jeden anderen Kurs ist das vorliegende Buch eine nützliche Begleitlektüre. Wo Grundkenntnisse fehlen oder aufgefrischt werden sollten, kann dies durch einen Blick in das Auffrischungskapitel geschehen. Dort sind eigentlich alle benötigten Kenntnisse in knapper Form zusammengestellt, so dass das Buch auch ganz eigenständig benutzt werden kann. Das Auffrischungskapitel dient zudem dem Zweck, die Sprachregelungen zu vereinheitlichen und einige besondere Begriffe dieses Kurses (wie z. B. „zwetschgenförmiges Streuungsdiagramm“) einzuführen. Eine weitere kleine Eigenheit des Buches ist es, dass als Dezimaltrennzeichen statt eines Kommas durchgängig (außer in wörtlichen Zitaten) ein Punkt verwendet wird. Der Grund dafür ist, dass in der Statistik oftmals Programmpakete benutzt werden, die nur in international üblicher Form vorliegen. Wer Lehrbücher in deutscher Sprache bevorzugt, kann zu vielen bewährten Titeln – wie beispielsweise [1], [5], [8] oder [13] – greifen. Ein sehr ausführliches und lesenswertes neueres Buch ist [7]. Auch ältere Darstellungen – wie [10], [11] oder [17] – lohnen wegen ihrer vielen guten Gedanken und Beispiele nach wie vor einen Blick. Ganz modern und ganz gezielt kann man sich zu Einzelfragen im Internet kundig machen, wobei der Autor insbesondere Wikipedia hervorheben möchte.

1.2.3 Der Fahrlehrer

Der Fahrschüler sitzt nicht allein im Auto. Neben ihm sitzt der Fahrlehrer und passt auf. Natürlich haben in diesem Buch die Lösungsvorschläge diese Funktion. Der Lernende kann hieraus selbst ersehen, inwieweit er wichtige Aspekte der Aufgabenstellung erfasst hat. Die Lösungen sind absichtlich sehr ausführlich gehalten. Nach Erfahrung des Autors gibt es nämlich viele Defizite bei der Darstellung von Sachverhalten, die man „im Prinzip“ verstanden hat. Zwischen einer numerisch richtigen Rechnung oder einem Schlagwort und einer akzeptablen Präsentation von Ergebnissen können Welten liegen. Aber Letzteres ist eine für die Anwendung der Statistik im Berufsleben unabdingbare Kompetenz. Deswegen werden Multiple-Choice-Aufgaben auch weitgehend vermieden. Wer nur Multiple-Choice-Aufgaben anklickt – dies ist oft bei Lernprogrammen der Fall – entwickelt nicht die Fähigkeit, Sachverhalte eigenständig und korrekt zu formulieren. (Auch die Fragen- und Antwortenkataloge etwa für Abiturprüfungen sind nach Meinung des Autors in dieser Hinsicht eher kontraproduktiv, weil dabei meist nur ein Rechenergebnis, aber keine ausführliche verbale Herleitung angegeben wird.) Idealerweise sollte der Leser also die Lösung selbst ausformulieren und erst nachher mit dem Lösungsvorschlag vergleichen. Dies ist übrigens auch der Grund, weshalb die Lösungen nicht unmittelbar unterhalb der jeweiligen Aufgabe angegeben sind. Es ist sonst noch schwerer, der Versuchung zu widerstehen, schon einmal in die Lösung zu schauen.

Kurz und knapp: Die hier gesammelten Klausuren legen Wert auf korrektes Denken und korrektes Darstellen. Wenig Wert wird auf kompliziertes Rechnen gelegt, denn das kann der Computer besser. Deswegen sind alle Klausuren auch „im Kopf“ (d. h. ohne Taschenrechner) zu bearbeiten. Lediglich die im Anhang A abgedruckte Tabelle der Normalverteilung wird benötigt. Das Buch kann insofern überall – insbesondere in Pausen oder bei Bahnfahrten – schnell hervorgeholt und benutzt werden.

Zusätzlich geben viele Fahrlehrer mehr oder weniger kluge Weisheiten an den Fahrschüler weiter (z. B. über „Herren mit Hut“ am Steuer oder ähnliches). Auch dafür haben wir ein Analogon: die Kommentare zu einigen Lösungen. Hier finden sich Anmerkungen, die die jeweilige Aufgabenlösung in einen weiteren Kontext einordnen, auf zusätzliche wichtige Aspekte aufmerksam machen oder die der Autor einfach irgendwie loswerden wollte. Sie gehören nicht wirklich zur Aufgabenlösung, können aber für manche Leser ganz besonders wertvoll sein. Formal sind sie durch farbige Unterlegung gekennzeichnet, damit man sie entweder leicht finden oder leicht überspringen kann. Hier ist ein Beispiel:

> Kommentar: Es ist gut, bei Klausuren möglichst viele und „in Flensburg“ möglichst wenige Punkte zu erzielen.

Ein mitteilungsfreudiger Fahrlehrer kommentiert nicht nur während der Fahrstunden, sondern auch während des theoretischen Unterrichts. Analog gibt es Kommentare auch im Auffrischungskapitel.

1.2.4 Fazit

Der Ansatz dieses Buches unterscheidet sich möglicherweise von anderen Kursen, die Sie besuchen. Es geht hier nicht so sehr um Informationsvermittlung, sondern um Kompetenzerwerb. Sie sollen also nicht in erster Linie über etwas „orientiert“ oder „informiert“ werden. Der Anspruch ist vielmehr, sich eine neue Fertigkeit so anzueignen, dass man sie in verschiedensten Situationen einsetzen kann. Wie beim Schreiben und Lesen oder beim Autofahren ist das zeitaufwendig und erfordert Übung. Aber es lohnt sich. Kompetenzen verlernt man – im Unterschied zu Informationen – auch nicht so schnell wieder.

Die Analogie zum Schreiben und Lesen oder zum Autofahren ist noch in einer anderen Hinsicht passend: Statistisches Denken ist ebenfalls eine Schlüsselqualifikation, die in den verschiedensten Gebieten und Kontexten benötigt wird.

