Über dieses Buch

Dieses Buch bietet – nun in seiner 2., überarbeiteten Auflage – eine Einführung in die Statistik. Abgedeckt werden die Grundlagen der deskriptiven und induktiven Statistik, welche von mir in der Veranstaltung „Wirtschaftsstatistik“ der Bachelor-Studiengänge an der Fakultät Wirtschaftswissenschaft und Wirtschaftsingenieurwesen der HTWK Leipzig gelehrt werden. Im Vordergrund steht dabei die Anwendung statistischer Methoden. Die mathematischen Anforderungen sind auf ein Mindestmaß reduziert. Neben vielen Beispielen zur Anwendung und Verdeutlichung der statistischen Methoden werden zu jedem Kapitel Übungsaufgaben präsentiert. Aufgrund der knappen Zeit, die im Studium für Statistik zur Verfügung steht, ist eine Konzentration auf die essentiellen statistischen Konzepte notwendig. In Kap. 13 werden daher Hinweise für weiterführende Literatur gegeben.

Die Einführung in die Statistik wird in diesem Buch ergänzt mit einer Einführung in `R`. `R` ist eine mittlerweile sehr populäre, leistungsfähige und kostenfreie Statistik-Software. Die Kombination der Einführung in die Statistik mit einem Grundkurs in `R` ermöglicht es, die statistischen Methoden direkt am Rechner anzuwenden und damit besser zu verstehen. Die Einführung in die Statistik-Software `R` ist an die zuvor betrachteten statistischen Methoden gekoppelt und entspricht im Wesentlichen dem Statistik-Teil der Veranstaltung „Quantitative Methoden“ in den Master-Studiengängen an der Fakultät Wirtschaftswissenschaft und Wirtschaftsingenieurwesen der HTWK Leipzig. Jedes Kapitel im Buch ist daher in drei Teile gegliedert: (i) Statistische Methoden, (ii) Übungsaufgaben und (iii) Anwendung mit `R`. Dabei sind die Übungsaufgaben auch ohne Software lösbar.

In Kap. 14 (Anhang) sind Tabellen mit Verteilungen ausgewählter Zufallsvariablen zu finden. Ausführliche Lösungen zu den Übungsaufgaben, die im Buch verwendeten Datensätze und den `R`-Code für Abbildungen und Berechnungen in den Kapiteln gibt es im Internet unter http://bsturm.htwk-leipzig.de/ oder auf Nachfrage per E-Mail (bodo.sturm@htwk-leipzig.de) beim Autor.

Dieses Buch wäre nicht entstanden ohne die großartige und unermüdliche Unterstützung meines Vaters, Martin Sturm, der die Darstellung inhaltlich und sprachlich deutlich verbessert hat. Vielen Dank für Deine Hilfe! An der Fehlerkorrektur und der Aufbereitung der Lösungen zu den Übungsaufgaben haben weiterhin Stefanie Burkhardt, Andrea Gauselmann, Natascha Götze, Mathis Kirchner und Philipp Radomski mitgewirkt. Auch ihnen gilt mein Dank. Alle verbleibenden Fehler gehen natürlich auf mein Konto.

Leipzig, im Dezember 2018

Bodo Sturm

IMAGINE a world without statistics.
Governments would fumble in the dark,
investors would waste money and
electorates would struggle to hold their political leaders to account.
The Economist, 25.2.2012

1 Einführung

Zu Beginn dieser Einführung soll zunächst geklärt werden, was man unter Statistik versteht. Im Anschluss daran werden einige grundlegende Begriffe in der Statistik erläutert. Schließlich wird die Statistik-Software `R` vorgestellt. Mit `R` wird im Rahmen dieses Buchs gearbeitet.

1.1 Was ist Statistik?

Der Begriff *Statistik* beschreibt grundsätzlich zwei unterschiedliche Aspekte. Erstens, in einer umgangssprachlichen Bedeutung, die Zusammenstellung von Daten, die bestimmte Bereiche der menschlichen Zivilisation beschreiben. Dies kann die Entwicklung der Bevölkerung oder des Bruttoinlandsprodukts eines Landes sein, die Umsatzstatistik eines Unternehmens oder das Wählerverhalten bei einer Abstimmung. In diesem Sinne ist auch das Churchill[1] zugeschriebene Zitat "*I only believe in statistics that I doctored by myself*" („Ich glaube nur an Statistiken, die ich selbst gefälscht habe") zu verstehen. Statistiken sind also mit dem Verdacht der Fälschung belastet – dies ist natürlich nicht Gegenstand dieses Buchs! Zweitens, die Gesamtheit der Methoden zur Beschaffung und Auswertung von statistischen Daten. Diese *statistische Methodenlehre* ist eine wissenschaftliche Disziplin und Gegenstand dieser Einführung. Die Auswertung unterteilt man im Allgemeinen in Analyse und Interpretation. Einer wissenschaftlichen Betrachtung sind vor allem die Analysemethoden zugänglich. Sie sind daher auch unser Schwerpunkt. Wesentliche Methoden wie Lagemaße, Streuungs- und Zusammenhangsmaße und lineare Regression werden in diesem Buch vorgestellt. Hierzu zählen auch Methoden zur Analyse von Stichproben wie Konfidenzintervalle und Hypothesentests. Die Antwort auf die Frage „Was ist Statistik?" könnte man also folgendermaßen formulieren: Statistik beschreibt einerseits die Ergebnisse statistischer Arbeit und andererseits das methodische Vorgehen zur Beschaffung und Auswertung von Daten über Massenphänomene. Die Statistik als wissenschaftliche Disziplin stellt für letzteres das notwendige Instrumentarium bereit.

