Sandro Scheid
Stefanie Vogl

Data Science

Grundlagen, Methoden und Modelle der Statistik

HANSER

Scheid / Vogl

Data Science

Sandro Scheid
Stefanie Vogl

Data Science

Grundlagen, Methoden und Modelle der Statistik

HANSER

Autoren:

Dr. Sandro Scheid, Hochschule für angewandte Wissenschaften, München
Prof. Dr. Stefanie Vogl, Hochschule für angewandte Wissenschaften, München

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

Internet: www.hanser-fachbuch.de

Lektorat: Frank Katzenmayer
Herstellung: Anne Kurth
Covergestaltung: Max Kostopoulos
Coverkonzept: Marc Müller-Bremer, www.rebranding.de, München
Titelbild: © gettyimages.de/gremlin
Satz: Dr. Sandro Scheid, Prof. Dr. Stefanie Vogl
Druck und Bindung: CPI books GmbH, Leck
Printed in Germany

Print-ISBN 978-3-446-46663-0
E-Book-ISBN 978-3-446-47001-9

Inhalt

Vorwort

In diesem Buch, das aus Vorlesungen für Studierende der Betriebswirtschaftslehre entstanden ist, werden wichtige Methoden der Datenanalyse vorgestellt. Statistische Verfahren werden stets dann benötigt und eingesetzt, wenn im Rahmen empirischer Fragestellungen Daten erhoben, dargestellt und analysiert werden sollen. In allen empirischen Wissenschaften – wir nennen beispielhaft die Wirtschaftswissenschaften – hat die Statistik daher eine große praktische Bedeutung. Zum Studium dieser Wissenschaften gehört deshalb auch eine intensive Beschäftigung mit Statistik.

Dieses Buch ist ein idealer Begleiter zu den Statistikvorlesungen des Bachelor-Studiums an Hochschulen und Universitäten und für das Nacharbeiten statistischer Themen im Masterstudium. Für Praktiker bietet es die Gelegenheit, sich im Selbststudium mit statistischen Fragestellungen zu befassen.

Das Buch setzt keine besonderen mathematischen Kenntnisse voraus. Der Leser benötigt die üblichen Grundkenntnisse der Elementarmathematik, wie sie an Fachoberschulen und Gymnasien unterrichtet wird. Die darüber hinausgehenden Anforderungen beschränken sich im Wesentlichen auf das Rechnen mit dem Summenzeichen.

Das Buch behandelt im ersten Teil (Kapitel 1 bis 3) die deskriptive Statistik, mit der die Datenanalyse beginnen sollte. Im zweiten Teil (Kapitel 4 bis 7) wird die Wahrscheinlichkeitsrechnung behandelt, die als Grundlage für die weiteren Kapitel benötigt wird. Teil drei behandelt die klassischen Themen der induktiven Statistik (Kapitel 8 bis 11). Danach werden im vierten Teil (Kapitel 12 bis 16) verschiedene weiterführende Methoden der Datenanalyse behandelt.

Zu beachten ist, dass die Rechnungen zu den komplexeren Beispielen teilweise mit Hilfe von Software durchgeführt wurden. Hier sind, um die Kumulation von Rundungsfehlern zu vermeiden, Zwischenergebnisse gerundet angegeben. Intern wurde aber mit einer höheren Genauigkeit weitergerechnet. Endergebnisse können sich also beim Rechnen mit gerundeten Zwischenergebnissen leicht unterscheiden.

An dieser Stelle wollen wir unseren Dank aussprechen. Unser Dank gilt zunächst den Studierenden der Hochschule München, die durch zahlreiche Fragen und Anmerkungen zur Gestaltung des Lehrtextes beigetragen haben. Weiter danken wir ganz herzlich Frau Julia Pelger, die durch ihre Korrekturvorschläge eine unersetzliche Arbeit geleistet hat, und Herrn Oliver Hien, durch dessen konstruktive Anregungen und Kommentare der Text weiter wesentlich verbessert werden konnte.

Unser ganz besonderer Dank gilt Herrn Frank Katzenmayer und Frau Christina Kubiak, Hanser Verlag, für die exzellente Betreuung des Buchprojekts.

Kochel am See und Konzell, Sommer 2021 Sandro Scheid und Stefanie Vogl

1 Deskriptive Statistik

1 Grundlagen

Die deskriptive Statistik stellt Methoden bereit, um Untersuchungsgegenstände, die aus einer Vielzahl von einzelnen Objekten bestehen, zu beschreiben. Hinter dieser abstrakten Beschreibung stehen vielfältige unterschiedliche Anwendungen. Beispielsweise soll das Wahlverhalten der deutschen Bürger beschrieben werden oder die Nebenwirkungen bei der Einnahme eines Medikamentes oder das Interesse der Konsumenten an einem neuen Produkt. Die drei genannten Beispiele haben gemeinsam, dass das konkrete Interesse auf eine Gesamtheit bezogen ist. Dies sind die wahlberechtigten Personen in Deutschland, die Patienten, die das Medikament einnehmen oder der Kundenkreis des neuen Produkts. Diese Gesamtheiten sollen durch die statistische Analyse quantitativ beschrieben werden. Dazu werden an den Einheiten – in den genannten Beispielen sind dies Personen – Untersuchungsmerkmale betrachtet. Eine deskriptive Auswertung sollte das Untersuchungsziel und auch die Gesamtheit genauer beschreiben. Wichtige Begriffe, die dabei nützlich sind, werden in diesem Kapitel behandelt.

1.1 Aufgaben der deskriptiven Statistik

Die deskriptive Statistik beschäftigt sich mit der Aufbereitung und Komprimierung von Daten, die in den verschiedensten Bereichen anfallen oder erhoben werden. Mithilfe der deskriptiven Statistik können Sachgebiete quantitativ beschrieben werden. Dabei werden viele Daten im Hinblick auf die Eigenschaften, die von Interesse sind, zusammengefasst. Zum Aufgabengebiet der deskriptiven Statistik zählen im Einzelnen:

- das Zusammenfassen und Ordnen der Daten in Tabellen,
- das Erstellen von Diagrammen und
- das Berechnen charakteristischer Kenngrößen oder Maßzahlen.

Beispiel 1.1

- Der Absatz der verschiedenen Modelle eines Kfz-Herstellers soll graphisch dargestellt werden.
- Die Einkommensstruktur der Bundesbürger soll komprimiert dargestellt werden. Von Interesse ist hier der Anteil der Bevölkerung an der Gesamtbevölkerung, der zu einer Einkommensklasse gehört, wobei das Einkommen in verschiedene Größenklassen aufgeteilt ist.
- Ein Kraftfahrzeugversicherer möchte die eingegangenen Schadensfälle verschiedenen Merkmalen zuordnen. So ordnet er sie dem Kraftfahrzeugtyp und dem Alter der Versicherungsnehmer zu.