1.3 Tipps zum Umgang mit diesem Buch

Arbeiten Sie mit diesem Buch, um statistisches Denken für Ihre sonstige Arbeit zu erlernen oder als Vorbereitung, um eine Pflichtklausur zu bestehen? Glücklicherweise stellt sich diese Frage gar nicht, weil die Empfehlungen in beiden Fällen gleich lauten. Denn unser Ausgangspunkt ist ja gerade, dass es keinen Unterschied zwischen den Klausuren und der eigentlich benötigten Statistik und folglich auch keine Diskrepanz zwischen den beiden oben genannten Zielen geben sollte – außer natürlich dem formalen Prüfungscharakter einer Klausur. Die Empfehlungen in Abschnitt 1.3.1 sind also für alle Leser gültig. In Abschnitt 1.3.2 gibt es dann noch einige spezielle Tipps für die Stresssituation „Klausur“.

1.3.1 Tipps zur Klausurvorbereitung

1. Blättern Sie als erstes das Auffrischungskapitel durch. Das meiste wird Ihnen aus Ihrem Statistikkurs bekannt sein oder unmittelbar einleuchten. Sollten dennoch Lücken verbleiben, können Sie diese gegebenenfalls durch die empfohlene Literatur füllen.

2. Bearbeiten Sie dann die Übungsklausuren. Diese sind in keiner speziellen Reihenfolge angeordnet; daher spielt es überhaupt keine Rolle, mit welcher Sie beginnen. Ähnlich wie bei Fahrstunden kommt es nur darauf an, sich den fallstudienartigen Aufgaben auszusetzen und daraus zu lernen.

3. Am meisten profitieren Sie, wenn Sie die Lösungen selbst ausarbeiten und erst danach mit den Lösungsvorschlägen vergleichen. Sie benötigen nur Bleistift und Papier sowie die Normalverteilungstabelle im Anhang A. Alle Rechnungen lassen sich leicht ohne Taschenrechner ausführen. Es ist sinnvoll, die Lösungen recht ausführlich auszuarbeiten und nachher selbstkritisch zu prüfen, welche Einzelschritte man erkannt oder vielleicht übersehen hat.

4. Geben Sie nicht zu schnell auf, wenn Ihnen die Klausuren zu Beginn schwer fallen. Denken Sie an Ihre erste Fahrstunde! Nach einigen Klausuren werden Sie von alleine mehr Übersicht auch in komplizierteren Fragen entwickeln. Das ist genau die Kompetenz, die wir entwickeln wollen. Denken Sie auch daran, dass Kompetenzerwerb mehr Zeit und Übung benötigt als reine Informationsvermittlung.

5. Sollten Sie gezielt ein Themengebiet durch Aufgaben wiederholen wollen oder umgekehrt bei einer Aufgabe ganz festhängen, können Sie die Kreuzreferenztabelle aus Anhang B zu Rate ziehen. Davon sollten Sie aber nur sehr sparsam Gebrauch machen. Deswegen wurde diese Tabelle

auch bewusst auf die Aufgaben der Klausuren mit den Abschnittsnummern 3.1, 3.3, 3.5, 3.7, 3.9 und 3.11 beschränkt.

1.3.2 Tipps zur Klausurbearbeitung

Statistikklausuren sind Prüfungs- und Stresssituationen. Deswegen sind hier noch einige Tipps zusammengefasst, die der Autor beim Betreuen und Auswerten von vielen Klausuren immer wieder bestätigt gefunden hat.

1. Lesen Sie die Aufgaben genau durch. Oftmals ist es viel einfacher als man glaubt.

2. Bearbeiten Sie diejenigen Aufgaben, zu denen Sie eine Lösung direkt sehen, als erste. Sie schaffen sich so ein Polster und können gelassen und mit einem Erfolgserlebnis an die schwierigeren Aufgaben gehen. Zudem sind Sie für die weiteren Aufgaben durch eine positive Erfahrung „beflügelt“. Wenn Sie mit (für Sie) schwierigen Aufgaben beginnen, haben Sie nachher eventuell nicht genügend Zeit und geraten in Panik. Außerdem können Sie die negative Erfahrung mit der schwierigen Aufgabe vielleicht nicht schnell genug überwinden.

3. Verbeißen Sie sich in einer realen Klausur nicht in eine Aufgabe, mit der Sie nicht zurechtkommen. Jede Klausur enthält einige Redundanz. Es wird also nicht erwartet, dass man alle Aufgaben bearbeitet. In jeder Klausur können 120 Punkte erreicht werden. Die Punktezahl einer Aufgabe gibt in etwa die Bearbeitungszeit in Minuten an, falls man mit dem Stoff gut vertraut ist. Viel mehr Zeit sollten Sie (zumindest nach einer Gewöhnungsphase) nicht darauf verwenden.

4. Wenn Sie zu Nervosität bei Klausuren neigen, ist es eine gute Idee, die Übungsklausuren als „echte Probedurchläufe“ zu inszenieren. Schirmen Sie sich dazu für genau 120 Minuten von Störungen ab, und stellen Sie sich vor, die Übung sei schon der Ernstfall. Besser noch: Machen Sie dies in einer Arbeitsgruppe, in der dann später Ihre Klausur von jemand anderem korrigiert wird. Es ist einfach etwas anderes, ob man eine Lösung nur für sich notiert oder für eine zweite Person aufschreibt. Zudem hilft die inszenierte Klausursituation tatsächlich, eine gewisse „Routine“ und damit mehr Gelassenheit auch für die reale Klausur zu erreichen. Das Motto ist hier: Es ist besser, sich vorher selbst ein wenig unter Druck zu setzen, als nachher wirklich unter Druck zu kommen.