Statistik lässt sich unterscheiden in (i) *deskriptive* oder beschreibende Statistik und (ii) *induktive* oder schließende Statistik. Im Fall der deskriptiven Statistik werden die relevanten Daten der zu untersuchenden Grundgesamtheit gesammelt und ausgewertet. Alle Aussagen beziehen sich nur auf diese Daten: Hochrechnungen oder Verallgemeinerungen auf eine größere Datenmenge sind nicht zulässig. Im Fall der induktiven Statistik werden Daten nur von einem i.d.R. über einen Zufallsprozess ausgewählten Teil der Grundgesamtheit, auch „Stichprobe" genannt, beschafft. Von dieser Stichprobe schließt man mit mathematischen Methoden wie der Wahrscheinlichkeitsrechnung und

1 Winston Churchill (1874-1965), britischer Politiker.

Grenzwertsätzen auf die Grundgesamtheit, z.B. bei Meinungsumfragen oder Materialprüfungen. Man spricht in diesem Fall auch von *Inferenz*.

1.2 Grundlagen

Bevor wir mit der Datenanalyse beginnen, müssen einige statistische Grundbegriffe erläutert werden.

Grundbegriffe der Statistik

Der Begriff *Daten* ist vorwiegend im Plural gebräuchlich. Es handelt sich hierbei um Informationen, die Ausprägungen von Merkmalen betreffen, die von statistischen Einheiten beobachtet werden können. Dabei kann die Ausprägung eines Merkmals quantitativ sein, z.B. bei einer Altersangabe. Statistische Daten können aber auch qualitativer Natur sein, z.B. „Frau“ als mögliche Ausprägung des Merkmals Geschlecht. Eine *statistische Einheit* – auch als *Merkmalsträger* bezeichnet – ist das zu untersuchende Einzelobjekt, welches Gegenstand der statistischen Untersuchung ist, z.B. Unternehmen, Land oder Wahlberechtiger. Eine *Grundgesamtheit* oder statistische Masse ist die Gesamtheit der zu untersuchenden statistischen Einheiten. Wesentlich ist hierbei die Unterscheidung von Bestands- und Bewegungsmassen. Bestandsmassen wie z.B. die Mitarbeiter eines Unternehmens werden zu einem bestimmten Zeitpunkt erfasst. Neueinstellungen oder Entlassungen sind dagegen Bewegungsmassen, die einen Zeitraum betreffen. Bewegungsmassen sind Veränderungen von Bestandsmassen. Eine Grundgesamtheit oder statistische Masse muss mit Hilfe von Kriterien sachlich, räumlich und zeitlich abgegrenzt werden, z.B. die Wahlberechtigten in Deutschland bei der nächsten Bundestagswahl. Als *Merkmal* bezeichnet man eine Eigenschaft einer statistischen Einheit, mit der diese in eine Untersuchung eingeht, z.B. die Parteienpräferenz. *Merkmalsausprägungen* sind dann die verschiedenen Kategorien oder numerischen Werte eines Merkmals. Bei der Frage nach der Parteienpräferenz wären dies z.B. die Kategorien {CDU/CSU, SPD, Bündnis90/Die Grünen, Die Linke, FDP, AfD}.

Wie werden nun statistische Daten erhoben? Grundsätzlich unterscheidet man zwei Möglichkeiten, Merkmalsausprägungen festzustellen: (i) *Totalerhebung* und (ii) *Stichprobe*. Die Erhebung selbst geschieht meist in Form von reiner Beobachtung, Messung oder Befragung. Bei einer Totalerhebung werden die Ausprägungen der interessierenden Merkmale aller statistischen Einheiten einer Grundgesamtheit erhoben. Eine Totalerhebung hat den Vorteil, dass man wirklich „alles“ hinsichtlich der relevanten Merkmale weiß. Oft ist es jedoch viel zu teuer und unpraktisch eine Totalerhebung durchzuführen. Daher wird häufig nur ein Teil der Grundgesamtheit betrachtet. Eine Teilgesamtheit heißt „Stichprobe“, wenn bei der Auswahl der Elemente der *Zufall* wesentlich beteiligt war. Basierend auf den Informationen aus der Stichprobe sollen dann Aussagen über die Merkmale in der Grundgesamtheit getroffen werden. Die Vorteile einer Stichprobenerhebung liegen auf der Hand: Man spart Zeit und Kosten. Der Nachteil der Stichprobenlösung besteht darin, dass man mit unvollständigen Daten arbeiten muss und damit eine gewisse Unsicherheit – auch „Stichprobenfehler“ genannt – verbunden ist.

Allerdings kann diese Unsicherheit bei Anwendung eines klar definierten Zufallsprinzips durch Wahrscheinlichkeitsaussagen beschrieben und kontrolliert werden.

Merkmale →

	ID	Einkommen	Alter	HZB	Geschlecht	Semester	…
Merkmalsträger ↓	1	450	20	1.3	Frau	2	…
	2	700	22	2.0	Frau	1	…
	3	600	24	1.4	Mann	7	…
	…	…	…	…	…	…	…
	…		Merkmalsausprägungen				…
	…	…	…	…	…	…	…
	160	800	25	2.3	Frau	4	…

Tab. 1.1: Datentabelle und Grundbegriffe der Statistik

Beispiel 1.1 Betrachtet sei eine Befragung von 160 zufällig ausgewählten Studenten der HTWK Leipzig im Sommersemester 2012.[2] Angenommen, wir wollen basierend auf dieser Stichprobe Aussagen über alle HTWK-Studenten treffen. Dann umfasst die Grundgesamtheit alle Studenten der HTWK Leipzig im Sommersemester 2012, also etwa 6000 Personen. Die Merkmalsträger in unserer Untersuchung sind offensichtlich die 160 Studenten, die an der Befragung teilgenommen haben. Die bei der Befragung erhobenen Merkmale könnten z.B. sein: Einkommen, Alter, Note der Hochschulzugangsberechtigung (HZB), Geschlecht, Semester usw. Jedem Teilnehmer wird darüber hinaus eine Identifikationsnummer (ID) zugewiesen. Die möglichen Merkmalsausprägungen variieren offensichtlich in Abhängigkeit des erhobenen Merkmals. Die ID hat als Ausprägung die ganzen Zahlen $\{1, \ldots, 160\}$. Das Einkommen wird in €-Beträgen je Monat gemessen, das Alter in Jahren. Die Ausprägungen der HZB sind auf eine Dezimalstelle[3] gerundete Noten. Das Geschlecht hat hier die Ausprägungen {Frau, Mann} und das Semester hat die Ausprägungen {1, …, 10}, wenn man sich auf Angaben innerhalb der Regelstudienzeit für Bachelor und Master beschränkt.