■

Deskriptive Statistiken werden in unterschiedlichsten Bereichen benötigt. So stellen etwa die staatlichen Statistikämter unter anderem Statistiken der Bevölkerungsstruktur, der Erwerbstätigkeit, des produzierenden Gewerbes, der Sozialleistungen für Handel und Verkehr oder für das Gesundheitswesen zur Verfügung. Im Bereich Betriebsstatistik erfassen Unternehmen etwa Auftragseingänge, Produktion, Erzeugerpreise sowie Lohn- und Materialkosten. In der Forschung werden mittels Statistik Forschungsgegenstände wie z. B. die Verträglichkeit von Medikamenten, das Wahlverhalten, die Belastungsfähigkeit unterschiedlicher Materialien, das Kaufverhalten der Geschlechter oder das Armutsrisiko von Patchworkfamilien beschrieben.

1.2 Grundgesamtheit und Stichprobe

Ausgangspunkt einer statistischen Fragestellung ist ein Untersuchungsgegenstand, der sich in der Regel auf viele einzelne Objekte bezieht. Dabei sollte der Untersuchungsgegenstand präzise formuliert und außerdem sachlich, räumlich und zeitlich abgegrenzt sein, damit eindeutig feststeht, auf welche konkreten Objekte er sich bezieht. Für die Vielzahl an möglichen Untersuchungsgegenständen sollen nun exemplarisch drei Fälle illustriert werden.

Beispiel 1.2

Eine Firma überlegt, ein neues Produkt einzuführen. Sie möchte wissen, wie es auf dem Markt ankommen würde. Um das herauszufinden, führt das Unternehmen eine Umfrage durch.
Untersuchungsgegenstand ist also die Akzeptanz eines neuen Produkts auf dem Markt. Dazu muss der potenzielle Absatzmarkt festgelegt werden. Welche geographische Größe ist relevant, ist der Markt national, europaweit oder global definiert? Sollen nur Ballungszentren bedient werden oder die ganze Fläche? Welches Geschlecht und welche Altersgruppen kommen infrage? Welche Einkommensklassen? Welche Vertriebswege sollen genutzt werden?

■

Beispiel 1.3

Ein Kfz-Versicherer möchte die Schadensfälle des abgelaufenen Geschäftsjahres nach Schadenshöhe und geografischer Häufigkeit charakterisieren.
Untersuchungsgegenstand ist die Gesamtheit der im abgelaufenen Geschäftsjahr eingegangenen Schadensfälle. Somit müssen Beginn und Ende des Geschäftsjahres durch konkrete Zeitpunkte markiert werden. Sollen sämtliche Versicherungstarife analysiert werden oder gilt das Interesse nur einer Auswahl?

■

Wie an den Beispielen deutlich wird, bezieht sich der Untersuchungsgegenstand stets auf eine Menge von Objekten. Diese werden in der Statistik als Untersuchungseinheiten bezeichnet. Die Menge aller Untersuchungseinheiten bildet dann die Grundgesamtheit.

Untersuchungseinheit
Untersuchungseinheiten sind die Objekte, auf die sich die statistische Analyse bezieht.

Grundgesamtheit
Die Menge aller Untersuchungseinheiten wird als Grundgesamtheit oder Population bezeichnet.

Meist werden in der Praxis nicht alle Untersuchungseinheiten, auf die sich eine Fragestellung bezieht, analysiert. Dies wäre aus organisatorischen und zeitlichen Gründen oft viel zu aufwendig oder sogar vollkommen unmöglich. Häufig ist das auch gar nicht notwendig. Die moderne Statistik stellt nämlich Methoden zur Verfügung, die es ermöglichen, basierend auf einer relativ kleinen Auswahl von Untersuchungseinheiten allgemein gültige Aussagen bezüglich einer weitaus größeren Grundgesamtheit herzuleiten.

Stichprobe
Eine Stichprobe bezeichnet eine Auswahl an Untersuchungseinheiten einer Grundgesamtheit.

Die Vorgehensweise, Ergebnisse einer Teilgesamtheit auf eine übergeordnete Gesamtheit zu verallgemeinern, ist Aufgabe der induktiven Statistik. Die deskriptive Statistik dient ausschließlich nur zur Beschreibung einer vollständig bekannten Grundgesamtheit oder zur Beschreibung einer Stichprobe. Interpretationen, die darüber hinausgehen, sind nicht Gegenstand der Untersuchungen und bleiben der induktiven Statistik vorbehalten.

1.3 Merkmale und Skalenniveaus

Bei einer Untersuchung werden an den Untersuchungseinheiten Merkmale betrachtet, die dem jeweiligen Interesse entsprechen. Merkmale werden mit Großbuchstaben $X, Y, Z, \ldots$ notiert. Allgemein sind an einer statistischen Einheit in der Regel mehrere Merkmale beobachtbar.

Beispiel 1.4

Sind die Untersuchungseinheiten Personen, so könnten folgende Merkmale von Interesse sein:
X = das Monatseinkommen,
Y = die bei der letzten Bundestagswahl gewählte Partei,
Z = das Geschlecht.

■

Beispiel 1.5

Sind die Untersuchungseinheiten Schadensereignisse, die bei einer Versicherung innerhalb eines Zeitraums eingehen, so könnten folgende Merkmale von Interesse sein:
X = die Schadenshöhe,
Y = der Zeitpunkt des Schadens,
Z = die Versicherungshöhe des Versicherungsnehmers.

■

Neben dem Begriff Merkmal wird häufig auch der Begriff Variable verwendet. Beide Begriffe werden synonym angewandt. Der Begriff Variable deutet schon darauf hin, dass die Variable oder das Merkmal für die verschiedenen Untersuchungseinheiten verschiedene Werte annehmen kann. Die Werte, die die Merkmale annehmen, werden als Merkmalsausprägungen bezeichnet. Die möglichen Ausprägungen eines Merkmals bilden den Merkmals- oder Zustandsraum. Die folgende Übersicht enthält nochmals eine Zusammenfassung der beschriebenen Begriffe.

Merkmal
Unter einem Merkmal versteht man bestimmte Aspekte oder Eigenschaften einer Untersuchungseinheit.

Merkmalsausprägung
Den konkreten Wert, den eine Untersuchungseinheit für ein Merkmal annimmt, bezeichnet man als Merkmalsausprägung.

Merkmalsraum
Die Menge der möglichen Ausprägungen eines Merkmals wird mit Merkmals- oder Zustandsraum bezeichnet.

In der Literatur ist es üblich, folgende Begriffe synonym zu verwenden:
Untersuchungseinheit = Merkmalsträger = Objekt = Fall (Case),
Merkmal = Variable = Attribut.

Beispiel 1.6

Im *Beispiel 1.4* wurden Personen als Einheiten betrachtet. Die genannten Merkmale könnten die Werte
X = Monatseinkommen: 1 700 €, 2 100 €, 1 900 €, ...
Y = gewählte Partei: CDU, SPD, Grüne, FDP, ...
Z = Geschlecht: männlich, weiblich
annehmen.