2 Auffrischungen aus der Theorie

2.1 Grundlagen

2.1.1 Ein einführendes Beispiel statistischen Denkens

Wie gewinnt man gesicherte Erkenntnisse? Keine einfache, aber doch schon eine sehr alte Frage. Betrachten wir dazu ein Beispiel aus der Bibel [2] (Daniel 1, 8–17), das der Statistiker Mosteller (in [9], S. 881) einmal analysiert hat:

> Daniel war entschlossen, sich nicht mit den Speisen und dem Wein der königlichen Tafel unrein zu machen, und er bat daher den Oberkämmerer darum, sich nicht unrein machen zu müssen. Gott ließ ihn beim Oberkämmerer Wohlwollen und Nachsicht finden. Der Oberkämmerer aber sagte zu Daniel: Ich fürchte mich vor meinem Herrn, dem König, der euch die Speisen und Getränke zugewiesen hat; er könnte finden, dass ihr schlechter ausseht als die anderen jungen Leute eures Alters; dann wäre durch eure Schuld mein Kopf beim König verwirkt. Da sagte Daniel zu dem Mann [...]: Versuch es doch einmal zehn Tage lang mit deinen Knechten! Lass uns nur pflanzliche Nahrung und Wasser zu trinken geben. Dann vergleiche unser Aussehen mit dem der jungen Leute, die von den Speisen des Königs essen.[...] Der Aufseher nahm ihren Vorschlag an und machte mit ihnen eine zehntägige Probe. Am Ende der zehn Tage sahen sie besser und wohlgenährter aus als all die jungen Leute, die von den Speisen des Königs aßen. Da ließ der Aufseher ihre Speisen und auch den Wein, den sie trinken sollten, beiseite und gab ihnen Pflanzenkost.

Der Text ist sehr aufschlussreich. Erkenntnis wird hier durch einen Versuch[1], genauer durch einen Vergleich[2], gewonnen.

Sollte man daraufhin zum Vegetarier werden? Für Daniel und seine Freunde ist das sicher der richtige Weg, aber gilt dies auch allgemein? Da gibt es Zweifel, denn das Ergebnis ist nicht zwingend: Daniel lehnte die nichtvegetarischen Speisen aus religiösen Gründen ab. Er wünschte sich eine vegetarische Ernährung, und es könnte durchaus sein, dass einem diejenigen Speisen gut bekommen, die man gerne essen möchte. Oder seine stärkere Religiosität könnte zur besseren Gesundheit geführt haben. Der Vergleich müsste für eine verallgemeinerungsfähige Aussage also anders gestaltet werden.

[1] Versuch macht klug.
[2] Vergleich macht reich.

Aus heutiger Sicht würde man zunächst eine Gruppe von Probanden suchen, die an einem entsprechenden Versuch teilnehmen würden. Diese Gruppe würde man dann in eine Kontrollgruppe und eine Behandlungsgruppe unterteilen. Die Kontrollgruppe würde wie bisher verpflegt, die Behandlungsgruppe mit vegetarischer Kost. Die Aufteilung darf nicht den Probanden selbst überlassen bleiben, damit sich deren Vorlieben nicht auswirken können. Sie muss also vom Versuchsleiter vorgenommen werden. Dies ist charakteristisch für ein sogenanntes kontrolliertes Experiment. Am besten geschieht die Aufteilung randomisiert, d. h. durch einen Zufallsmechanismus (z. B. Münzwurf), um jegliche Verzerrung soweit als möglich auszuschließen. Man wird weiterhin ausschließen wollen, dass Vorurteile der Beurteiler das Resultat verfälschen, und daher eine Verblindung vornehmen. Die Beurteiler wissen dann also nicht, ob ein Teilnehmer, dessen Zustand sie beurteilen, aus der Kontroll- oder der Behandlungsgruppe stammt. Wünschenswert wäre, dass auch die Versuchsteilnehmer selbst das nicht wissen. Bei Ernährungsversuchen ist das nicht durchführbar, bei Medikamententests kann es aber oft durch Gabe eines Scheinmedikamentes (Placebo) erreicht werden. Gesicherte Erkenntnisse gewinnt man also durch einen Vergleich unter möglichst (bis auf die zu untersuchende Behandlung) identischen Bedingungen.

Nicht immer lässt sich ein kontrolliertes Experiment durchführen. Soll etwa die Gefährlichkeit von Asbest untersucht werden, ist es wenig wahrscheinlich, dass sich Teilnehmer finden, die sich der Behandlungsgruppe zuweisen lassen. Man muss sich dann mit Beobachtungsstudien begnügen, bei denen die Zuordnung zur Behandlungs- oder Kontrollgruppe nicht in der Hand des Versuchsleiters liegt. Dadurch besteht aber die Gefahr, dass die beiden Gruppen sich auch in anderer Hinsicht als nur der zu untersuchenden Behandlung unterscheiden – nämlich zum Beispiel in den Kriterien, die die Aufteilung bewirkt haben. In Beobachtungsstudien können sich also weitere Einflüsse (sogenannte vermengende Faktoren) unauflösbar mit dem Behandlungseinfluss vermischen. Aus diesem Grunde kann man aus Beobachtungsstudien zwar nützliche Hinweise erhalten, aber keine beweiskräftigen Schlüsse ziehen. Auch das obige Bibelbeispiel ist eine Beobachtungsstudie. In diesem Fall hatte sich die Behandlungsgruppe durch Eigenauswahl gebildet, da Daniel und seine Freunde um vegetarische Ernährung gebeten hatten. Man kann aber z. B. nicht sagen, ob sich das bessere Aussehen wirklich aus der vegetarischen Ernährung oder einer stärkeren Beachtung religiöser Vorschriften ergab, die ja auch andere Lebensbereiche betreffen dürfte.

Um Entwicklungen im Zeitablauf zu betrachten, eignen sich besonders Längsschnittstudien, bei denen eine Gruppe im Zeitablauf beobachtet wird. Hat man etwa eine Gruppe von Menschen mit Geburtsjahr 1950 in den Zeitpunkten 1970 und 2010 untersucht, wird man feststellen, dass deren Haare im Zeitablauf vom Alter zwanzig bis zum Alter sechzig grauer geworden sind.

Hat man stattdessen im Rahmen einer Querschnittstudie im Jahre 2010 eine Gruppe von Sechzigjährigen und eine Gruppe von Zwanzigjährigen untersucht, wird man auch feststellen, dass die Haare der Sechzigjährigen grauer sind. Man kann aber nicht sagen, ob Menschen im Laufe ihres Lebens grauere Haare bekommen oder ob vielleicht Menschen im Jahre 1950 mit graueren Haaren zur Welt kamen als im Jahre 1990.