In Tab. 1.1 sind diese Angaben in einer *Datentabelle* zusammengefasst. Unterstellt wird dabei eine Befragung, die alle in Tab. 1.1 enthaltenen Merkmale enthält. Die Zeilen der Datentabelle betreffen die Merkmalsträger, hier die Studenten, von denen jeder eine eigene ID hat. In der Kopfzeile der Spalten stehen die Merkmale. In den Zellen der Datentabelle stehen die jeweiligen Merkmalsausprägungen. Grundsätzlich wäre es auch möglich, einen Teil dieser Angaben aus vorhandenen Unterlagen der Hochschule zu gewinnen. Man spricht dann von einer „Sekundärstatistik“ im Unterschied zur „Primärstatistik“. Ob dieser Weg im vorliegenden Fall zweckmäßig ist, wollen wir hier offen lassen – es geht hier nur darum, auf die Möglichkeit einer sekundärstatistischen Erhebung hinzuweisen, die im Regelfall billiger ist als eine direkte Befragung.

2 In diesem Buch werden Personenbezeichnungen aus Gründen der besseren Lesbarkeit i.d.R. lediglich in der männlichen Form verwendet. Dies schließt das weibliche oder ein drittes Geschlecht mit ein.

3 In diesem Text wird als Dezimaltrennzeichen der Punkt „ . “ und nicht – wie sonst üblich im deutschsprachigen Raum – das Komma „ , “ verwendet. Der Grund hierfür liegt darin, dass die Software R mit dem Punkt als Dezimaltrennzeichen arbeitet.

Unterscheidung von Merkmalen

Merkmale können nach verschiedenen Kriterien unterschieden werden. Eine Unterscheidung, die in der Literatur häufig verwendet wird, ist die in *qualitative* und *quantitative* Merkmale. Ein Merkmal ist qualitativ, wenn die Eigenschaften von Untersuchungseinheiten nur der Beschaffenheit nach variieren. Es kommt also alleine auf den Unterschied zwischen den Ausprägungen an. Beispiele sind Geschlecht, Religionszugehörigkeit und Rechtsform von Unternehmen. Dagegen spricht man von quantitativen Merkmalen, wenn die Eigenschaften der Untersuchungseinheiten sich zahlenmäßig unterscheiden. Die Merkmalsausprägungen sind von vornherein Zahlen, mit oder ohne Maßeinheit. Beispiele sind Alter, Kinderzahl und Einkommen. Hinsichtlich der Art der Informationen über die Ausprägungen von Merkmalen unterscheiden wir in den folgenden Kapiteln daher auch qualitative und quantitative Daten.

Merkmale werden zusätzlich nach ihrer *Skalierung* unterschieden, also danach, welches Niveau der Messbarkeit vorliegt. Die Unterscheidung nach Art der Skalierung ist wichtig, da i.d.R. Auswertungsmethoden in der Statistik nur für bestimmte Merkmalsskalierungen anwendbar sind. Um das richtige Auswertungsinstrument auszuwählen, muss man daher die Skalierung des jeweiligen Merkmals kennen. Unterschieden werden mit aufsteigender Messbarkeit: (i) nominal, (ii) ordinal und (iii) kardinal skalierte Merkmale.

Bei einem *nominal*[4] skalierten Merkmal ist nur die Angabe einer *qualitativen* Verschiedenheit möglich – darüber hinaus liefert die Skalierung in diesem Fall keine Information. Insbesondere ist keine natürliche Ordnung der Merkmalsausprägungen möglich. Beispiele sind Familienstand, Geschlecht und Wirtschaftszweigklassifikation. In Tab. 1.1 sind die Merkmale Geschlecht mit den Ausprägungen {Frau, Mann} und ID mit {1, 2, …, 160} nominal skaliert.

Bei einem *ordinal*[5] skalierten Merkmal ist zusätzlich zur qualitativen Verschiedenheit eine natürliche *Rangordnung* gegeben. Der Informationsgehalt ist daher höher als bei nominal skalierten Merkmalen. Beispiele sind Zeugnisnoten, Zustimmung zu einer Aussage (gemessen z.B. mit {stimme zu, indifferent, stimme nicht zu}) und Produktgüteklassen. In Tab. 1.1 ist das Merkmal HZB ordinal skaliert, da die Noten als Ausprägung in eine natürliche Reihenfolge gebracht werden können (z.B. eine 1 ist besser als eine 2).[6] Beobachtete Ausprägungen nominal und ordinal skalierter Merkmale werden auch als *qualitative* (bzw. *kategoriale*) *Daten* bezeichnet.

Bei einem *kardinal* skalierten Merkmal (auch „metrisch“ skaliert genannt) kann zusätzlich zu den Eigenschaften ordinal skalierter Merkmale eine quantitative, d.h. mengenmäßige, Verschiedenartigkeit festgestellt werden. Darüber hinaus besitzen diese Merkmale immer eine spezielle Maßeinheit. Der Informationsgehalt ist relativ hoch. Beispiele sind

4 Abstammung vom lateinischen „nomen“ = „Name“, d.h., es geht nur um Unterschiedlichkeit.

5 Abstammung vom lateinischen „ordo“ = „Reihe(nfolge), Ordnung“.

6 Das ordinal skalierte Merkmal HZB bzw. Zeugnisnoten ist ein Merkmal mit zahlenmäßigen Ausprägungen. Wir führen es hier mit auf, obwohl es der Form nach quantitativ ist. Das Gleiche gilt für das Merkmal ID.

Einkommen, Alter, Messungen in cm-g-sec und Temperatur. In Tab. 1.1 sind die Merkmale Alter, Einkommen und Semester kardinal skaliert. Differenzen und Verhältnisse können für diese Merkmale gebildet und interpretiert werden. Beobachtete Ausprägungen kardinal skalierter Merkmale werden auch als *quantitative* (bzw. *numerische*) *Daten* bezeichnet.