■

Das Beispiel illustriert, dass die Art der Ausprägungen unterschiedlich sein kann. Im Folgenden sollen die verschiedenen möglichen Arten von Merkmalsausprägungen beschrieben werden. Je nach Art der Merkmalsausprägungen werden sich später die möglichen statistischen Methoden, die zur Auswertung der Daten herangezogen werden können, unterscheiden.
Eine grobe Einteilung der Merkmalsausprägungen ergibt sich durch die Unterscheidung in zahlenmäßige und nicht zahlenmäßige Werte. Die verwendeten Begriffe sind:

Qualitative Merkmale
Die Ausprägungen unterscheiden sich in ihrer Art. Sie können nicht durch eine Größenordnung beschrieben werden, aber in verschiedene Kategorien eingeteilt werden.

Quantitative Merkmale
Die Ausprägungen lassen sich durch Zahlen beschreiben. Sie besitzen eine Ausprägung, bei der die Größe interpretiert werden kann.

Oft werden die Kategorien von qualitativen Merkmalen im praktischen Gebrauch mit Zahlen belegt. So stehen z. B. Steuer- oder Kundennummern für verschiedene Personen oder Steuerklassen für unterschiedliche Personengruppen. Bankleitzahlen stehen für Banken, Postleitzahlen für Orte. Wichtig für das Verständnis des Begriffs des qualitativen Merkmals ist hierbei, dass die jeweiligen Zahlen in diesen Fällen nur der Unterscheidung dienen. Eine Anordnung der Merkmale im Sinne einer Größe ist hier nicht sinnvoll.

Beispiel 1.7

- Qualitative Merkmale: Geschlecht, Familienstand, Konfession, Steuerklasse, Farbe eines Autos, Rechtsform eines Unternehmens.
- Quantitative Merkmale: Körpergröße, Alter, Einkommen, Umsatz, Temperatur.

■

Quantitative Merkmale lassen sich weiter unterteilen, je nachdem, ob die Merkmalsausprägungen nur diskrete oder kontinuierliche Werte annehmen können.

Diskrete Merkmale
Die Merkmalsausprägungen nehmen nur bestimmte, separate Zahlenwerte an. Es werden nur endlich viele oder abzählbar unendlich viele Werte angenommen.

Stetige Merkmale
Es können Werte aus einem reellen Zahlenintervall angenommen werden. Die Werte sind kontinuierlich. Man stellt sich vor, dass sie auf beliebig viele Nachkommastellen messbar sind.

Bemerkung:

- Alle qualitativen Merkmale sind trivialerweise diskret. Quantitative Merkmale sind dann diskret, wenn die Merkmalsausprägungen durch einen Zählvorgang ermittelt werden können.
- Merkmale, die sich sehr fein unterteilen lassen, aber im Prinzip diskrete Merkmale sind, werden auch quasi-stetige Merkmale genannt und zu den stetigen Merkmalen gezählt.

Beispiel 1.8

- Diskrete Merkmale: Anzahl der Semester eines Studierenden, die Anzahl der Angestellten in einem Betrieb, Anzahl der Fehltage eines Arbeitnehmers.
- Stetige Merkmale: Körpergröße, Temperatur, Zeit, Benzinverbrauch.
- Quasi-stetige Merkmale: Einkommen eines Haushalts, Umsatz einer Firma, Quadratmeterpreis einer Wohnung.

■

Je nach Art des betrachteten Merkmals können die Merkmalsausprägungen anhand verschiedener Skalen gemessen werden. Im Folgenden sind die verschiedenen Skalenniveaus, geordnet nach dem Informationsgehalt, beginnend beim niedrigsten, beschrieben.

Skalenniveaus

- Nominales Skalenniveau
  Bei einem nominalen Skalenniveau lassen sich die verschiedenen Ausprägungen des Merkmals lediglich unterscheiden. Es gibt keine natürliche Anordnung der Merkmalsausprägungen. Man spricht von einem nominalen Merkmal.
- Ordinales Skalenniveau
  Es gibt eine Rangfolge bzw. Ordnung innerhalb der Ausprägungen des Merkmals. Der Abstand zwischen den Ausprägungen ist aber nicht sinnvoll interpretierbar. Man spricht von einem ordinalen Merkmal.
- Metrisches Skalenniveau
  Die Ausprägungen des Merkmals lassen sich der Größe nach anordnen. Zudem sind die Abstände der Ausprägungen interpretierbar. Man spricht von einem quantitativen bzw. metrischen Merkmal.

Beispiel 1.9 Skalen

- Beispiele für Merkmale, die auf einer nominalen Skala gemessen werden, sind: Geschlecht, Familienstand, Konfession, Augenfarbe, Rechtsform eines Unternehmens, Branche eines Unternehmens.
- Beispiele für Merkmale, die auf einer ordinalen Skala gemessen werden, sind: Leistungsbeurteilung, Bewertung bei einem Schönheitswettbewerb, Zufriedenheit mit einem Produkt mit beispielsweise den Ausprägungen: sehr zufrieden, zufrieden, unzufrieden, sehr unzufrieden.

- Beispiele für Merkmale, die auf einer metrischen Skala gemessen werden, sind: Körpergröße, Wartezeit auf den Bus, Gewinn eines Unternehmens, Einkommen eines Arbeitnehmers.

■

Für die metrischen Merkmale sind weitere Unterteilungen möglich.

Arten metrischer Skalen

- Intervallskala
  Nur die Differenzen zwischen Ausprägungen können interpretiert werden.
- Verhältnisskala
  Das Merkmal besitzt einen absoluten Nullpunkt. Als Folge davon ergibt eine Aussage wie „die eine Ausprägung ist doppelt so hoch wie die andere“ Sinn. Allgemein sind die Verhältnisse der verschiedenen Ausprägungen interpretierbar.

Beispiel 1.10 Metrische Skalen

- Ein Merkmal, das auf einer Intervallskala gemessen wird, ist die Temperatur. Abstände sind hier interpretierbar. Eine Aussage wie „heute ist es 3 Grad wärmer als gestern“ ist sinnvoll. Bei null Grad Celsius handelt es sich nicht um einen natürlichen Nullpunkt. Angenommen heute sind 6 °C und gestern 3 °C gemessen worden, so ist es nicht sinnvoll zu sagen, dass es heute doppelt so warm ist wie gestern.
- Ein Beispiel für ein Merkmal, das auf einer Verhältnisskala gemessen wird, ist das Einkommen eines Arbeitnehmers. Hier bietet es sich an, Verhältnisse zu bilden. Aussagen wie „Peter verdient doppelt so viel wie Dieter“ oder „Petra verdient 15 % mehr als Susanne“ sind zulässig.

