utb.

Andreas Behr

Grundwissen Deskriptive Statistik

mit Aufgaben, Klausuren und Lösungen

3. Auflage

utb 4825

Eine Arbeitsgemeinschaft der Verlage

Brill | Schöningh – Fink · Paderborn
Brill | Vandenhoeck & Ruprecht · Göttingen – Böhlau · Wien · Köln
Verlag Barbara Budrich · Opladen · Toronto
facultas · Wien
Haupt Verlag · Bern
Verlag Julius Klinkhardt · Bad Heilbrunn
Mohr Siebeck · Tübingen
Narr Francke Attempto Verlag – expert verlag · Tübingen
Psychiatrie Verlag · Köln
Ernst Reinhardt Verlag · München
transcript Verlag · Bielefeld
Verlag Eugen Ulmer · Stuttgart
UVK Verlag · München
Waxmann · Münster · New York
wbv Publikation · Bielefeld
Wochenschau Verlag · Frankfurt am Main

Prof. Dr. Andreas Behr lehrt Statistik an der Universität Duisburg-Essen.

Andreas Behr

Grundwissen Deskriptive Statistik

mit Aufgaben, Klausuren und Lösungen

3., überarbeitete und erweiterte Auflage

UVK Verlag · München

Umschlagabbildung: © megakunstfoto · iStockphoto
Autorenbild: © privat

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

3., überarbeitete und erweiterte Auflage 2023
2., überarbeitete Auflage 2019
1. Auflage 2017

DOI: https://doi.org/10.36198/9783838561752

– ein Unternehmen der Narr Francke Attempto Verlag GmbH + Co. KG
Dischingerweg 5 · D-72070 Tübingen

Internet: www.narr.de
eMail: info@narr.de

Einbandgestaltung: siegel konzeption | gestaltung
CPI books GmbH, Leck

utb-Nr. 4825
ISBN 978-3-8252-6175-7 (Print)
ISBN 978-3-8385-6175-2 (ePDF)

Vorwort zur dritten Auflage

Für die dritte Auflage wurden im Text lediglich kleinere Korrekturen vorgenommen. Im Anhang wurden für die Überprüfung des Lernstands und die Klausurvorbereitung vier Übungsklausuren mit Lösungshinweisen ergänzt.

Andreas Behr August 2023

Digitale Zusatzmaterialien
Die im Text verwendeten Daten können Sie unter *www.utb.de* auf Titelebene des Buches bei *Bonus-Material* herunterladen.

Vorwort zur zweiten Auflage

Für die zweite Auflage wurden alle beispielhaften empirischen Analysen mit Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 2018 (ALLBUS) angefertigt. (GESIS - Leibniz-Institut für Sozialwissenschaften (2019): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2018. GESIS Datenarchiv, Köln. ZA5270 Datenfile Version 2.0.0, doi:10.4232/1.13250.) Die empirischen Aufgaben und deren Lösungshinweise beziehen sich nun ebenfalls durchgängig auf Daten das ALLBUS. Das verwendete Datenfile steht für die Leserinnen und Leser des Buches zum Download bereit. Zudem wurden im Text, in den Übungsaufgaben und den Lösungshinweisen in der ersten Auflage verbliebene Fehler und Ungenauigkeiten für die zweite Auflage korrigiert. Alle Ergebnisse wurden in R berechnet. Bei der Angabe von Zwischenergebnissen im Text ist zu beachten, dass durchgängig mit nicht gerundeten Zwischenergebnissen gerechnet wurde. Hierdurch ergeben sich mitunter geringfügige Abweichungen der dargestellten Ergebnisse von den Ergebnissen, die bei Verwendung gerundeter Zwischenergebnisse resultieren.

Auch bei der zweiten Auflage gilt mein besonderer Dank Christoph Schiwy für seine Unterstützung bei der Erstellung des Buches mit LaTeXund knitr. Für die Durchsicht des Manuskriptes möchte ich mich bei Gerald Fugger, Marco Giese, Donald Teguim Kamdjou, Fiona Ewald, Lucy Hong und Erik Berns bedanken.