Kardinal skalierte Merkmale werden wiederum unterteilt in zwei unterschiedliche Skalenarten: (i) Intervallskala und (ii) Verhältnisskala. Merkmale sind intervallskaliert, wenn die Skala keinen natürlichen Nullpunkt hat und dementsprechend nur Differenzen aber keine Verhältnisse sinnvoll zu interpretieren sind. Beispiele sind Temperatur und Jahreszahl.[7] Verhältnisskalierte Merkmale verfügen dagegen über einen natürlichen Nullpunkt, daher sind neben Differenzen auch Quotienten sinnvoll. Beispiele sind Körpergewicht und Einkommen. In Tab. 1.1 sind die kardinal skalierten Merkmale Alter, Einkommen und Semester auf einer Verhältnisskala messbar.

Beispiel 1.2 Die Temperatur wird in Europa üblicherweise in Grad Celsius, ^{o}C, gemessen. In Nordamerika ist dagegen die Angabe der Temperatur in Grad Fahrenheit, ^{o}F, üblich. Die Umrechnung von ^{o}C in ^{o}F erfolgt mit der linearen Transformation $^{o}F = \frac{9}{5}\ ^{o}C + 32$. Daher macht es keinen Sinn, zu sagen „40 ^{o}C sind doppelt so warm wie 20 ^{o}C“. In Grad Fahrenheit wären beide Temperaturen 104 ^{o}F bzw. 68 ^{o}F – die Temperatur wäre nur auf etwa das 1.5fache gestiegen. Die Ursache hierfür ist der fehlende natürliche Nullpunkt für das Merkmal Temperatur – beide Messkonzepte haben einen unterschiedlichen Nullpunkt gewählt. Im Allgemeinen sprechen wir von einer linearen Transformation, wenn zwischen einer Variablen X und einer Variablen Y ein funktionaler Zusammenhang der Form $Y = a + bX$ mit $a, b = const.$ besteht.

Wir haben kardinal skalierte Merkmale als quantitativ und nominal skalierte als qualitativ bezeichnet. Noch etwas mehr zu den ordinal skalierten Merkmalen: Ordinal skalierte Merkmale sind im Allgemeinen qualitativer Natur wie z.B. das Merkmal Zustimmung zu einer Aussage mit den Ausprägungen {stimme zu, indifferent, stimme nicht zu}. Auch Prüfungsnoten mit den Ausprägungen {1.0, …, 5.0} sind streng genommen qualitativ, auch wenn hier zahlenmäßige Ausprägungen vorliegen. So ist die Bildung eines arithmetischen Mittels hier unzulässig, wird doch dabei der Abstand zwischen einer 1 und 2 gleich bewertet wie der zwischen einer 4 und 5. Dennoch ist es allgemeine Praxis, mit Durchschnittsnoten zu arbeiten. In Tab. 1.1 sind nur die Merkmale Geschlecht und HSZ qualitativ, alle anderen Merkmale sind quantitativ.

Eine weitere Unterscheidung ist die in *diskrete* und *stetige* Merkmale. Zur Erläuterung wird Tab. 1.2 mit weiteren Beispielen genutzt. Hier werden sowohl die Skalierung als auch der Merkmalstyp (diskret, stetig) für ausgewählte Merkmale betrachtet.

---

7 Es gibt nur wenige Skalen, die keinen natürlichen Nullpunkt besitzen. Zu den beiden Beispielen: (i) Eine Temperatur von 0 Grad Celsius bedeutet nicht, dass es keine Temperatur gibt, sondern nur eine thermische Molekülbewegung, bei der Wasser gerade taut. (ii) Das Jahr 0 bedeutet nicht die Abwesenheit der Zeitdauer, sondern nur, dass in diesem Jahr eine neue Zählung der Jahre begann.

Merkmals-typ	Skalierung		
	kardinal	ordinal	nominal
diskret	Anzahl Kinder pro Haushalt, Geldbeträge	Evaluation einer Vorlesung mit {sehr gut, gut, befriedigend, ungenügend}	Nationalität, Geschlecht
stetig	Entfernung zw. Wohnung und Arbeitsplatz, Alter	-	-

Tab. 1.2: Merkmalstyp und Skalierung

Ein Merkmal ist *diskret*, wenn es abzählbar viele Ausprägungen annehmen kann. Die Ausprägung ist i.d.R. exakt bestimmbar, es gibt keine Abgrenzungsschwierigkeiten zu anderen Merkmalsausprägungen. Beispiele sind Anzahl Kinder pro Haushalt, Evaluation einer Vorlesung und Nationalität. Auch Geldbeträge (z.B. in € oder €Cent) sind nicht beliebig teilbar und daher diskret. Ein *stetiges* Merkmal kann dagegen jeden beliebigen reellen Wert zumindest in einem bestimmten Intervall annehmen. Die Ausprägungen sind nicht mehr abzählbar, sondern werden durch die Messgenauigkeit bestimmt und sind immer nur Näherungswerte. Beispiele sind Entfernung zwischen Wohnung und Arbeitsplatz, aber auch Alter (gemessen in beliebig kleinen Zeiteinheiten), Alkoholgehalt in Getränken, CO_2-Konzentration, Länge und Gewicht. Die mit der Darstellung stetiger Merkmale verbundene Problematik wird uns später insb. im Kap. 3 noch ausführlich beschäftigen. Hier soll nur angemerkt werden, dass man sich i.d.R. mit einer Bildung von Klassen, einer sog. *Klassierung*, behilft, z.B. sind für das stetige Merkmal „Alter" drei Klassen (unter 18 Jahre, 18 bis unter 65, 65 und darüber) denkbar. Durch die Klassierung wird ein stetiges Merkmal in ein diskretes Merkmal überführt. In der Praxis kommt auch der umgekehrte Vorgang zur Klassierung vor, nämlich die Behandlung eines eigentlich diskreten Merkmals (mit sehr vielen möglichen Ausprägungen) als stetiges Merkmal. Man spricht dann auch von einem quasistetigen Merkmal. Ein Beispiel hierfür sind Geldbeträge. Nominal und ordinal skalierte Merkmale sind dagegen immer diskrete Merkmale.