■

2 Betrachtung eines Merkmals

Um sich einen Überblick bezüglich wesentlicher Eigenschaften eines Merkmals anzueignen, beginnt man mit der Häufigkeitsverteilung. Diese Verteilung beschreibt, wie häufig die einzelnen Merkmalsausprägungen in der Grundgesamtheit zu finden sind. Häufigkeiten lassen sich für jedes Merkmal und jedes Skalenniveau ermitteln. In diesem Kapitel werden – getrennt für diskrete und stetige Merkmale – Häufigkeitsbegriffe erörtert und graphische Darstellungen vorgestellt.
Ferner werden Methoden erarbeitet, mit denen sich die charakteristischen Eigenschaften eines einzelnen Merkmals beschreiben lassen. Die geeigneten Methoden sind abhängig von der Art des jeweiligen Merkmals, insbesondere von dessen Skalenniveau. Man unterscheidet bei diesen statistischen Kenngrößen oder Maßzahlen hierbei Lagemaße und Streuungsmaße. Den Abschluss des Kapitels bildet die Fragestellung der Konzentration von Merkmalen auf Merkmalsträger.

2.1 Häufigkeitsverteilungen bei diskreten Merkmalen

2.1.1 Absolute und relative Häufigkeitsverteilung

Zu den diskreten Merkmalen können wir hier alle qualitativen sowie die quantitativ-diskreten Merkmale zählen. Die Anzahl der unterschiedlichen Merkmalsausprägungen ist in der Regel wesentlich kleiner als der Umfang der Grundgesamtheit und damit überschaubar. So gehören beispielsweise zum qualitativen Merkmal „Geschlecht“ die Ausprägungen „weiblich“, „männlich“ und „divers“ . Durch einfaches Abzählen lässt sich ermitteln, wie häufig die Ausprägungen in der Grundgesamtheit vertreten sind.

Wir bezeichnen mit

N den Umfang der Grundgesamtheit, bestehend aus N Untersuchungseinheiten,

x_i die Merkmalsausprägung des Merkmals X, die bei der i-ten Untersuchungseinheit beobachtet wurde.

Nummeriert man die Untersuchungseinheiten fortlaufend von 1 bis N durch, dann enthält die Urliste die statistischen Daten so, dass jeder Untersuchungseinheit i die Merkmalsausprägung x_i zugeordnet ist.

Beispiel 2.1 Haushaltsgröße von Privathaushalten

Für $N = 40$ Privathaushalte liegen für das Merkmal „Anzahl Personen“ folgende Daten vor:

3	5	2	3	3	2	3	5	3	5	5	3	2	1	4	3	4	4	2	4
3	3	1	4	5	3	2	4	1	4	1	2	2	3	1	3	1	4	2	2

Die Daten liegen in Form einer Urliste vor, die man auch in standardisierter Form mit 40 Zeilen und zwei Spalten notieren könnte, wobei die erste Spalte lediglich zur Numerierung der Untersuchungseinheiten (Objekte) dient. Einem Objekt entspricht ein Haushalt, der das Merkmal X = „Anzahl Personen“ besitzt. Beispielsweise besitzt der erste Haushalt (erstes Objekt) die Merkmalsausprägung $x_1 = 3$, der zweite Haushalt (zweites Objekt) die Merkmalsausprägung $x_2 = 5$, usw., wie man der ersten Zeile in der obigen Urliste entnimmt. ■

Um Fragen der Form „Wie viele Haushalte sind 4-Personen-Haushalte in dieser Grundgesamtheit?“ beantworten zu können, ordnet man den in der Grundgesamtheit vorkommenden unterschiedlichen Merkmalsausprägungen ihre absoluten Häufigkeiten zu, d.h. man ermittelt durch einfaches Abzählen, wie viele 4-Personen-Haushalte es in der Grundgesamtheit gibt.

Allgemein formuliert man diesen Sachverhalt folgendermaßen:
Ein diskretes Merkmal X habe $M \leq N$ verschiedene Ausprägungen $x_1, \ldots, x_M$. Die absolute Häufigkeit einer Ausprägung x_j wird mit h_j bezeichnet. Der Buchstabe j ist der sogenannte Laufindex, der zwischen 1 und M variiert. Die Summe aller absoluten Häufigkeiten h_j entspricht der Anzahl der Untersuchungseinheiten.

Absolute Häufigkeiten

$h(x_j) = h_j$	absolute Häufigkeit der Ausprägung x_j
$h_1, \ldots, h_M$	absolute Häufigkeitsverteilung
$0 \leq h_j \leq N$	$j = 1, \ldots, M$
$h_1 + \ldots + h_M = \sum_{j=1}^{M} h_j = N$	

Will man den Anteil der 4-Personen-Haushalte in der Grundgesamtheit ermitteln, so benötigt man die Anzahl N der Untersuchungseinheiten. Man spricht dann von der relativen Häufigkeit, mit der die Merkmalsausprägung in der Grundgesamtheit vorkommt.

Relative Häufigkeiten

$f(x_j) = f_j$	relative Häufigkeit der Ausprägung x_j
$f_j = \frac{h_j}{N}$	$j = 1, \ldots, M$
$f_1, \ldots, f_M$	relative Häufigkeitsverteilung
$0 \leq f_j \leq 1$	$j = 1, \ldots, M$
$f_1 + \ldots + f_M = \sum_{j=1}^{M} f_j = 1$	

In der Praxis gewinnt man die Häufigkeiten am einfachsten durch das Erstellen einer Strichliste oder – weniger mühsam – mittels einer geeigneten Statistiksoftware.

Beispiel 2.2 Privathaushalte (Fortsetzung)

Für die Daten der 40 Privathaushalte aus *Beispiel 2.1* ergeben sich folgende Häufigkeiten:

Haushaltsgröße x_j	Strichliste	absolute Häufigkeiten h_j	relative Häufigkeiten f_j
1	//////	6	6/40 = 0,150
2	//////////	9	9/40 = 0,225
3	///////////	12	12/40 = 0,300
4	////////	8	8/40 = 0,200
5	/////	5	5/40 = 0,125
$\sum$		40	1,0

Die Anzahl der 4-Personen-Haushalte unter den betrachteten 40 Privathaushalten beträgt also 8, bzw. 20 % der Haushalte sind 4-Personen-Haushalte.

■

Beispiel 2.3 Studienwahl

Neu eingeschriebene Studierende an einer kleinen Hochschule verteilen sich auf die in der Tabelle aufgeführten Fächergruppen. Folgende Anzahlen ergeben sich für ein ausgewähltes Semester:

Fächergruppe	Neueinschr., Anzahl	Neueinschr. in %
Recht-, Wirtschafts- und Sozialwissenschaften	453	49
Ingenieurwissenschaften	232	25
Mathematik, Naturwissenschaften	147	16
Agrar-, Forst- und Ernährungswissenschaften	88	10
$\sum$	920	100,0

Objekt = neu eingeschriebene Studierende an einer kleinen Hochschule
Grundgesamtheit = alle neu eingeschriebene Studierende an einer kleinen Hochschule in einem ausgewählten Semester
Untersuchungsmerkmal (qualitativ, nominalskaliert) X = Fächergruppe
Merkmalsausprägungen: x_1 = Recht-, Wirtschafts- und Sozialwissenschaften, x_2 = Ingenieurwissenschaften, x_3 = Mathematik, Naturwissenschaften, x_4 = Agrar- Forst- und Ernährungswissenschaften
Die zu den Merkmalsausprägungen $x_1, \ldots, x_4$ gehörenden absoluten Häufigkeiten $h_1, \ldots, h_4$ findet man in der Spalte „Neueinschr., Anzahl“, die relativen Häufigkeiten (in Prozent) $f_1, \ldots, f_4$ in der Spalte „Neueinschr. in %“.