Andreas Behr August 2019

Vorwort zur ersten Auflage

Der vorliegende Text soll Einblicke in die Grundlagen der Deskriptiven Statistik vermitteln. Er ist entstanden auf der Grundlage von Vorlesungsfolien und Skripten meiner Lehrveranstaltungen an den Universitäten in Frankfurt/M., Münster und Essen. Als didaktisches Konzept wurde versucht, die vorgestellten Methoden mit Hilfe einfachster Zahlenbeispiele transparent darzustellen, bevor sie auf einen Datensatz, der Informationen über 1000 Personen enthält und aus der Panel Study of Income Dynamics (USA) stammt, angewendet werden.

Der Text enthält neben der Darstellung der ausgewählten statistischen Methoden jeweils am Kapitelende kurze Blöcke, in denen Code zur Berechnung der numerischen Ergebnisse und zur Erstellung der Graphiken der statistischen Programmierumgebung R präsentiert wird. Die dargestellten und besprochenen Ergebnisse lassen sich damit recht einfach reproduzieren. Ein einführender Text in die statistische Analyse mit R ist Behr, Andreas / Pötter, Ulrich, Einführung in die Statistik mit R, 2. Auflage, Vahlen Verlag, München, 2011.

Aus Platzgründen wurde in der Regel ein etwas vereinfachter R-Code angegeben, so dass die im Text enthaltenen Graphiken nicht mit den aus dem angegeben R-Code resultierenden identisch sind. Zu beachten ist, dass die dargestellten Ergebnisse gerundet wurden, wodurch sich u.U. geringfügige Abweichungen von exakten oder weniger stark gerundeten Ergebnissen – etwa bei Verwendung des angegebenen R-Codes – erklären. In Anlehnung an die übliche Darstellung in statistischer Software wird im gesamten Text als 1000er Trennzeichen ein Komma und als Dezimaltrennzeichen ein Punkt verwendet.

Am Ende jedes Kapitels befinden sich Übungsaufgaben, mit deren Hilfe die in dem jeweiligen Kapitel besprochenen Inhalte vertieft und deren Anwendung geübt werden kann. Am Ende des Buches finden sich gekürzte Lösungen der Übungsaufgaben. Zudem enthält das Buch eine Formelsammlung, in der die wichtigsten Formeln des Textes zusammengestellt sind. Üblich ist die Bereitstellung derartiger Formelsammlungen als Hilfe in Klausuren. Formeln, die in der Formelsammlung enthalten sind, sind im Text grau hinterlegt, womit auf deren herausgehobene Bedeutung verwiesen wird.

Für die eigenständige Überprüfung des Kenntnisstands sind zudem zwei Klausuren im Text enthalten. Auch für diese finden sich am Ende des Buches kurze Lösungshinweise.

Bedanken möchte ich mich bei Götz Rohwer für Hinweise und Beiträge, insbesondere zu den Kapiteln 2 und 11; und bei Christoph Schiwy, ohne dessen Unterstützung in LaTeXund knitr das Buch nicht entstanden wäre. Zudem danke ich Katja Theune, Lucy Hong, Neele Daun, Jurij Weinblat, Gerald Fugger und Kevin Gründker für die Durchsicht des Manuskripts.

Andreas Behr 2017

Inhaltsverzeichnis

1
Einführung

Mit Hilfe der Methoden der Deskriptiven Statistik sollen Daten, die für eine Anzahl an Einheiten (Personen, Unternehmen, etc.) gewonnen wurden, so dargestellt und beschrieben werden, dass ihr Informationsgehalt einfach und anschaulich sichtbar wird. Ausgangspunkt sind Werte einer oder mehrerer statistischer Variablen, mit denen Eigenschaften der Einheiten erfasst sind. In diesem einleitenden Kapitel erläutern wir dies Ziel, geben einige grundlegende Definitionen an sowie ein Beispiel, das auch in späteren Kapiteln verwendet wird.

1.1 Einleitung

In diesem Buch beschäftigen wir uns mit der deskriptiven (beschreibenden) Statistik. Vordringlich geht es darum, Methoden zu besprechen, mit denen vorliegende Daten anschaulich dargestellt und wesentliche Charakteristika der Verteilung der Daten herausgearbeitet werden können. Methoden der Wahrscheinlichkeitsrechnung und der Inferenzstatistik werden in einem anderen Buch (Grundwissen: Induktive Statistik) dargestellt.