Kommentar: „Denken heißt Vergleichen“ ist ein bekannter Aphorismus von Walther Rathenau ([12], S. 32). Dieser Spruch überrascht zunächst. Denn natürlich muss man beim Vergleichen denken, aber dass beides nahezu identisch ist, erschließt sich nicht sofort. Wenn der Leser aber zum Beispiel darüber nachdenkt, ob er „reich“ ist, dann werden die meisten das mit Blick auf die Weltbevölkerung bejahen, mit Blick auf Fußballstars aber verneinen. Man sieht: Außer in Formalwissenschaften (Logik, Mathematik) ist relevantes Denken ohne Vergleichen in der Tat kaum möglich. Für uns hat das eine aufmunternde Konsequenz: Wenn relevantes Denken und Vergleichen identisch sind, dann ist die „Wissenschaft vom Vergleichen“ (also die Statistik) eben auch die „Wissenschaft vom relevanten Denken“ – ein erhebendes Gefühl!

2.1.2 Grundstruktur statistischer Überlegungen

Etwas abstrakter ist die Grundstruktur aller statistischen Überlegungen in der Abbildung 2.1 beschrieben:

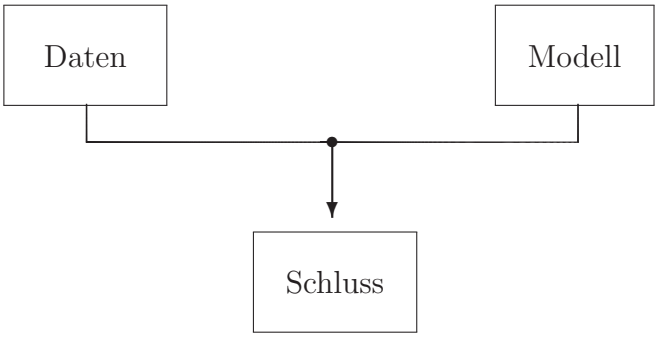

Abbildung 2.1: Grundmodell der Statistik

Im Wesentlichen ist es stets so, dass aus Daten mit Hilfe von Modellen ein Schluss gezogen wird. Man muss sich dabei jederzeit bewusst sein, in welchem Ausmaß Daten und Modell letztlich zur Schlussfolgerung beitragen. Insbesondere die Rolle des Modells wird häufig unterschätzt und muss daher

jeweils genau diskutiert werden. Stellt man etwa durch Marktforschung fest, dass in einem Land niemand Taschenuhren besitzt, kann die Schlussfolgerung sein, dass dies kein Markt für Taschenuhren ist, da niemand welche kauft. Sie kann aber auch ganz im Gegenteil lauten, dass dies ein großartiger Markt für Taschenuhren ist, weil noch niemand eine hat. Trotz gleicher Daten gelangt man also zu völlig verschiedenen Schlüssen. Das Modell ist gewissermaßen der Scheinwerfer, mit dem die Daten beleuchtet werden. Je nach der Farbe des Scheinwerferlichtes erscheinen auch die Daten in der entsprechenden Farbe. Deswegen ist die Wahl des richtigen Modells von ausschlaggebender Bedeutung.

Kommentar: Ein bemerkenswerter Fall unterschiedlicher Modellannahmen ergab sich während des Zweiten Weltkrieges bei Studien zur besseren Panzerung amerikanischer Flugzeuge gegen Flakbeschuss. Der Statistiker A. Wald (vgl. [16]) ließ die Militärs zunächst feststellen, wo sich bei beschossenen Maschinen die Einschussstellen befanden. Anschließend schlug er zur allgemeinen Überraschung vor, diejenigen Stellen stärker zu panzern, an denen man keine Einschüsse festgestellt hatte. Der Grund für die Verwunderung der Militärs waren unterschiedliche Modellannahmen. Die Militärs gingen davon aus, dass die Flugzeuge eine Stichprobe aus den beschossenen Flugzeugen waren und dass daher Flugzeuge an den festgestellten Einschussstellen besonders häufig getroffen wurden. Wald ging dagegen davon aus, dass die Treffer in etwa gleichmäßig über das Flugzeug verteilt waren. Flugzeuge, die an wirklich gefährlichen Stellen getroffen wurden, kehrten aber nicht zurück, so dass an diesen Stellen auch keine Einschüsse festgestellt wurden. Die untersuchten Flugzeuge waren also seiner (realistischeren) Modellvorstellung nach eine Stichprobe von Flugzeugen, die nur an relativ unproblematischen Stellen getroffen worden waren. Je nach Modell markieren also die Daten (Einschussstellen) entweder besonders gefährdete oder relativ ungefährdete Flugzeugteile.

Mit den Daten und ihrer Aufbereitung beschäftigt sich die deskriptive Statistik. Die Ausarbeitung und Untersuchung von Modellen ist die Domäne der Wahrscheinlichkeitstheorie. In der analytischen Statistik werden dann beide Elemente zusammengeführt.

Daten begegnen dem Statistiker zumeist in numerisch kodierter Form. Es ist aber wesentlich, sich über das Messskalenniveau im Klaren zu sein, auf dem sie aufgenommen wurden. Auf einer Nominalskala werden einfach nur Eigenschaften nebeneinander gestellt wie z. B. „männlich“ und „weiblich“. Durch eine Ordinalskala ist hingegen eine Rangordnung vorgegeben, z. B. eine Klassifikation als „schwacher“, „mittlerer“ oder „starker“ Raucher. Die Abstände zwischen den Kategorien müssen dabei nicht unbedingt interpretierbar sein. Sind sie dies, liegt eine Intervallskala vor. Den Unterschied kann man gut an