■

Bemerkung:
Die relative Häufigkeit wird oft in Prozentwerten angegeben. Da der Ausdruck Prozent „von Hundert“ bedeutet, sind derartige Angaben nur bei einer hinreichend großen Grundgesamtheit sinnvoll. Wenn man bei kleineren Grundgesamtheiten mit weniger als 40 Untersuchungseinheiten Prozente berechnet, täuscht man eine Größe vor, die in Wirklichkeit nicht vorhanden ist. In diesem Fall sollte man anstelle der Prozentangaben einfache Quotienten bevorzugen – wie in *Beispiel 2.2*: Die relative Häufigkeit der 1-Personen-Haushalte beträgt 6/40. In jedem Fall ist die Anzahl der Untersuchungseinheiten, d. h. die Größe der Grundgesamtheit, mit anzugeben.

2.1.2 Graphische Darstellung

Für die Darstellung der Häufigkeitsverteilungen $h_1, \ldots, h_M$ bzw. $f_1, \ldots, f_M$ sind Tabellen oder Graphiken üblich. Tabellarische Darstellungen haben wir bereits im vorigen Abschnitt benutzt. Derartige Tabellen sind keine Urlisten. Vielmehr sind sie bereits eine Aggregation bzw. Auswertung von Urlisten.
Da sich Menschen oft leichter von visuellen Eindrücken als von Zahlentabellen überzeugen lassen, werden insbesondere bei Präsentationen in der Praxis häufig graphische Darstellungen des Zahlenmaterials verwendet.
Häufige Darstellungsformen qualitativer oder quantitativ-diskreter Merkmale sind Säulen- und Kreisdiagramme.

Säulendiagramm
Die Ausprägungen $x_1, \ldots, x_M$ des Merkmals X werden auf der horizontalen Merkmalsachse aufgetragen und über den Merkmalen Rechtecke mit Höhen $h_1, \ldots, h_M$ bzw. $f_1, \ldots, f_M$ und identischen Breiten gezeichnet.

Kreisdiagramme eignen sich besonders zur Darstellung von Häufigkeiten qualitativer Merkmale, insbesondere von Häufigkeiten nominalskalierter Merkmale, da bei einem Kreisdiagramm keine Ordnung in den Daten dargestellt werden kann.

Kreisdiagramm
Die Flächen der Kreissektoren im Kreisdiagramm sind proportional zu den Häufigkeiten. Für den Winkel α_j des j-ten Kreissektors gilt: $\alpha_j = f_j \cdot 360^\circ$.

Beispiel 2.4 Säulendiagramm und Kreisdiagramm ■

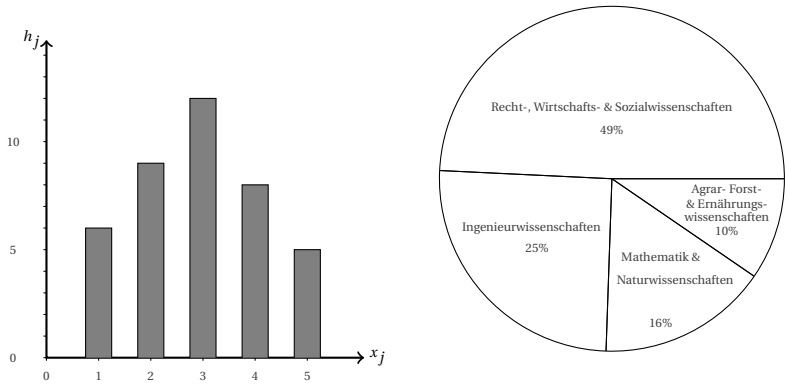


Bild 2.1 Säulendiagramm für die Haushaltsgröße und Kreisdiagramm für das Merkmal „Fächergruppe“ bei neu eingeschriebenen Studierenden

2.2 Häufigkeitsverteilungen bei stetigen Merkmalen

2.2.1 Prinzip der Klassenbildung

In vielen Fällen, insbesondere bei stetigen und quasi-stetigen Merkmalen, ist es oft nicht möglich, die N beobachteten Merkmalswerte der Urliste auf eine deutlich kleinere Menge von x_1 , ... , x_M (unterschiedlichen) Werten zu komprimieren. Häufigkeitsverteilungen, die mit den Methoden des vorhergehenden Abschnitts erstellt werden, haben dann nur eine geringe Aussagekraft, da jede beobachtete Merkmalsausprägung die gleiche Häufigkeit besitzt.

Um eine interpretierbare und überschaubare Häufigkeitsverteilung zu erhalten, fasst man mehrere Merkmalsausprägungen zu einer Klasse zusammen, d.h. die Werte der Urliste werden M verschiedenen Klassen $K_1, K_2, \ldots, K_M$ zugeordnet. Dabei sind unterschiedliche Klassenbreiten erlaubt. Das ursprüngliche Merkmal, wie etwa Zeit, Größe oder Gewicht, wird so zu einem diskreten Merkmal, das nur noch die M verschiedenen „Werte“ K_j annehmen kann. Der Preis, den man mit einer Diskretisierung bzw. Klassenbildung bezahlt, ist ein Informationsverlust, da die Verteilung der Werte innerhalb einer Klasse nicht mehr berücksichtigt wird.

Wir führen folgende Bezeichnungen ein:

M	Anzahl der Klassen,
a_j	untere Klassengrenze der Klasse K_j,
b_j	obere Klassengrenze der Klasse K_j,
$b_j - a_j$	Klassenbreite der Klasse K_j.

Damit lassen sich die absoluten und relativen Häufigkeiten je Klasse wie bei den diskreten Merkmalen berechnen.

Absolute Häufigkeiten

h_j Anzahl der Beobachtungen x_j in der Klasse K_j
$a_j \le x_j < b_j, \quad j = 1, \ldots, M$

Relative Häufigkeiten

$f_j = \dfrac{h_j}{N}$ Anteil der Beobachtungen x_j in der Klasse K_j
$a_j \le x_j < b_j, \quad j = 1, \ldots, M$

Also werden alle Merkmalswerte, die größer oder gleich der Klassenuntergrenze a_j und kleiner als die Klassenobergrenze b_j sind, in der Klasse j gezählt. Die Häufigkeitstabelle enthält somit die Klassennummer K_j, für die wir kürzer einfach j schreiben, die Klassengrenzen a_j und b_j und manchmal auch die Klassenbreite $b_j - a_j$, die absoluten Häufigkeiten h_j und die relativen Häufigkeiten f_j der Klassen. Dieser Aufbau ist in der folgenden Tabelle dargestellt:

Tabelle 2.1 Klassierte Häufigkeitstabelle

Klassennr. j	Klasse j	Klassenbreite	h_j	f_j
1	$[a_1, b_1)$	$b_1 - a_1$	h_1	f_1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
M	$[a_M, b_M)$	$b_M - a_M$	h_M	f_M
$\sum$			N	1

Bemerkung:
Wenn eine Intervallgrenze durch eine runde Klammer angegeben wird, bedeutet dies, dass der Grenzwert nicht im Intervall enthalten ist. Eine eckige Klammer ([oder]) zeigt an, dass der Grenzwert zum Intervall gehört.