1.1.1 Ziele

Das **Ziel** besteht darin, Einblicke in die Methoden und die Probleme der statistischen Begriffsbildung, der Datengewinnung und der Datenauswertung zu geben. Obwohl Fragen der Operationalisierung in der empirischen Wirtschaftsforschung von ganz zentraler Bedeutung sind, wird im Rahmen dieser Einführung nur in begrenztem Umfang darauf eingegangen und der Schwerpunkt auf die statistische Auswertung von Daten gelegt. Fragen der Operationalisierung müssen in der Praxis jeweils gesondert für das aktuelle Forschungsprojekt behandelt werden und sind nur eingeschränkt einer allgemeinen Behandlung zugänglich. Ein Grundwissen über statistische Methoden der Datenanalyse in Form von tabellarischen und grafischen Darstellungen und der Charakterisierung durch Kennzahlen sollte jedoch jeder Wirtschafts- und Sozialwissenschaftler besitzen.

1.1.2 Motivation

Die Statistik kann zwar einerseits als eine Hilfswissenschaft für die Wirtschaftswissenschaften verstanden werden, sie hat jedoch andererseits eine zentrale Funktion. Die meisten Phänomene, die in den Wirtschaftswissenschaften interessieren, sind einer unmittelbaren Beobachtung oder Erfahrung nicht zugänglich. Erst durch eine adäquate Begriffsbildung und Datenerhebung werden diese Phänomene empirisch zugänglich. Zu denken ist hier z. B. an das Niveau der Arbeitslosigkeit, die allgemeine Entwicklung von Verbraucherpreisen, die Mietpreisentwicklung und dergleichen mehr. In diesem Sinne kann die Statistik als ein „Sinnesorgan“ der Wirtschaftswissenschaften verstanden werden.

Als eine weitere Motivation lässt sich die zunehmende Datenverfügbarkeit und damit einhergehend die zunehmende Bedeutung von Datenanalysen anführen. Die Fähigkeit, Ergebnisse von Datenanalysen verstehen und interpretieren und die dabei verwendeten Methoden kritisch hinterfragen zu können, ist sicherlich von herausragender Bedeutung.

1.2 Variablen und Häufigkeiten

In diesem Abschnitt erläutern wir einige Begriffe, die für alle weiteren Kapitel von grundlegender Bedeutung sind.

1.2.1 Variablen und Daten

Deskriptive Statistik beginnt mit Daten. Diese Daten sind fast immer in der Form einer Datenmatrix gegeben, deren Schema folgendermaßen verdeutlicht werden kann:

i	x_i	y_i	z_i
1	x_1	y_1	z_1
2	x_2	y_2	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	z_n

Jede Zeile bezieht sich auf eine Einheit (z. B. eine Person oder ein Unternehmen). Die erste Spalte enthält eine Nummer, die die jeweilige Einheit angibt. Die Anzahl der Einheiten wird durch die Zahl n angegeben.

Die weiteren Spalten enthalten die Werte von Variablen. Im obigen Schema gibt es drei Variablen: X, Y und Z. Dies ist eine allgemeine Konvention: Variablen werden durch kursive Großbuchstaben bezeichnet, ihre Werte durch entsprechende Kleinbuchstaben. So ist x_i der Wert, den die Variable X bei der Einheit i annimmt; und entsprechend sind y_i und z_i zu verstehen. Diese Werte sind die eigentlichen Daten, aber wir betrachten sie nicht isoliert, sondern als Werte von Variablen, die für die jeweilige Gesamtheit der Einheiten definiert sind.

Dementsprechend kann der Begriff ‚**Variable**‘ in zwei Bedeutungen verwendet werden. Einerseits bezieht er sich auf die Spalten einer Datenmatrix; bei einer formalen Betrachtung handelt es sich dann um Spaltenvektoren. Andererseits kann man mit dem Begriff eine Abbildung bezeichnen, die jeder Einheit einen bestimmten Wert in einem **Merkmalsraum** zuordnet, d.h. in einer Menge möglicher Merkmalsausprägungen.

1.2.2 Merkmalsarten und Skalenniveaus

In der Statistik ist es allgemein üblich, Merkmalswerte durch Zahlen zu repräsentieren (so dass man mit ihnen rechnen kann). Natürlich muss ihre Bedeutung angegeben werden, z. B. dass es sich um Monatslöhne in Euro handelt.