den Schulnotenskalen im angelsächsischen und im deutschen System veranschaulichen. Nach angelsächsischer Gepflogenheit werden diese mit „A“ bis „F“, nach deutscher dagegen mit „1“ bis „6“ bezeichnet. Im angelsächischen System ist zwar ein „A“ besser als ein „B“ und dieses besser als ein „C“, es ist aber nicht gesagt, dass die Abstände gleich groß sind. Daher kann man die Werte auch nicht mitteln; ein „A“ und ein „C“ sind nicht notwendigerweise gleich gut wie zwei „B“. Im deutschen System wird dies hingegen impliziert, sofern es nicht nur als numerische Codierung für ordinale Daten verstanden wird. Weitere typische Beispiele für Intervallskalen sind Temperaturskalen. Hier ist noch zu beachten, dass Verhältnisaussagen wie diejenige, dass 20 Grad Celsius „doppelt so warm“ wie 10 Grad Celsius wäre, nicht sinnvoll sind, weil dies nach Umrechnung der Werte in Grad Fahrenheit nicht mehr gilt. Falls hingegen ein fester Nullpunkt existiert – wie etwa bei Gewichten oder Geldeinheiten –, so sind auch Verhältnisaussagen sinnvoll: Doppelt so viel in Euro ist eben auch doppelt so viel in Dollar. Man spricht dann von einer Ratioskala. Intervallskalen und Ratioskalen werden auch zusammenfassend als metrische Skalen bezeichnet.

Schließlich werden Merkmale, die auf einer metrischen Skala gemessen werden, auch als quantitativ bezeichnet; auf einer Nominal- oder Ordinalskala gemessene Merkmale heißen auch qualitativ.

2.1.3 Statistik und Information

Information ist das Kernmaterial der Statistik: Stets wird Information aufgenommen, zu relevanter Information verdichtet und in dieser Form weitervermittelt. Es ist daher sinnvoll, der Funktion von Informationsträgern – insbesondere der Darstellung von Information durch Graphiken und Tabellen – gezielte Aufmerksamkeit zu widmen.

Die graphische Darstellung ist in dem Klassiker von E. R. Tufte [14] ausgiebig untersucht. Tufte vertritt einen sehr substanzorientierten Ansatz. Demnach soll Information mit „möglichst wenig Tinte“ und so klar wie möglich präsentiert werden. Ein Beispiel soll diesen Ansatz verdeutlichen.

In der Frankfurter Allgemeinen Sonntagszeitung vom 11. April 2010, Seite 41, findet man einen Artikel von Nadine Oberhuber mit dem Titel „Der Lockruf der Versicherer“, in dem darüber berichtet wird, dass bei Neuabschlüssen von Lebensversicherungen der Anteil von Versicherungen gegen Einmalbeitrag zu Lasten desjenigen mit laufender Beitragszahlung zugenommen hat. Der Artikel enthält auch die in Abbildung 2.2 gezeigte Graphik. Betrachtet man zunächst die linke der drei Teilgraphiken, so ist anzumerken, dass sich die Kurvenwerte jeweils zu 100 % addieren und somit eine der Kurven überflüssig ist. Zudem verbinden die Kurven nur einige wenige diskrete Werte. Denn ins-

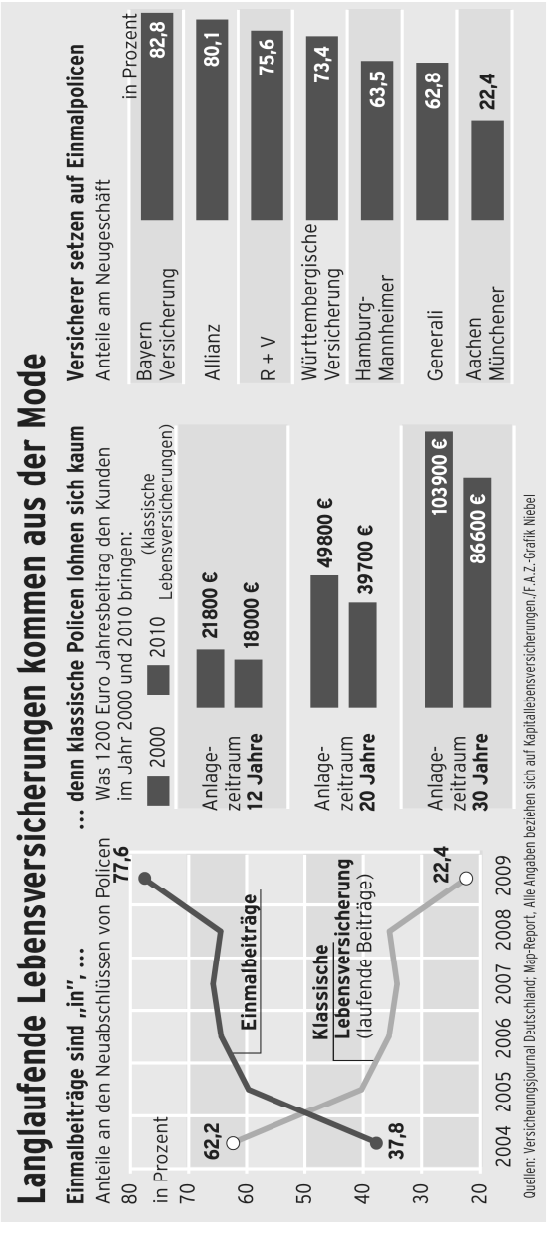

Abbildung 2.2: Beispiel einer Graphik
(Quelle: F.A.Z.–Grafik/Andreas Niebel, in: F.A.S., 11. April 2010, S. 41)

gesamt werden nur zwei numerisch angegebene Werte (37.8 % und 77.6 %) oder – nach nicht ganz einfachem Ablesen – sechs Zahlenwerte dargestellt. Dies ließe sich durch einen Satz („Der Anteil der Einmalbeiträge stieg im Zeitraum von 2004 bis 2009 von 37.8 % auf 77.6 %.“) oder eine Tabelle vermutlich klarer kommunizieren. Die meiste „Tinte“ in der Graphik (sehr ausführliche Achsenbeschriftungen, Farbschattierungen als Hintergrund, zweite Kurve) bezieht sich nicht auf informationstragende Elemente.

Kommentar: Bei der Beurteilung von Graphiken muss man auch bedenken, dass die reine Informationsvermittlung nicht unbedingt das alleinige Ziel sein muss. Gerade im Bereich des Journalismus kann es auch ein wesentliches Ziel sein, Aufmerksamkeit zu erwecken und vielleicht „Farbe“ hinein zu bringen. Dies mag erklären, weshalb manche Graphiken fast „Cartoon-Charakter“ haben.