Für die Anzahl der Klassen und die Klassenbreite gibt es kaum feste Regeln. Bei sehr vielen schmalen Klassen ist die Darstellung unübersichtlich und die Struktur der Verteilung schwer erkennbar. Dagegen ist eine geringe Anzahl von breiten Klassen mit einem hohen Informationsverlust verbunden und charakteristische Eigenschaften der Verteilung werden eventuell

verdeckt. Am besten ist es, wenn es sachlogische Zusammenhänge gibt, die die Klassengrenzen definieren. Gibt es diese Zusammenhänge nicht, muss man die Klassengrenzen willkürlich setzen. Damit ist die Struktur einer klassierten Häufigkeitsverteilung von dem Ersteller der Statistik abhängig. Dennoch gibt es für die Anfertigung von klassierten Häufigkeitstabellen folgende „Faustregel“:

Regeln für die Anfertigung von klassierten Häufigkeitstabellen

- Die Klassenzahl M richtet sich nach dem Umfang der Grundgesamtheit N. Als Anhaltspunkt gilt: $M \approx \sqrt{\textit{Anzahl der Merkmalswerte } N}$.
- Die Klassenzahl soll mindestens 4 und höchstens 15 betragen.
- Die Klassenbreiten sind so zu wählen, dass keine leeren Klassen auftreten.

Beispiel 2.5 Bedienzeiten

Im Rahmen einer Kundenzufriedenheitsanalyse werden bei 20 Kunden die Bedienzeiten X [min] an einem Postschalter gemessen:

2,30	1,94	0,11	5,70	5,28	2,91	0,89	4,20	0,30	0,23
5,09	7,90	3,47	1,60	0,40	6,20	0,90	4,35	3,21	1,10

Zu den Daten lässt sich exemplarisch die folgende klassierte Häufigkeitstabelle angeben:

Klassennr. j	Klasse j	f_j	h_j
1	0≤ ... <1	6	6/20 = 0,30
2	1≤ ... <2	3	3/20 = 0,15
3	2≤ ... <5	6	6/20 = 0,30
4	5≤ ... <8	5	5/20 = 0,25
$\sum$		20	1,00

25 % der Kunden hatten beispielsweise eine Bedienzeit von 5 bis unter 8 Minuten. Es ist jedoch nicht mehr erkennbar, wie sich die Zeiten innerhalb der Klasse verteilen.

■

2.2.2 Histogramm

Häufigkeiten metrischer Daten werden üblicherweise durch ein Histogramm dargestellt. Das Histogramm geht von klassierten Daten aus, zeichnet für die Klassen Rechtecke, deren Fläche proportional zu der jeweiligen absoluten bzw. relativen Klassenhäufigkeit ist.

Für die Höhe des Rechtecks über der j-ten Klasse gilt dann:

$$h_j^* = \frac{h_j}{b_j - a_j} \qquad \text{bzw.} \qquad f_j^* = \frac{f_j}{b_j - a_j}.$$

Die nachfolgende Abbildung zeigt den Zusammenhang zwischen der absoluten Häufigkeit h_j und der Höhe h_j^* graphisch:

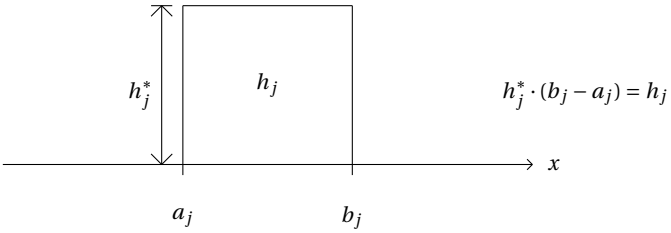


Bild 2.2 Zusammenhang zwischen absoluter Häufigkeit und Höhe

Histogramm
Die Veranschaulichung der Häufigkeiten h_j bzw. f_j der Merkmalsausprägungen x_j der Klassen j, $j = 1, ..., M$ heißt Histogramm. Dazu werden Rechtecke über den Klassen

$$[a_j, b_j), \quad j = 1, ..., M$$

mit einer Höhe proportional zu

$$h_j^* = \frac{h_j}{(b_j - a_j)} \quad \text{bzw.} \quad f_j^* = \frac{f_j}{(b_j - a_j)}$$

gezeichnet. f_j^* wird auch als Häufigkeitsdichte bezeichnet und gibt an, wie dicht die Beobachtungen im entsprechenden Intervall liegen.

Beispiel 2.6 Bedienzeiten (Fortsetzung)

Histogramm für das Merkmal Bedienzeiten X [min] aus *Beispiel 2.5*
Für die Höhe der Rechtecke h_j^* bzw. f_j^* erhalten wir:

Klassennr. j	Klasse j	h_j	f_j	h_j^*	f_j^*
1	$0 \leq .. < 1$	6	0,30	6,0	0,3
2	$1 \leq .. < 2$	3	0,15	3,0	0,15
3	$2 \leq .. < 5$	6	0,30	2,0	0,1
4	$5 \leq .. < 8$	5	0,25	1,67	0,083

Daraus ergibt sich dann das folgende Histogramm:

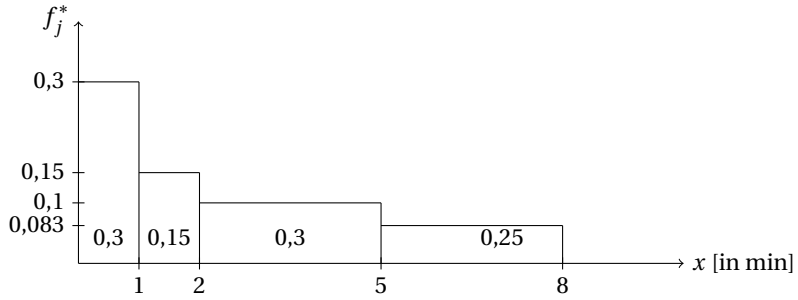


Bild 2.3 Histogramm für das Merkmal Bedienzeiten

■

2.3 Statistische Maßzahlen

In diesem Abschnitt werden Methoden vorgestellt, mit denen sich die charakteristischen Eigenschaften eines einzelnen Merkmals durch aussagekräftige statistische Kennzahlen oder Maßzahlen beschreiben lassen. Man unterscheidet hierbei Lagemaße, Streuungsmaße und Formmaße. Diese Maßzahlen lassen sich auch in geeigneter Form, wie etwa im Box-Plot, visualisieren.