In Tab 1.1 sind die Merkmale Einkommen, HZB, Geschlecht und Semester diskret. Das Merkmal Alter ist stetig, weil in beliebig kleinen Zeiteinheiten messbar. Wenn das Merkmal Alter jedoch bereits in Jahren erhoben wird, so dass es abzählbar viele Ausprägungen gibt, wäre es als diskret einzustufen.

Die Begriffe *Merkmal* und *Variable* werden häufig synonym verwendet, obwohl sie streng genommen nicht dasselbe bedeuten. Bezeichnen wir mit u eine statistische Einheit, die über $M(u)$ Merkmalsausprägungen verfügt. Eine statistische Variable – oder kurz *Variable* – ordnet dann den Merkmalswerten $M(u)$ reelle Zahlen x zu. Somit ist eine Variable eine Funktion X der statistischen Einheit u, d.h.

$$x = X(u) = Funktion(M(u)) \qquad (1.1)$$

Variablen werden deshalb gerne benutzt, weil man mit Zahlen einfach besser arbeiten kann. Da häufig Merkmalsausprägungen bereits als Zahlen vorliegen, sind in diesem Fall Merkmal und Variable identisch. Zum Beispiel ist das Merkmal Einkommen in Tab. 1.1 bereits identisch mit der Variable Einkommen. Ebenso gilt dies für die Merkmale Alter und HZB. Das Merkmal Geschlecht (mit den Ausprägungen $Frau$ und $Mann$) muss, um

als Variable dargestellt zu werden, umcodiert werden. Hier bietet sich die Zuweisung $Frau = 1$ und $Mann = 0$ an. Die Variable Geschlecht hätte dann die Ausprägungen $\{0,1\}$. Eine solche Umcodierung zu einer binären Variable ist nicht ohne Tücken. Um eine Verwechslung der Ausprägungen auszuschließen, sollte die Variable dann besser mit „Frau“ bezeichnet werden. So wird deutlicher, dass die Variablenausprägung 1 (0) der Merkmalsausprägung $Frau$ ($Mann$) zugeordnet ist.

1.3 Aufgaben

1.1 Merkmale Geben Sie zu den folgenden Merkmalen Beispiele für statistische Einheiten und Merkmalsausprägungen an. Nennen Sie Merkmalsart (qualitativ, quantitativ), Merkmalstyp (diskret, stetig) und Skalierung (nominal, ordinal, kardinal): a) Haarfarbe, b) Verdienst, c) Abiturnote in Mathe, d) Geschlecht, e) Beruf, f) Kontobewegungen in € pro Monat.

1.2 Untersuchungen Betrachten Sie die folgenden statistischen Untersuchungen. Nennen Sie jeweils die statistische Einheit und mögliche Merkmalsausprägungen. Welcher Merkmalstyp und welche Skalierung liegen vor? Haben wir es mit qualitativen oder quantitativen Merkmalen zu tun? Im Fall von quantitativen Merkmalen, in welchen Einheiten wird die Variable gemessen?

a) Eine Studie untersucht die Autos von Mitarbeitern einer großen Firma. Erhoben werden Baujahr, Herstellerland und Typ.
b) Ein Bericht einer Verbraucherschutzorganisation listet 41 Kühlschränke auf mit den Merkmalen Marke, Kosten, Höhe, Breite, Tiefe, Typ, jährliche Energiekosten, Gesamteinschätzung (gut, ausgezeichnet usw.) und Reparaturanfälligkeit der Marke (Anzahl notwendiger Reparaturen je Kühlschrank in den letzten fünf Jahren).
c) Das Umweltbundesamt analysiert den Kraftstoffverbrauch von PKW. Folgende Merkmale werden erhoben: Hersteller, Typ, Gewicht, PS, Verbrauch innerstädtisch, Verbrauch Autobahn.

1.4 Statistik mit R

R ist eine kostenfrei verfügbare Statistik-Software, die sich hervorragend für Lehre und Forschung eignet. R stellt ein umfassendes und flexibles System zur statistischen Analyse und graphischen Darstellung von Daten dar. Eine Einführung in die Statistik kann mit R ohne Programmierkenntnisse mit den angebotenen Funktionen bewältigt werden. Der Vorteil, eine Einführung in die Statistik mit R zu kombinieren, liegt auf der Hand. Studenten können Theorie und Praxis der Statistik sowohl mit „Stift & Zettel“ als auch am Rechner erlernen und anwenden. R kann dann im Laufe des Studiums, in Praktika und nach dem Studium im Unternehmen oder an der Hochschule weiter genutzt werden – ohne Lizenzkosten. Einmal in R geschriebener Programm-Code kann abgespeichert und wiederverwendet werden.

R wird durch das CRAN („Comprehensive R Archive Network“) bereitgestellt. Das CRAN (http://www.r-project.org/) ist ein Server-Netzwerk, welches u.a. den Quellcode, Zusatzpakete (wird – wenn nötig – später erläutert) und Dokumentationen zum Download anbietet.

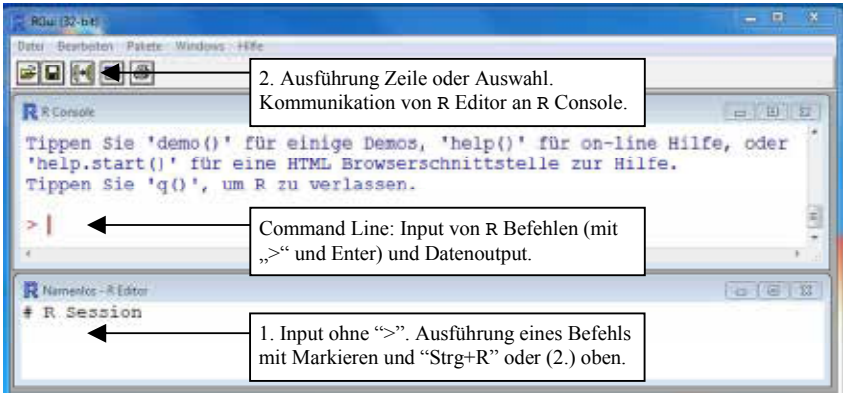

Abb. 1.1: Console und Editor

R wird für verschiedene Betriebssysteme angeboten, z.B. Windows und Mac OS. In dieser Einführung wird R 3.2.2 genutzt. Auf der CRAN Homepage ist der Download von R schrittweise erläutert. Nach dem Speichern und Ausführen der exe-Datei auf dem Rechner legt das Installationsprogramm ein R Symbol zum Start des Programms auf dem Bildschirm ab. R wird gestartet durch Doppelklicken auf dieses Symbol. Es öffnet sich eine einfache Benutzeroberfläche (R GUI = R Graphical User Interface) mit der sog. Console (vgl. Abb. 1.1).