Merkmale haben ein bestimmtes **Skalenniveau**. Bei **nominalen Merkmalen** kann lediglich die Unterschiedlichkeit festgestellt werden, aber verschiedene Ausprägungen können nicht sinnvoll angeordnet werden und Abstände zwischen den Ausprägungen haben keine bestimmte Bedeutung. Nominale Merkmale sind z. B. das Geschlecht oder der Beruf.

Bei einem **ordinalen Merkmal** lassen sich die verschiedenen Ausprägungen in eine sinnvoll interpretierbare Reihenfolge bringen, jedoch haben auch in diesem Fall die Abstände keine bestimmte Bedeutung. Insbesondere bei subjektiven intensitätsmäßigen Auskünften findet oft die Ordinalskala Anwendung, etwa bei Wertungen wie ‚gut‘, ‚mittel‘, ‚schlecht‘ o.ä.

Ein Merkmal ist **kardinal skalierbar**, oft auch **metrisches Merkmal** genannt, wenn die verschiedenen Ausprägungen unterscheidbar sind, in eine Rangfolge gebracht werden können und die Abstände eine bestimmte Bedeutung haben. Bei einer Intervallskala existiert kein absoluter Nullpunkt, so dass zwar Abstände aber nicht sinnvoll Verhältnisse interpretiert werden können, wie etwa bei der Temperatur. Bei einer Verhältnisskala existiert ein absoluter Nullpunkt, etwa bei Gewichten oder Längenangaben.

1.2.3 Absolute und relative Häufigkeiten

Mit den Methoden der deskriptiven Statistik interessiert man sich nicht für die Merkmalswerte bestimmter (identifizierbarer) Einheiten, sondern nur dafür, mit welchen Häufigkeiten Merkmalswerte

in der jeweiligen Gesamtheit der n Einheiten (oder in Teilgesamtheiten) auftreten. Man unterscheidet absolute und relative Häufigkeiten. Die absolute Häufigkeit, mit der eine Variable einen Wert x annimmt, ist die Anzahl der Einheiten, die diesen Merkmalswert aufweisen. Die relative Häufigkeit ist der entsprechende Anteil, also die absolute Häufigkeit geteilt durch n.

Wenn einfach von Häufigkeiten gesprochen wird, sind in diesem Buch stets relative Häufigkeiten gemeint. Als grundlegende Notation verwenden wir $\mathrm{P}(X = x)$, womit die Häufigkeit gemeint ist, mit der die Variable X den Wert x annimmt. Ganz analog bedeutet $\mathrm{P}(X = x, Y = y)$ die Häufigkeit, mit der X den Wert x *und* Y den Wert y annimmt.

Zur Illustration betrachten wir eine Gesamtheit von $n = 8$ Einheiten. Für die Variable X gibt es folgende Merkmalswerte (z. B. Altersjahre): $x_1 = 1$, $x_2 = 2$, $x_3 = 2$, $x_4 = 4$, $x_5 = 4$, $x_6 = 4$, $x_7 = 7$ und $x_8 = 16$. Dann kann man bespielsweise folgende Häufigkeiten ermitteln:

$$\mathrm{P}(X = 1) = 1/8, \quad \mathrm{P}(X = 4) = 3/8, \quad \mathrm{P}(X = 9) = 0.$$

Offenbar kann man auch x-Werte verwenden, die bei den Einheiten nicht vorkommen; dann ist die Häufigkeit Null.

Schließlich verwenden wir auch manchmal eine Notation, die sich auf mehrere mögliche Merkmalswerte bezieht: $\mathrm{P}(X \in A)$, womit die Häufigkeit dafür gemeint ist, dass X irgendeinen Wert in der Menge A annimmt. Beispielsweise findet man mit den eben angegebenen Werten, dass $\mathrm{P}(X \in \{1, 4\}) = 1/2$ ist.

Beziehen wir uns auf die Elemente eines explizit definierten Merkmalsraums, bezeichnen wir diese mit $\tilde{x}_j$ $(j = 1, \ldots, J)$ und ihre Häufigkeiten mit $f_j = \mathrm{P}(X = \tilde{x}_j)$. Mit $n_j = f_j n$ bezeichnen wir die absolute Häufigkeit.