2.3.1 Lagemaße

Verteilungen eines Merkmals geben detaillierte Informationen, welche Merkmalswerte wie häufig in einer Grundgesamtheit anzutreffen sind. Lageparameter hingegen dienen dazu, die Eigenschaften einer Verteilung in komprimierter Form wiederzugeben, indem sie alle Merkmalswerte auf einen einzigen repräsentativen Wert reduzieren, der stellvertretend für alle Merkmalswerte steht. Insbesondere sind Lagemaße beim Vergleich mehrerer Grundgesamtheiten beliebt.

Beispiel 2.7

Familie A möchte bei ihrem nächsten Sommerurlaub im Juni unbedingt am Meer baden. Sie erfährt, dass im Juni die durchschnittliche maximale Tagestemperatur in Palermo bei +25 °C und in Sylt bei +15 °C liegt. Ohne die Verteilung der Temperaturen in den beiden Orten näher zu kennen, fällt die Entscheidung leicht: Familie A fährt nach Palermo. ■

Modus

Dieses Lagemaß gibt an, welche Merkmalsausprägung am Häufigsten vorkommt. Bei einer stetigen oder klassifizierten Variablen ist der Modus die Klasse, in der die Werte am dichtesten liegen, also die Dichte den größten Wert annimmt.

Modus
X ist ein diskretes Merkmal:

$$x_{mod} = \text{häufigster Wert des Merkmals } X, \tag{2.1}$$

X ist ein stetiges bzw. klassiertes Merkmal mit der Dichte f_j^*:

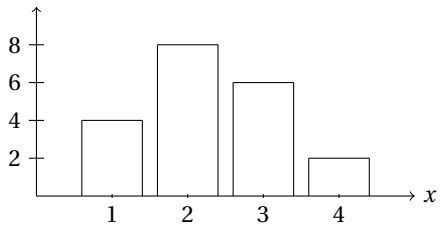
$$x_{mod} = \text{Klasse } K_j \text{ mit größter Häufigkeitsdichte } f_j^*. \tag{2.2}$$

Der Modus ist das wichtigste Lagemaß für kategoriale Merkmale und bereits auf Nominalskalenniveau sinnvoll. In der Darstellung durch ein Säulendiagramme ist der Modus die Merkmalsausprägung mit der höchsten Säule. Bei stetigen Merkmalen ist der Modus die Klasse mit der größten Dichte.

Beispiel 2.8

Eine Kleinstadt beabsichtigt, ihren Marktplatz neu zu gestalten. Dazu wird eine Umfrage unter 20 Marktständen durchgeführt. Unter anderem wird nach der Zahl X der Beschäftigten gefragt. Das Ergebnis ist in der folgenden Tabelle bzw. in dem folgendem Diagramm zusammengefasst.

Beschäftigte x_j	$h(x_j)$
1	4
2	8
3	6
4	2
Gesamt	20



Für das Merkmal X ist $x_{mod} = 2$, d. h. Verkaufsstände mit 2 Beschäftigten kommen unter den 20 befragten Marktständen am häufigsten vor.

■

Beispiel 2.9

Für die Daten des *Beispiels 2.5* liefert die Tabelle bzw. das Histogramm aus *Beispiel 2.6* als Modus die Bedienzeitklasse 0-1 Minute. Die Klasse 2-5 Minuten besitzt zwar eine genauso große relative Häufigkeit, jedoch ist hier die Dichte geringer.

■

Median

Der Median oder Zentralwert ist ein Merkmalswert, der die Grundgesamtheit in zwei Hälften teilt, wobei in der einen Hälfte die Objekte mit den größeren Merkmalswerten, und in der anderen Hälfte die kleineren Merkmalswerte liegen. Um diese Maßzahl zu ermitteln, sind die Beobachtungswerte der Größe nach zu sortieren. Die geordneten Werte werden mit tief gestellten, in eckigen Klammern gesetzten Indizes versehen, sodass gilt:

$$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[N]}$$

Demnach ist $x_{[1]}$ der kleinste Wert der Beobachtungsreihe; $x_{[N]}$ ist der größte Wert. Die sortierten Beobachtungswerte nennt man auch Rangliste. Das dazugehörige Merkmal muss mindestens ordinal skaliert sein, da für nominal skalierte Daten keine sinnvolle Reihenfolge angegeben werden kann.

Beispiel 2.10

Fünf Angestellte einer Firma vergleichen ihr Bruttomonatseinkommen. Folgende Einkommen wurden festgestellt:

Angestellter Nr. i	1	2	3	4	5
Einkommen x_i in €	3 400	3 800	4 100	3 700	3 200

Für die geordnete Reihe der Merkmalswerte ergibt sich:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$
3200	3400	3700	3800	4100

Die Grundgesamtheit umfasst 5 Objekte, d. h. eine ungerade Zahl von Objekten. Eine Aufteilung in zwei gleich große Hälften zu jeweils exakt 50 % ist nicht möglich. Der Merkmalswert 3 700 € mit der Ordnungsnummer 3, der in der „Mitte“ steht, kommt der Idee des Medians am nächsten. Wir stellen fest, dass 60 % der Merkmalswerte kleiner oder gleich $x_{[3]} = 3\,700$ € und 60 % der Merkmalswerte größer oder gleich $x_{[3]} = 3\,700$ € sind.

■

Beispiel 2.11

Zu den fünf Angestellten aus *Beispiel 2.10* kommt ein weiterer Angestellter hinzu. Er verdient 3 300 €. Die Grundgesamtheit umfasst jetzt 6 Objekte, d. h. eine gerade Zahl von Objekten. Für die geordnete Reihe der Merkmalswerte erhält man:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[5]}$
3 200	3 300	3 400	3 700	3 800	4 100

Zwar ist eine Aufteilung in zwei gleich große Hälften zu jeweils exakt 50 % möglich, jedoch gibt es diesmal keinen Merkmalswert, der genau in der „Mitte“ steht. Für jeden Wert x aus dem Intervall [3 400, 3 700], d. h. für jedes x mit $x_{[3]} \leq x \leq x_{[4]}$ gilt, dass 50 % der Merkmalswerte kleiner oder gleich x und 50 % der Merkmalswerte größer oder gleich x sind. Dies zeigt, dass der Median in bestimmten Fällen nicht eindeutig ist. Es ist üblich, in solchen Fällen die Mitte des Intervalls als Median zu nehmen. ■

Beispiel 2.11 zeigt, dass man eine Grundgesamtheit nicht immer in zwei gleich große Hälften mit genau 50 % kleineren Werten und genau 50 % größeren Werten aufteilen kann. Dies liegt daran, dass man die Mitte selbst einer Seite zuordnen muss. Wenn man die Mitte zweimal vergibt, also beiden Seiten zuordnet, entstehen zwei „Hälften“, die jeweils einen Anteil von mindestens 50 % besitzen.

Median
Der Median x_{med} ist dadurch charakterisiert, dass mindestens 50 % der Merkmalsausprägungen einen Wert kleiner oder gleich dem Wert x_{med} und mindestens 50 % einen Wert größer oder gleich dem Wert x_{med} annehmen.