In der rechten der drei Teilgraphiken ist ebenfalls fraglich, ob die Graphikelemente (Balken) zur Veranschaulichung der Information beitragen. Sie verwirren eher, weil die Länge der Balken nicht zu den dargestellten Werten proportional ist. (Dies ist auch dann nicht der Fall, wenn man einen von 0 verschiedenen Startwert annimmt.) Überdies verwundert, dass der Wert für die AachenMünchener abweichend angegeben ist (durch eine schwarze Zahl hinter dem Balken statt durch eine weiße Zahl im Balken). Möglicherweise ist dies gerade als Hinweis darauf gedacht, dass dieser Wert und sein Balken auf einer anderen Skala dargestellt werden. Solche Brüche in der Datendarstellung führen aber fast immer zu (manchmal beabsichtigter!) Verwirrung und sollten auf jeden Fall vermieden werden. Man kann sich probeweise abwechselnd einmal die Ziffern und einmal die Balken „wegdenken“ und dann beurteilen, welchen Anteil die beiden Elemente an der Informationsvermittlung haben. Den Beitrag der Balken wird man als negativ einstufen. Eine Tabelle der sieben Werte wie etwa Tabelle 2.1 wäre sicherlich ausreichend.

Unternehmen	Prozent
Bayern Versicherung	83
Allianz	80
R + V	76
Württembergische Versicherung	73
Hamburg-Mannheimer	64
Generali	63
AachenMünchener	22

Tabelle 2.1: Anteile der Einmalpolicen am Neugeschäft in Prozent

Bei der Gestaltung einer solchen Tabelle können Techniken helfen, die A. S. C. Ehrenberg in [4] untersucht hat. Ein Kernpunkt seiner Aussagen ist, dass eine Tabelle nicht nur – vielleicht nicht einmal in erster Linie – archivarische Funktion hat. Die Zahlenwerte sollen also nicht akribisch genau, sondern möglichst deutlich vermittelt werden. Dazu ist Runden auf zwei signifikante Stellen fast immer besser als zu große numerische Genauigkeit, weil der Mensch nur mit etwa zwei Ziffern wirklich zu arbeiten vermag, während weitere Stellen eher ablenken. Man sollte also in diesem Beispiel auf die Dezimalstellen verzichten, die ohnehin kaum ins Gewicht fallen. (Wenn nötig, kann die Dokumentation der genauen Zahlen in einem Anhang oder einer Datenbasis erfolgen.) Wie sinnvoll dieser Ratschlag ist, kann man ersehen, wenn man unseren obigen Satz in diesem Sinne neu formuliert: „Der Anteil der Einmalbeiträge stieg im Zeitraum von 2004 bis 2009 von etwa 38 % auf etwa 78 %.“ Das Wesentliche (Differenz und Verhältnis der beiden Zahlen) tritt nun klarer hervor. Das wurde bei der Gestaltung von Tabelle 2.1 schon berücksichtigt. Zugleich wurde in Tabelle 2.1 eine Anordnung der Größe nach (und nicht etwa der nichtssagenden alphabetischen Reihenfolge nach) gewählt. Auf diese Weise wurde eine in der Graphik enthaltene und durchaus hilfreiche Idee auch in der Tabelle genutzt.

Selbstverständlich wird oft eine Graphik einer Tabelle auch in Hinsicht auf Informationsvermittlung weit überlegen sein. Einen interessanten Fall zeigt Abbildung 2.3, die ihrer Form nach eine Graphik, ihrer Struktur nach aber eher eine Tabelle ist. Sie gibt Gelegenheit, eine gerade zur Darstellung von Konzentration im Bereich der Wirtschaft sehr nützliche Graphik vorzustellen: die Lorenzkurve. Für die Daten aus Abbildung 2.3 ist diese in Abbildung 2.4 gegeben. Die farbig markierten Punkte • geben darin in kumulierter Form die Werte aus der Abbildung 2.3 wider. So tragen z. B. die untersten 37.1 % (= 22.7 % + 14.4 %) der Steuerzahler 2.5 % (= 0.1 % + 2.4 %) zur Einkommensteuer bei. Diese kumulierten Werte sind für alle in Abbildung 2.3 angegebenen Daten in Tabelle 2.2 zusammengefasst. Die ganze Kurve entsteht dann durch lineare Interpolation. Die Lorenzkurve zeigt auf einen Blick die gesamte Konzentration und ermöglicht auch leicht Vergleiche zwischen verschiedenen Situationen. Würden alle die gleiche Steuer entrichten, fiele die Lorenzkurve mit der Sehne von (0 %, 0 %) nach (100 %, 100 %) zusammen. Das wäre der Fall vollständiger Gleichheit. Umgekehrt würde im Fall vollständiger Konzentration (nur der Reichste zahlt alle Steuern) die Lorenzkurve beliebig weit in das Achsenkreuz hineingezogen. Die Fläche zwischen der Sehne und der beobachtenen Lorenzkurve ist daher auch ein Maß für die Konzentration. Geeignet normiert – durch Multiplikation mit 2 und manchmal auch dem Faktor $(n/(n-1))$ – wird es als Gini-Koeffizient bezeichnet. Dabei bezeichnet n die Anzahl der Dateneinheiten, die den Prozentwerten zugrunde liegen, falls diese vorliegt. Im Beispiel wäre n die Gesamtzahl der Steuerpflichtigen, die aber nicht angegeben ist.