1.2.4 Stabdiagramme

Durch die Häufigkeiten $\mathrm{P}(X = x)$ wird die Verteilung der Variablen X beschrieben. Viele Methoden der deskriptiven Statistik haben das Ziel, anschauliche und informative Bilder solcher Verteilungen zu liefern. Wenn es nicht zu viele unterschiedliche Merkmalswerte gibt, kann man Stabdiagramme verwenden, bei denen die X-Achse die möglichen Merkmalswerte und die Y-Achse die zugehörigen

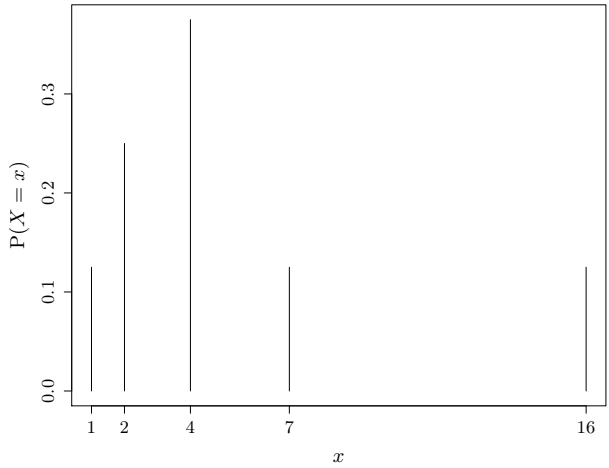

Abbildung 1.1: Relative Häufigkeiten des Zahlenbeispiels.

Häufigkeiten angibt. Abbildung 1.1 zeigt das Stabdiagramm für die 8 Beispielswerte. R-1-1

1.2.5 Klassierung

Wenn es sehr viele unterschiedliche Merkmalswerte gibt, ist es oft hilfreich, Merkmalsklassen zu verwenden. Wenn z. B. eine Variable die monatlichen Einkommen von Haushalten erfasst, könnten Einkommensklassen gebildet werden, und die klassierte Variable erfasst dann nur, in welcher Einkommensklasse sich ein Haushalt befindet. In dem oben angeführten Zahlenbeispiel könnten vier Klassen gebildet werden: $\tilde{x}_1^* = \{1, 2\}$, $\tilde{x}_2^* = \{3, 4\}$, $\tilde{x}_3^* = \{5, 6\}$, $\tilde{x}_4^* = \{7, 8\}$. Die klassierte Variable X^* nimmt dann einen dieser vier Werte an, und es gilt: $\mathrm{P}(X^* = \tilde{x}_j^*) = \mathrm{P}(X \in \tilde{x}_j^*)$; zum Beispiel $\mathrm{P}(X^* = \tilde{x}_1^*) = 3/8$.

1.3 Ein Beispiel mit Einkommensdaten

In diesem Abschnitt erläutern wir einen Beispielsdatensatz, der dann in den meisten folgenden Kapiteln zur Illustration von Konzepten und Methoden verwendet wird.

1.3.1 Datenquelle: Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)

Als Beispieldatensatz verwenden wir Daten der Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) des Jahres 2018.[1] Der ALLBUS wird seit 1980 in der Regel alle zwei Jahre durchgeführt und enthält Daten über Einstellungen, Verhaltensweisen und Sozialstruktur der Bevölkerung in der Bundesrepublik Deutschland.[2] Der ALLBUS ist angelehnt an den General Social Survey (GSS), der in den USA seit 1972 regelmäßig durchgeführt wird. Die Anzahl der Befragten in den veröffentlichten Daten liegt zwischen 2,800 und 3,500, in 2018 liegen Informationen für 3,477 Befragte vor. Das Untersuchungsgebiet des ALLBUS ist Deutschland und die Grundgesamtheit sind alle Personen, die zum Befragungszeitpunkt in Privathaushalten lebten und vor dem 01.01.2000 geboren sind.

Die Auswahl erfolgt als zweistufige, disproportional geschichtete Zufallsauswahl in Westdeutschland (inkl. West-Berlin) und Ostdeutschland (inkl. Ost-Berlin). In der ersten Auswahlstufe wurden Gemeinden in Westdeutschland und in Ostdeutschland mit einer Wahrscheinlichkeit proportional zur Zahl ihrer erwachsenen Einwohner ausgewählt. In der zweiten Auswahlstufe wurden Personen aus den Einwohnermeldekarteien zufällig gezogen. Die einzelnen Querschnittsdatensätze, haben neben einem Kernfrageprogramm wechselnde inhaltliche Schwerpunkte und dienen der Untersuchung von Einstellungen und Verhaltensweisen der deutschen Bevölkerung.