R wird zumeist interaktiv benutzt. Man stellt R eine Frage und R gibt die Antwort. Fragen werden gestellt und beantwortet in der Command Line der Console. Die Eingabe erfolgt hinter dem > Zeichen der Command Line und wird in der Console mit Enter abgeschlossen (vgl. Abb. 1.1). Die Console wird dabei sowohl für die Eingabe von Befehlen und Daten als auch für die Ausgabe genutzt.

Eine gute Gelegenheit, um erste Erfahrungen mit der Statistik-Software zu sammeln, besteht darin, R als Rechner zu benutzen. Betrachten wir die folgenden Beispiele

```
> 2 + 3         # Addition
[1] 5
> 2/3           # Division
[1] 0.6667
```

```
> 2 - 3         # Subtraktion
[1] -1
> 2*3           # Multiplikation
[1] 6
> 2^3           # Exponent, „2 hoch 3", auch = 2*2*2
[1] 8
> 4^2 - 3*2     # Kombination von Operationen
[1] 10
```

Befehle in R enthalten oft Kommentare. Alles hinter dem # Zeichen ist ein Kommentar (wie im Beispiel oben) und wird von der Console ignoriert. Dabei spielt es keine Rolle, ob der Kommentar alleine in einer Zeile oder hinter einem Befehl steht. Kommentare eignen sich gut, um den R Code besser zu strukturieren und zu erläutern.

Das Symbol `[1]` in der Ausgabe des Befehls zeigt die Nummer des Elements in der Ausgabe an. Alle oben gezeigten Ausgaben haben nur ein Element, daher beginnt die Ausgabe mit `[1]`.

Da eine gemeinsame Ein- und Ausgabe oft verwirrend ist, nutzt man üblicherweise einen separaten Editor für die ausschließliche Eingabe von Befehlen und Daten. Im Editor erfolgt die Eingabe ohne das > Zeichen. Der Editor öffnet sich über die Steuerungsleiste der Console mit **Datei => Neues Skript**. Die Nutzung von Console und Editor ist in Abb. 1.1 veranschaulicht. In diesem Buch nutzen wir die Darstellung über die Console, also mit dem > Zeichen, wenn Input und Output *zusammen* dargestellt werden. Wird nur Input dargestellt, erfolgt die einfachere Darstellung im Editor (d.h. ohne das > Zeichen).

Die Ausführung des Befehls im Editor erfolgt entweder mit Markieren und „Strg+R" oder durch Markieren und Mausklicken auf das Symbol „Ausführung Zeile oder Auswahl" in der Kopfzeile der R GUI (vgl. Abb. 1.1). Beispielsweise ist die Rechenoperation $2\frac{1}{(2+4\cdot 2)^2} = \frac{1}{50} = 0.02$ im Editor (ohne das > Zeichen) folgendermaßen darzustellen

```
2*(1/(2 + 4*2)^2)         # sollte 0.02 ergeben
```

Einige wichtige mathematische Funktionen sind

```
> sqrt(2)             # Quadratwurzel
[1] 1.414214
> log(100)            # natürlicher Logarithmus
[1] 4.60517
> exp(1)              # Eulersche Zahl e, „e hoch 1"
[1] 2.718282
> log(2.718282)       # Umkehrrechnung
[1] 1
```

R ist eine objektbasierte Software. Objekte können z.B. Vektoren und Matrizen sein. Um die Erstellung von Objekten zu veranschaulichen, erzeugen wir zunächst einfache Zahlenreihen. Hierfür wird der Befehl `seq()` verwendet, wobei Beginn, Ende und Abstand zwischen den Elementen anzugeben ist.

```
> seq(1, 10, 1)  # sequence, ganze Zahlen von 1 bis 10
 [1]  1  2  3  4  5  6  7  8  9 10
```

Der gleiche Output lässt sich mit `1:10` und dem `c()` Befehl erzeugen, wobei beim letzteren alle Elemente angegeben werden müssen.

```
> 1:10                              # ganze Zahlen von 1 bis 10
 [1]  1  2  3  4  5  6  7  8  9 10
> c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) # kombiniert Zahlen zu Vektor
 [1]  1  2  3  4  5  6  7  8  9 10
```

Zur Definition von Objekten wird der sog. *Zuweisungsoperator* `<-` verwendet. Wir definieren im Folgenden ein Objekt `x` und lassen dieses ausgeben.

```
> x <- seq(1, 6, 0.5); x    # Das „;“ spart den Zeilenumbruch
 [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
```

Hierbei ist zu beachten, dass (i) R case sensitiv ist (es macht also einen Unterschied, ob wir `x` oder `X` schreiben) und (ii) für die Ausgabe eines Objektes der Objektname ausgeführt werden muss. Im Beispiel oben steht der Objektname `x` hinter dem Semikolon `;`, was den Zeilenumbruch spart. Man hätte also auch schreiben können

```
> x <- seq(1, 6, 0.5)         # Definition von x
> x                           # Ausgabe von x
 [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
```

Eine nützliche Eigenschaft von R ist, dass man mit Vektoren rechnen kann. So können z.B. mit den Vektoren `a` (die ganzen Zahlen von 1 bis 8) und `b` (11 bis 18) folgende Berechnungen durchgeführt werden

```
> a <- 1:8; a           # Definition und Ausgabe von a
[1] 1 2 3 4 5 6 7 8
> b <- 11:18; b         # Definition und Ausgabe von b
[1] 11 12 13 14 15 16 17 18
> 1/a                   # Division
[1] 1.0000 0.5000 0.3333 0.2500 0.2000 0.1667 0.1429 0.1250
> a^2                   # Quadrat
[1]  1  4  9 16 25 36 49 64
> c <- a + b; c         # Addition, Definition und Ausgabe von c
[1] 12 14 16 18 20 22 24 26
```