Die Berechnungsmethoden, die wir in den vorigen Beispielen benutzt haben, können wir wie folgt notieren:

Berechnung des Median bei vorliegender Urliste
Sortiere die Urliste nach aufsteigenden Merkmalswerten:

$x_{[1]} \leq x_{[2]} \leq ... \leq x_{[N]}$

Dann erhält man:

$$x_{med} = \begin{cases} x_{[\frac{N+1}{2}]} & \text{für } N \text{ ungerade} \\ \frac{1}{2} \cdot (x_{[\frac{N}{2}]} + x_{[\frac{N}{2}+1]}) & \text{für } N \text{ gerade} \end{cases} \tag{2.3}$$

Bemerkung:
Falls das betrachtete Merkmal nur ordinal skaliert ist, so ist bei der Berechnung des Median x_{med} gemäß *Formel (2.3)* zu beachten, dass die Mittelung von $x_{[\frac{N}{2}]}$ und $x_{[\frac{N}{2}+1]}$ für den Fall N gerade nicht sinnvoll ist, es sei denn $x_{[\frac{N}{2}]}$ und $x_{[\frac{N}{2}+1]}$ sind gleich. Im Fall verschiedener Werte erfüllen sowohl $x_{[\frac{N}{2}]}$ als auch $x_{[\frac{N}{2}+1]}$ die Forderung an den Median, sodass dieser nicht mehr eindeutig bestimmt werden kann.

Beispiel 2.12

Wir betrachten die Preise in [€] für eine Tasse Latte in verschiedenen Gaststätten in zwei Städten und bestimmen jeweils den Median nach *Formel (2.3)*:

Stadt 1

2,20	2,30	2,50	2,70

$$x_{med} = \frac{1}{2} \cdot (x_{[2]} + x_{[3]}) = 2,40 €$$

Stadt 2

1,80	1,90	2,00	2,20	2,30

$x_{med} = x_{[3]} = 2,00€$

■

Quantile

Der Median versucht eine Grundgesamtheit möglichst gut in zwei gleich große Hälften zu je 50 % aller Objekte aufzuteilen. Bei einem p-Prozent-Quantil verhält es sich ähnlich, jedoch können diesmal die beiden Teile der Grundgesamtheit auch unterschiedlich groß sein. Sei p eine Zahl zwischen Null und Eins und $x_{[1]} \leq x_{[2]} \leq ... \leq x_{[N]}$ die der Größe nach geordnete Urliste. Wir definieren ähnlich wie beim Median:

Quantile
Das p-Prozent-Quantil x_p ist dadurch charakterisiert, dass mindestens $p \cdot 100\,\%$ der Merkmalsausprägungen einen Wert kleiner oder gleich dem Wert x_p und mindestens $(1-p) \cdot 100\,\%$ einen Wert größer oder gleich dem Wert x_p annehmen.
Bei vorliegender Urliste gilt:

$$x_p = \begin{cases} x_{[<Np>]} & \text{wenn } Np \text{ nicht ganzzahlig} \\ \frac{1}{2} \cdot (x_{[Np]} + x_{[Np+1]}) & \text{für } Np \text{ ganzzahlig} \end{cases} \tag{2.4}$$

Das Symbol $< Np >$ bedeute „nächstgrößere ganze Zahl an Np“.

Spezielle Quantile sind das

- untere oder erste Quartil $x_{0,25}$. Dieses besagt, dass mindestens 25 % der Merkmalswerte kleiner oder gleich $x_{0,25}$ sind, während dementsprechend mindestens 75 % der Werte größer oder gleich $x_{0,25}$ sind.
- obere oder dritte Quartil $x_{0,75}$. Dieses besagt, dass mindestens 75 % der Merkmalswerte kleiner oder gleich $x_{0,75}$ sind, während dementsprechend mindestens 25 % der Werte größer oder gleich $x_{0,75}$ sind.
- mittlere oder zweites Quartil $x_{0,5}$. Dieses entpricht dem Median x_{med}.

Von Dezilen spricht man, falls $p = 0,1, 0,2, ... , 0,9$; von Perzentilen bei 2-stelligen Nachkommastellen $p = 0,01, 0,02, ... , 0,99$. Die Angabe eines Perzentils kann sehr hilfreich sein, um einen Messwert größenmäßig einzuordnen. So werden etwa in der Kinderheilkunde die individuellen Werte eines Kindes bezüglich Größe, Gewicht und Kopfumfang mit den altersgemäßen 3 %- und 97 %-Perzentilen verglichen, um zu beurteilen, ob es Auffälligkeiten in der Entwicklung gibt.

Beispiel 2.13

In einem Statistikseminar wurden die männlichen Studierenden gebeten ihre Körpergröße [in cm] anzugeben:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$
165	172	175	176	177	178	178	179	179

$x_{[10]}$	$x_{[11]}$	$x_{[12]}$	$x_{[13]}$	$x_{[14]}$	$x_{[15]}$	$x_{[16]}$	$x_{[17]}$	$x_{[18]}$
180	180	182	183	185	186	186	193	196

Wir bestimmen mithilfe obiger Rangliste einige Quantile bezüglich der Körpergröße nach *Formel (2.4)*:

1. Quartil: $Np = 0{,}25 \cdot 18 = 4{,}5$
$x_{0{,}25} = x_{[<4.5>]} = x_{[5]} = 177$ cm
3. Quartil: $Np = 0{,}75 \cdot 18 = 13{,}5$
$x_{0{,}75} = x_{[<13{,}5>]} = x_{[14]} = 185$ cm
9. Dezil: $Np = 0{,}9 \cdot 18 = 16{,}2$
$x_{0{,}9} = x_{[<16{,}2>]} = x_{[17]} = 193$ cm

Daraus folgt, dass ein 175 cm großer Student bezüglich seiner Körpergröße im unteren Viertel liegt, während ein 196 cm großer Student den oberen 10 % angehört.

■

Arithmetisches Mittel

Das am häufigsten benutzte Lagemaß der Verteilung eines quantitativen Merkmals ist das arithmetische Mittel, das umgangssprachlich oft einfach als Mittelwert bezeichnet wird. Zu seiner Berechnung werden alle Merkmalswerte addiert und deren Summe, der gesamte Merkmalsbetrag, durch die Zahl der Merkmalswerte N dividiert. Jeder Wert x_i geht mit dem gleichen Gewicht $1/N$ in die Berechnung ein. Das arithmetische Mittel darf nur dann berechnet werden, wenn die Summe bzw. die Differenz zwischen zwei Ausprägungen definiert ist. Dies setzt quantitative Merkmale voraus.

Arithmetisches Mittel
Wir bezeichnen mit

N	die Anzahl der Elemente der Grundgesamtheit bzw. die Zahl der Merkmalswerte
x_i	die Merkmalsausprägung des i-ten Elements, $i = 1, \ldots, N$

dann ist

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2.5}$$