1.3.2 Die Variablen im Datensatz

Der Datensatz enthält für $n = 1,747$ Personen die folgenden Variablen:

- *id*: Identifizierer der Personen, laufende Nummer von 1 bis 1,747

[1] GESIS - Leibniz-Institut für Sozialwissenschaften (2019): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2018. GESIS Datenarchiv, Köln. ZA5270 Datenfile Version 2.0.0, doi:10.4232/1.13250.

[2] Informationen finden sich auf der folgenden Webseite: https://www.gesis.org/allbus/allbus/.

- *geschlecht:* Geschlecht der Person, 0 Mann, 1 Frau
- *alter:* Das Lebensalter der Person
- *ostwest:* Indikator für die Region, 0 Westdeutschland, 1 Ostdeutschland
- *land:* Das Bundesland, BW (Baden-Württemberg), BY (Bayern), BE (Berlin), BB (Brandenburg), HB (Bremen), HH (Hamburg), HE (Hessen), MV (Mecklenburg-Vorpommern), NI (Niedersachsen), NW (Nordrhein-Westfalen), RP (Rheinland-Pfalz), SL (Saarland), SN (Sachsen), ST (Sachsen-Anhalt), SH (Schleswig-Holstein), TH (Thüringen)
- *bildung:* Die Anzahl der Ausbildungsjahre[3] wurde aus der Summe der Schul- und Ausbildungsjahre gebildet. Die Schuljahre wurden ausgehend von Angaben zur Schulausbildung berechnet (Kein Abschluss 7, Hauptschulabschluss 9, Realschulabschluss 10, Fachhochschulabschluss 12, Abitur 13, Andere 10). Die Ausbildungsjahre wurden ausgehend von Angaben zur Berufsausbildung berechnet (Lehre 1,5, Berufsfachschule/Gesundheitswesen 2, Beamtenausbildung 1,5, Fachhochschule 3, Universtität 5)
- *beruf:* Der Beruf wurde aus den Angaben der Berufsklassifikation nach ISCO 08 gewonnen (1 Fuehrungskraefte, 2 Akademiker, 3 Techniker, 4 Buerokraefte, 5 Dienstleister, 6 Bauern, 7 Handwerker, 8 Monteure, 9 Hilfsarbeiter
- *stunden:* Anzahl der monatlichen Arbeitsstunden (ermittelt als gerundeter Wert der 4,3-fachen wöchentlichen Arbeitszeit)
- *einkommen:* Das zusammengefasste monatliche Netto-Einkommen des Befragten.
- *stlohn:* Aus Monatseinkommen und monatlichen Arbeitsstunden berechneter Netto-Stundenlohn.

[3] Die Variable wurde entsprechend der Vorgehensweise beim Sozioökonomischen Panel gebildet. Vgl. John P. Haisken-DeNew und Joachim R. Frick, DTC Desktop Companionto the GermanSocio-Economic Panel (SOEP), Version 8.0 - Dec 2005, S. 69.

Tabelle 1.1: Ein Ausschnitt des Datensatzes.

id	geschlecht	alter	land	beruf	stlohn
1	0	62	BY	3	13.50
2	1	64	ST	3	6.98
3	0	22	NI	7	7.22
⋮	⋮	⋮	⋮	⋮	⋮
1745	1	60	HH	5	1.80
1746	0	54	BY	2	6.30
1747	0	49	NI	2	55.56

Tabelle 1.1 zeigt einen Ausschnitt des Datensatzes. Die erste Spalte (*id*) enthält eine durchgängige Nummerierung aller $n = 1747$ Personen. Für die ersten und letzten drei Personen sind in diesem Ausschnitt jeweils in einer Zeile die Ausprägungen der aufgeführten Merkmale angegeben. R-1-2