Wir können das eben Gelernte anwenden, wenn wir eine Funktion in R definieren und diese graphisch darstellen wollen (in späteren Kapiteln werden wir ausführlich auf graphische Darstellungen zu sprechen kommen). Angenommen, wir wollen die Funktion $y = f(x) = 3x^2 + 18 - x^3$ graphisch darstellen. Hierzu definieren wir zunächst einen Vektor `x` (die Zahlen von 0 bis 4, Abstand 0.1) und nutzen den Vektor `x`, um einen Vektor `y` zu erstellen. Die graphische Darstellung erfolgt mit `plot()`, der Grundfunktion für zweidimensionale Abbildungen in R. Hier wird nur die Eingabe des Befehls im Editor gezeigt (ohne >).

```
x <- seq(0, 4, 0.1)       # Definiere Vektor x
y <- 3*x^2 + 18 - x^3     # Definiere y
plot(x, y)                # Abbildung der Funktion
```

Abb. 1.2 zeigt die graphische Darstellung der Funktion $y = f(x) = 3x^2 + 18 - x^3$ für X in [0,4].

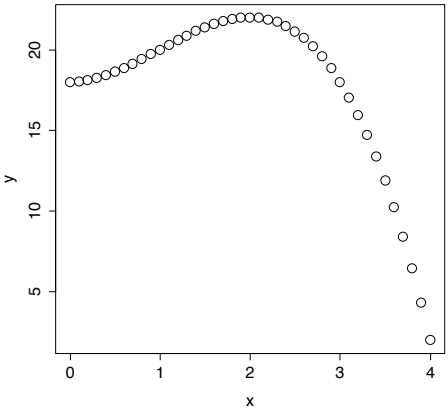

Abb. 1.2: Beispiel für graphische Darstellung einer Funktion

Der Zugriff auf Elemente in einem Vektor ist in R mit einer eckigen Klammer [] möglich. So kann z.B. das 4. Element von a folgendermaßen aufgerufen und geändert werden

```
> a[4]           # Aufruf des 4. Elements von a
[1] 4
> a[4] <- 77; a  # Überschreiben des 4. Elements mit 77
[1]  1  2  3 77  5  6  7  8
```

Daten werden in R zumeist in Tabellen gespeichert und bearbeitet. Man kann eine solche Datentabelle leicht mit dem Befehl data.frame() erzeugen. Im Folgenden wird die Tabelle d erzeugt. Diese beinhaltet als Merkmale die Buchstaben A bis E, die Zahl x=1 und die Variable Nr mit den Zahlen von 1 bis 5. Mit data.frame() werden diese Vektoren zu der Datentabelle d zusammengeführt.

```
> L5 <- LETTERS[1:5]; L5  # die ersten 5 Buchstaben im Alphabet
[1] "A" "B" "C" "D" "E"
> d <- data.frame(x = 1, Nr = 1:5, L = L5); d # Tabelle d
  x Nr L
1 1  1 A
2 1  2 B
3 1  3 C
4 1  4 D
5 1  5 E
```

Auf die Elemente der Tabelle d, die ja eine Matrix ist, kann wieder mit einer eckigen Klammer [] zugegriffen werden. So können einzelne Zeilen, Spalten oder auch Zellen angezeigt und ggf. bearbeitet werden.

```
> d[2, ] # Ausgabe 2. Zeile von d
  x Nr L
2 1  2 B
> d[, 3] # Ausgabe 3. Spalte von d
[1] A B C D E
Levels: A B C D E
> d[2,3] # Ausgabe 2. Zeile und 3. Spalte von d
[1] B
Levels: A B C D E
```

Das Merkmal `L` im Beispiel oben ist nominal skaliert. `R` gibt bei der Ausgabe solcher Merkmale immer die möglichen Ausprägungen (`Levels` genannt) an.

Zum Einlesen von Datensätzen in `R` gibt es verschiedene Möglichkeiten. Die wohl einfachste Option besteht darin, sog. csv-Dateien (csv = comma separated values) einzulesen. Betrachten wir als Beispiel die Datei `Bsp._1.1.csv`. In diesem Datensatz sind die Daten der ersten sechs Teilnehmer der Befragung in Beispiel 1.1 abgespeichert. Das Einlesen des Datensatzes erfolgt folgendermaßen: (i) Unter **Datei => Verzeichnis wechseln …** muss in der `R` GUI das *Arbeitsverzeichnis* angegeben werden, in dem die Datei liegt. `R` muss genau wissen, wo auf dem Rechner der Datensatz liegt. Alternativ kann das Arbeitsverzeichnis mit `setwd()` gesetzt werden. (ii) Mit dem Befehl `read.csv()` kann dann die Datei eingelesen werden. Dabei muss der vollständige Dateiname in Anführungszeichen gesetzt werden. Es ist sinnvoll, der eingelesenen Datei einen Objektnamen in `R` zuzuweisen (hier `data`). Auf dieses Objekt kann dann später in `R` zugegriffen werden. Die Objektbezeichnung sollte dabei anders lauten als die Variablenbezeichnungen im Objekt. Die Zuweisung erfolgt mit `<-`, dem Zuweisungsoperator. (iii) Das Objekt `data` kann nun in `R` angezeigt werden. Hierzu muss einfach der Objektname ausgeführt werden. (iv) Mit `names(data)` lassen sich die Variablennamen anzeigen. (v) Mit `attach(data)` wird das Objekt `data` zum aktuellen Objekt erklärt. Man sagt auch, dass ein Objekt an den Suchpfad in `R` „gebunden“ wird. Damit kann direkt mit Auswertungsfunktionen auf die Variablen zugegriffen werden, z.B. mit `mean(Alter)` das arithmetische Mittel der Variable Alter bestimmt werden.

```
> data <- read.csv("Bsp._1.1.csv") # Einlesen aus Arbeitsverz.
> data                             # Zeige Objekt
  ID Einkommen Alter  HZB Geschlecht Semester
1  1       450    20  1.3       Frau        2
2  2       700    22  2.0       Frau        1
3  3       600    24  1.4       Mann        7
4  4       500    19  2.0       Frau        1
5  5       650    25  2.7       Mann        8
6  6       800    19  1.0       Frau        1

> names(data)     # Zeige Variablennamen
[1] "ID"     "Einkommen"     "Alter"     "HZB"     "Geschlecht"
"Semester"
> attach(data)    # Definiere data als aktuelles Objekt
> mean(Alter)     # Berechne den Mittelwert von Alter
[1] 21.5
```