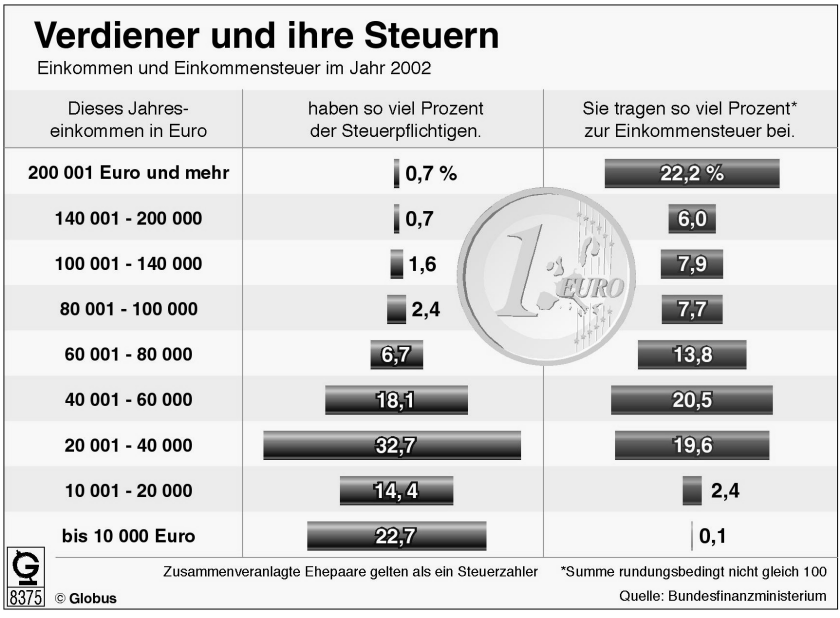

Abbildung 2.3: Einkommen und Einkommensteuer im Jahr 2002 (Quelle: picture alliance/Globus Infografik Bild-Nr. 11653953, gesehen in: Nordbayerischer Kurier, 22./23. März 2003, S. 6)

Einkommen (in tausend EUR)	Anteil an den Steuer-pflichtigen (in %)	Anteil an der Steuer (in %)	kumulierter Anteil an den Steuerpflich-tigen (in %)	kumulierter Anteil an der Steuer (in %)
$]0, 10]$	22.7	0.1	22.7	0.1
$]10, 20]$	14.4	2.4	37.1	2.5
$]20, 40]$	32.7	19.6	69.8	22.1
$]40, 60]$	18.1	20.5	87.9	42.6
$]60, 80]$	6.7	13.8	94.6	56.4
$]80, 100]$	2.4	7.7	97.0	64.1
$]100, 140]$	1.6	7.9	98.6	72.0
$]140, 200]$	0.7	6.0	99.3	78.0
$]200, \infty[$	0.7	22.2	100.0	100.2

Tabelle 2.2: Arbeitstabelle für die Lorenzkurve

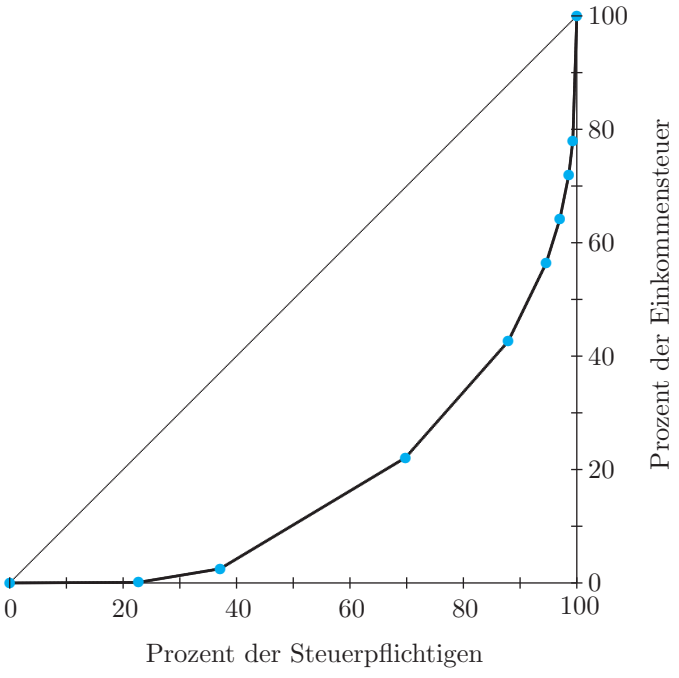

Abbildung 2.4: Lorenzkurve für die Einkommensteuer

2.2 Eindimensionale Daten

2.2.1 Deskriptive Techniken

Alle Daten lassen sich grundsätzlich numerisch kodieren. Umso wichtiger sind daher Techniken, um numerische Daten darzustellen. Wir wollen zusätzlich annehmen, dass die Daten sinnvoll auf einer Ratioskala gemessen wurden und beginnen mit einem einfachen Beispiel. Wir haben das Körpergewicht von zwanzig Personen (in kg) gemessen und dabei die in Tabelle 2.3 angegebenen Werte erhalten.

84, 64, 69, 57, 52, 67, 93, 72, 61, 74,
74, 55, 79, 82, 65, 61, 88, 68, 63, 77.

Tabelle 2.3: Beispieldaten für deskriptive Techniken

Falls wir möglichst wenig Modellvorstellungen benutzen möchten, bietet es sich an, Methoden der explorativen Datenanalyse zu benutzen. Diese Techniken haben die erklärte Absicht, die Daten weitgehend für sich selbst sprechen zu lassen und sie so aufzubereiten, dass man fast gezwungen ist, das

Wichtigste zur Kenntnis zu nehmen. Wir wollen uns auf drei Techniken beschränken und verweisen für mehr Details auf das grundlegende Buch von J. W. Tukey [15].

Zunächst wollen wir die Daten ordentlich aufschreiben. Dazu eignet sich ein Stem-and-Leaf-Display[1] wie in Abbildung 2.5.

```
5 | 257
6 | 11345789
7 | 24479
8 | 248
9 | 3

(5 | 2 bedeutet 52 kg)
```

Abbildung 2.5: Stem-and-Leaf-Display für die Beispieldaten

Die Grundidee ist darin, die Zehnerstelle als „Stamm“ abzutrennen und die Einerstelle als „Blätter“ übersichtlich zu notieren. Natürlich kann man auch andere Stellen als Stamm oder Blatt festsetzen, man muss seine Wahl nur unterhalb der Darstellung genau angeben. Schon ein flüchtiger Blick zeigt die Überlegenheit dieser Darstellung gegenüber einer einfachen Liste. Man hat zwar nach wie vor die vollen Daten, aber zugleich sofort eine Art Histogramm (s. u.), das die Besetzung verschiedener Bereiche andeutet. Besonders deutlich wird das, wenn man die Darstellung um 90 Grad gegen den Uhrzeigersinn dreht, so dass der Stamm zur Abszisse wird. Falls man viele Daten hat, kann man den Stamm auch weiter untergliedern. Ein Beispiel ist Abbildung 2.6.

```
5 *| 2
5 .| 57
6 *| 1134
6 .| 5789
7 *| 244
7 .| 79
8 *| 24
8 .| 8
9 *| 3

(5 *| 2 bedeutet 52 kg)
```

Abbildung 2.6: Erweitertes Stem-and-Leaf-Display für die Beispieldaten

[1] Der Begriff ist in der Statistik eingeführt und soll deshalb unübersetzt bleiben.