1.4 Aufgaben

1. Mit dieser Aufgabe soll der Umgang mit Summen und Produkten, die in der Statistik sehr häufig verwendet werden, in Erinnerung gerufen werden. Gegeben sind:

i	1	2	3	4
x_i	6	4	1	3
y_i	1	3	4	2

Berechnen Sie:

a) $\sum_{i=1}^{4} x_i$ c) $\prod_{i=1}^{4} x_i$ e) $\prod_{i=1}^{4} x_i^2 y_i^{0.5}$

b) $\sum_{i=1}^{4} x_i y_i$ d) $\prod_{i=1}^{4} x_i y_i$

2. Berechnen Sie möglichst einfach (Hinweise zu Summen finden Sie in der Formelsammlung S. 219):

a) $\sum_{i=1}^{20} (6 - 4i) + \sum_{i=1}^{20} (2i + 2) + \sum_{i=1}^{20} (-4 - 4i)$

b) $\sum_{i=1}^{30} \left(i^2 + 2i - 3\right) + \sum_{i=1}^{30} \left(3i^2 + 5i + 8\right) + \sum_{i=1}^{30} \left(4i^2 + 6i - 10\right)$

c) $\sum_{i=1}^{40} (1 + i)^2 + \sum_{i=1}^{40} (1 - i)^2$

3. Gegeben ist folgende Matrix $B = (b_{ij})$; $i = 1, \ldots, I$ ist der Zeilenindex und $j = 1, \ldots, J$ der Spaltenindex:

$$B = \begin{pmatrix} 1 & 4 & 4 & 7 & 8 & 4 \\ 2 & 3 & 6 & 6 & 2 & 3 \\ 6 & 9 & 7 & 6 & 7 & 2 \\ 5 & 7 & 8 & 8 & 9 & 6 \\ 4 & 6 & 2 & 3 & 4 & 5 \\ 3 & 5 & 2 & 3 & 7 & 7 \end{pmatrix}$$

Berechnen Sie:

a) $\sum_{i=1}^{2}\sum_{j=1}^{3} b_{ij}$ c) $\sum_{j=1}^{J} b_{2j}$ e) $\sum_{i=3}^{4}\sum_{j=5}^{6} b_{ij}$

b) $\sum_{i=2}^{2}\sum_{j=1}^{J} b_{ij}$ d) $\sum_{i=1}^{I}\sum_{j=1}^{2} b_{ij}$

4. Informieren Sie sich im Internet über den ALLBUS und versuchen Sie, folgende Fragen zu beantworten:

a) Was ist eine Querschnitts-, was eine Panelerhebung?

b) Wie werden die befragten Haushalte ausgewählt?

c) Hat jeder Haushalt in Deutschland die gleiche Chance ausgewählt zu werden?

d) Welche Informationen liefert der ALLBUS?

e) Welche Schwerpunkte hat das Frageprogramm des ALLBUS im Jahr 2018?

5. Geben Sie bei den nachfolgenden Variablen an, welches Skalenniveau sie besitzen: Geschlecht, Beruf, Warengruppe, Immobilienbesitz, Bonität, Einkommen, Vermögen.

6. Ermitteln Sie für die folgenden Werte einer Variablen X

$$1, 4, 5, 4, 5, 4, 5, 4, 6, 1, 2, 1, 1, 2, 1$$

die vorkommenden Merkmalsausprägungen ($\tilde{x}_j$) und deren absolute (n_j) und relative (f_j) Häufigkeiten.

7. Die folgende Tabelle enthält die Häufigkeiten der ALLBUS Monatslöhne (in Euro) von Personen in Westdeutschland für 6 Lohnklassen unterschiedlicher Klassenbreite.

Klasse	von	bis unter	abs. Häuf.
$\tilde{x}_1^*$	0	500	22
$\tilde{x}_2^*$	500	1000	142
$\tilde{x}_3^*$	1000	1500	183
$\tilde{x}_4^*$	1500	2500	454
$\tilde{x}_5^*$	2500	5000	369
$\tilde{x}_6^*$	5000	20000	68

Ermitteln Sie für die Lohnklassen die Häufigkeiten $\mathrm{P}(X^* = \tilde{x}_j^*)$.

1.5 R-Code

R-1-1

```
# Vektor a mit Merkmalswerten erstellen
x <- c(1,2,2,4,4,4,7,16)
# Anzahl an Merkmalswerten
n <- length(x)
# absolute Häufigkeiten
table(x)
# relative Häufigkeiten
f.x <- table(x) / n
f.x
# Stabdiagramm der relativen Häufigkeiten
plot(f.x)
```

R-1-2

```
# Daten einlesen
d <- read.csv2(file = "allbus2018.csv", stringsAsFactors = FALSE)
# Übersicht: Erste und letzte 6 Beobachtungen anzeigen
head(d)
tail(d)
# Anzahl an Beobachtungen
n <- nrow(d)
n
```