Der Befehl `attach()` ist nicht ohne Risiken. Da man immer nur ein Objekt an den Suchpfad binden kann, ist es möglich, dass – wenn man mehrere Objekte mit gleichen Variablenbezeichnungen eingelesen hat – Verwirrung darüber entsteht, welches Objekt denn nun das „aktuelle“ ist. Dieses Problem lässt sich vermeiden, wenn man auf `attach()` verzichtet und die jeweilige Variable zusammen mit dem Objekt direkt anspricht. Hierzu müssen Objekt und Variable mit dem `$`-Zeichen verknüpft werden, z.B.

```
> data$Alter       # direkter Zugriff auf Alter ohne attach()
[1] 20 22 24 19 25 19
> mean(data$Alter) # direkter Zugriff mit einer Funktion
[1] 21.5
```

Im Folgenden sollen einige sehr hilfreiche Befehle kurz erläutert werden. Mit `getwd()` gibt `R` das aktuelle Arbeitsverzeichnis, das *working directory*, aus. Sollte man also unsicher sein, in welchem Ordner `R` nach einem Datensatz sucht, kann mit diesem Befehl diese Unsicherheit schnell behoben werden. Wenn man das Arbeitsverzeichnis wechseln möchte (hier etwa auf das Verzeichnis `neu`) ist das mit `setwd()` möglich.

```
> getwd() # Wo ist das aktuelle Arbeitsverzeichnis?
[1] "C:/Users/Sturm/01_HTWK 31. Juli 2013"
> setwd("C:/Users/Sturm/neu")    # definiere Arbeitsverzeichnis
```

Mit `options(digits=)` kann die Anzahl der ausgegebenen Stellen begrenzt werden.

```
> options(digits = 4) # Beschränke die Ausgabe auf 4 Stellen
> exp(1)              # Eulersche Zahl, „e hoch 1“
[1] 2.718
```

Mit `options()` kann auch die Standardeinstellung bei der Ausgabe großer Zahlen (Darstellung über Exponent zur Basis 10) in eine herkömmliche Dezimaldarstellung verändert werden.

```
> 1000^2  # Standard = „1 x 10 hoch 6“
[1] 1e+06
> options(scipen="10") # options(scipen="0") ist der Standard
> 1000^2
[1] 1000000
```

Kommen wir noch einmal zum Befehl `attach()` zurück: Wenn man ein Objekt mit diesem Befehl an den Suchpfad gebunden hat, kann man diese Bindung mit `detach()` wieder rückgängig machen. Das ist dann sinnvoll, wenn man mehrere Objekte hat, die man an den Suchpfad binden möchte. Ist man unsicher, welches Objekt das „aktuelle“ ist, welches also als letztes an den Suchpfad gebunden wurde, hilft `search()`. Dieser Befehl gibt einen Vektor mit geladenen Objekten und Paketen aus. Das Objekt, welches an 2. Stelle der Ausgabe nach `search()` steht (hinter `".GlobalEnv"`) ist das „aktuelle“ Objekt auf dem Suchpfad. Zur Erläuterung ein kleines Beispiel (**…** ist gelöschter Output aus der Console).

```
> data <- read.csv("Bsp._1.1.csv") # Einlesen, neue Session
> search()      # Objekt data ist noch nicht attached
[1] ".GlobalEnv"        "package:stats"     …
```

```
> attach(data)  # Objekt data wird attached
> search()
 [1] ".GlobalEnv"      "data"            "package:stats"…

> detach(data)  # data wird detached
> search()      # Objekt data ist nicht mehr attached
[1] ".GlobalEnv"      "package:stats"     …
```

Die Hilfe in R lässt sich mit `?Befehl` aufrufen. Man muss also den Befehl kennen, um die Hilfe zu starten. Beispielsweise wird die Hilfe für `seq()` und `mean()` aufgerufen mit

```
?seq   # Hilfe zum Befehl seq()
?mean  # Hilfe zum Befehl mean()
```

Abschn. 15.2 listet die wichtigsten Befehle, die in diesem Buch verwendet werden, auf.

Man verlässt R, in dem man den Editor abspeichert: **Datei => Speichern in Datei …** im ausgewählten Verzeichnis. Es empfiehlt sich, die Editor-Datei als txt-Datei abzuspeichern, also zum Beispiel als `editor_1.txt`. Die abgespeicherte Editor-Datei kann dann später wieder im Editor aufgerufen und verwendet werden (vgl. unten).

Beim Verlassen von R wird man vom Programm gefragt, ob man den Workspace sichern möchte. Der Workspace beinhaltet quasi das Gedächtnis der aktuellen Session, also u.a. definierte Objekte und geladene Pakete. In dieser Einführung gehen wir davon aus, dass wir den Workspace nicht speichern und alle relevanten Informationen im Editor enthalten sind. Auf die Frage „Workspace sichern?“ also mit „Nein“ antworten.

Nach dem Starten von R kann die R GUI genutzt werden, um eine existierende Editor-Datei zu öffnen: **Datei => Öffne Skript ….** Die Datei im relevanten Verzeichnis (dabei Dateityp **Alle Dateien (*,*)** wählen) wird ausgewählt und geöffnet. Die existierenden Befehle können dann wieder ausgeführt werden. Bei einer neuen Session muss allerdings das Arbeitsverzeichnis ggf. neu angegeben werden (vgl. zum Ändern des Arbeitsverzeichnisses und zum Befehl `getwd()` oben).