Dort stehen jetzt hinter „∗“ die Einerstellen $0, \ldots, 4$ und hinter „.“ die Einerstellen $5, \ldots, 9$.

Für sehr große Datensätze sind Stem-and-Leaf-Displays nicht brauchbar. Man wird dann die Information ohnehin durch summarische Kennzahlen verdichten müssen. Allerdings sind Mittelwerte nicht für jeden Datensatz geeignet, da sie stark von Extremwerten beeinflusst sein können. Für den allgemeinen Fall haben sich daher Quantile sehr bewährt. Grob gesprochen versteht man unter dem $a\,\%$-Quantil den Wert, der den Datensatz so teilt, dass $a\,\%$ der Daten darunter und $(100 - a)\,\%$ der Daten darüber liegen. Das Zahlenbeispiel 1,2,2,3 zeigt aber, dass es gar nicht immer gelingt, einen solchen Wert zu finden. Sucht man etwa ein $50\,\%$-Quantil, so liegt „unter“ der Zahl 2 beispielsweise nur der Wert 1. Unterhalb von 2.000001 (oder jeder anderen Zahl, die größer ist als 2) aber bereits drei der vier Werte. Umgekehrt hätte man im Datensatz 1,2,3,4 sogar ganz viele $50\,\%$-Quantile, nämlich alle Werte zwischen 2 und 3. Um für alle Datensätze jeweils eine eindeutige Lösung sicherzustellen, vereinbaren wir die folgende Definition:

Ein $a\,\%$-Quantil ist *ein* Wert, „unter“ (im Sinne von „$\leq$“) dem mindestens $a\,\%$ der Daten und „über“ (im Sinne von „$\geq$“) dem mindestens $(100 - a)\,\%$ der Daten liegen. Falls es mehrere solche Werte gibt, bilden diese ein Intervall. *Das* $a\,\%$-Quantil ist der Intervallmittelpunkt.

Kommentar: Diese Definition erscheint auf den ersten Blick sehr kompliziert, so dass mancher vielleicht lieber eine „Rechenvorschrift“ hätte. Solche Vorschriften sind möglich, aber letztlich sogar komplizierter, weil viele Fälle unterschieden werden müssen. Die Vorteile der obigen Definition sind offensichtlich:

- Sie stellt sicher, dass das $a\,\%$-Quantil stets existiert und eindeutig ist.
- Sie zeigt immer klar auf, was die Funktion des $a\,\%$-Quantils ist.
- Sie gilt in dieser Form nicht nur für Datensätze, sondern auch für Verteilungen wie die Normalverteilung. Rechenvorschriften leisten dies nicht.

Es lohnt also, die kleine Komplexität der Definition in Kauf zu nehmen. Übrigens gibt es viele Witze über die Beschäftigung der Mathematiker mit den Fragen nach Existenz und Eindeutigkeit. In Wirklichkeit handelt es sich dabei jedoch keineswegs um eine weltfremde theoretische Spielerei. Menschen, die einen Lebenspartner suchen, werden sicher verstehen, dass die Frage nach Existenz und Eindeutigkeit eines solchen ziemlich relevant sein kann. Im gleichen Sinne versteht man den Bedeutungsunterschied der

in der Definition bewusst in Schrägschrift gesetzten Pronomina „*ein*“ und „*das*“ genau richtig, wenn man an den Unterschied zwischen den Sätzen „Eva ist *eine* Freundin von Adam“ und „Eva ist *die* Freundin von Adam“ denkt. Die Definition ist dann gar nicht mehr so kompliziert wie sie zunächst aussieht.

Besonders interessante Quantile sind das untere Quartil (das 25 %-Quantil), der Median (das 50 %-Quantil) und das obere Quartil (das 75 %-Quantil). Im Zusammenspiel mit Minimum (kleinster Wert) und Maximum (größter Wert) eines Datensatzes ergeben diese fünf Zahlen schon einen guten ersten Überblick und werden wie in Abbildung 2.7 notiert.

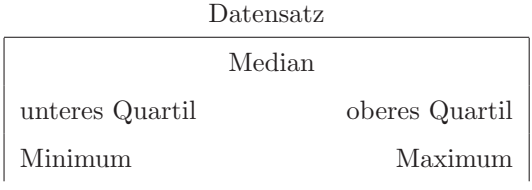

Abbildung 2.7: Aufbau einer 5-Number-Summary

Diese Darstellung heißt 5-Number-Summary[1]. Im konkreten Beispiel ergibt sich die Abbildung 2.8.

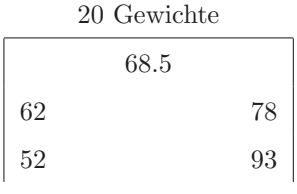

Abbildung 2.8: 5-Number-Summary für die Beispieldaten

Als drittes Konzept aus der explorativen Datenanalyse soll der Boxplot[1] vorgestellt werden. Dabei handelt es sich um eine Art zeichnerische Umsetzung der 5-Number-Summary, die man am besten am konkreten Beispiel versteht. Ein Boxplot wie in Abbildung 2.9 entsteht in drei Schritten:

1. Zunächst zeichnet man einen Kasten[2], der aus dem unteren Quartil, dem Median und dem oberen Quartil besteht.
   Dieser Kasten stellt die „inneren 50 % der Daten“ dar und zeigt u. a. wie symmetrisch der Median in diesem Bereich liegt.

[1] Auch dieser Begriff ist in der Statistik eingeführt und soll deshalb unübersetzt bleiben
[2] engl. box, daher stammt auch der Name.