Statistik im Klartext

Für Psychologen, Wirtschafts- und Sozialwissenschaftler

2., aktualisierte und erweiterte Auflage

Fabian Heimsch
Rudolf Niederer
Peter Zöfel

Pearson

EXTRAS ONLINE

Statistik im Klartext

Für Psychologen, Wirtschafts- und Sozialwissenschaftler

Statistik im Klartext

Für Psychologen, Wirtschafts- und Sozialwissenschaftler

2., aktualisierte und erweiterte Auflage

Fabian Heimsch
Rudolf Niederer
Peter Zöfel

Pearson

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.

10 9 8 7 6 5 4 3 2 1

22 21 20 19 18

ISBN 978-3-86894-325-2 (Buch)
ISBN 978-3-86326-812-1 (E-Book)

Programmleitung: Kathrin Mönch, kmoench@pearson.de
Korrektorat: le-tex publishing services GmbH, Leipzig
Herstellung: Philipp Burkart, pburkart@pearson.de
Coverabbildung: ammentorp, 123RF
Satz: le-tex publishing services GmbH, Leipzig
Druck und Verarbeitung: Drukarnia Dimograf, Bielsko-Biała

Printed in Poland

Inhaltsverzeichnis

Vorwort

Das vorliegende Buch ist eine Neuauflage von Peter Zöfels Werk „Statistik für Psychologen im Klartext“. Das Werk des leider verstorbenen Autoren wird im gesamten deutschsprachigen Raum an Universitäten, Fachhochschulen und höheren Fachschulen im Statistikunterricht eingesetzt. In diversen Statistikvorlesungen der unterschiedlichsten Studiengänge wie Psychologie, Wirtschaft, Mathematik, Erziehungs- und Sportwissenschaften sowie Informatik wird das Buch als Lehrmittel benutzt und findet sowohl bei Studierenden als auch Dozierenden großen Anklang. Der abgeänderte Titel – „Statistik im Klartext“ – ist an dieser Stelle somit nicht zu viel versprochen und soll gerade dem Umstand Rechnung tragen, dass eben sehr unterschiedliche Zielgruppen von der Benutzung des Buches profitieren können!

Aus genannten Gründen der Beliebtheit des Buches haben wir uns gerne für die Betreuung einer Neuauflage entschieden. Im Rahmen dieser Überarbeitung haben wir uns stets am Grundgedanken vom Autor Peter Zöfel orientiert, um den Inhalt eines Statistikbuches in einer Form darzubieten, dass auch mathematisch weniger Geübte einen Zugang finden. An dieser Stelle wollen wir als Neuaufleger unsere langjährige Unterrichtserfahrung einbringen und daher weitgehend auf theoretische Herleitungen verzichten. Vielmehr bemühen wir uns, die einzelnen Verfahren anhand einprägender Beispiele zu erläutern und wollen so auch die Tatsache berücksichtigen, dass im Präsenzunterricht immer weniger Zeit zur Verfügung steht und Studierende sich einen Großteil des Stoffes im Selbststudium erarbeiten dürfen. Es ist geplant, in einer weiteren Überarbeitung Übungen mit den Softwarepaketen SPSS und R einzufügen.

Die wenigsten Bücher tragen dem Umstand Rechnung, dass mittlerweile wohl kaum jemand mehr statistische Verfahren per Hand durchrechnet. Mit Hilfe der Statistiksoftware SPSS, eines der weltweit verbreitetsten Programme zur statistischen Datenanalyse, werden im Buch Beispiele zu diversen Verfahren erläutert. Die im Buch verwendeten Beispieldateien sind unter *www.pearson-studium.de* verfügbar und können von dort heruntergeladen werden. Den Abschluss der Kapitel bilden Übungsaufgaben, deren Lösungen im Anhang nachgesehen werden können.

Unser Dank gilt insbesondere Frau Kathrin Mönch vom Pearson Verlag für die Gelegenheit der Neuüberarbeitung dieses Buches und vielen Kolleginnen und Kollegen für wertvolle Anregungen und Kommentare zum Inhalt des vorliegenden Buches.

Zum Schluss geben wir leichtsinnigerweise eine E-Mail-Adresse bekannt für den Fall, dass Sie Fragen haben oder Anmerkungen zum Buch machen wollen.

Olten, im Januar 2018

Prof. Dr. Fabian Heimsch und Prof. Dr. Ruedi Niederer
Statistik-im-Klartext@fhnw.ch

Einführung

1

Dieses Buch soll und kann keine Unterhaltungslektüre sein, die Autoren wollen aber versuchen, das Beste aus dem doch eher trockenen Stoff zu machen und die zuweilen schwierige Materie vor allem durch passende Beispiele näher zu bringen. Das Verständnis soll hierbei in den Vordergrund gerückt werden und mathematische Formeln sollen als Hilfsmittel für das Erlangen dieses Verständnisses verstanden werden.

Der Begriff „Statistik“ wird heute im doppelten Sinn verwendet.

Zum einen versteht man unter Statistik Datensammlungen zu einem bestimmten Thema, zum Beispiel Bevölkerungsstatistiken, Preisstatistiken, Statistiken über Handel, Verkehr, Löhne und Gehälter, im Gesundheitswesen oder die Arbeitslosenstatistik. Neben diesen amtlichen Statistiken gibt es zahlreiche nichtamtliche und private Statistiken von Markt- und Meinungsforschungsinstituten, Wirtschaftsverbänden, Unternehmen, Forschungsinstituten oder auch Sportverbänden.

Zum anderen versteht man unter Statistik eine Wissenschaft, die sich mit der Analyse von Daten befasst, um Fragestellungen zu einem vorgegebenen Thema zu klären. Gerade auf dem Gebiet der Psychologie und Ökonomie sind statistische Methoden weit verbreitet. Im Mittelpunkt steht dabei die Datenanalyse, die man in einen beschreibenden Teil (deskriptive Statistik) und einen analytischen Teil (analytische Statistik) aufteilen kann. Mithilfe der deskriptiven Methoden werden die Daten durch Tabellen und Grafiken und durch die Berechnung bestimmter Kennwerte beschrieben. Mit der analytischen Statistik können ausgehend von einer Stichprobe allgemein gültige Schlüsse gezogen werden. Dazu stehen eine Vielzahl von Methoden (statistischen Testverfahren) zur Verfügung. Die analytische Statistik versucht somit, ausgehend von einer Stichprobe, gültige Aussagen für die Grundgesamtheit (Population) abzuleiten.

Mit diesem Aspekt der Statistik als einer mathematischen Wissenschaft beschäftigt sich dieses Buch, wobei auch stets ein Augenmerk darauf gerichtet wird, dass die zumeist sehr rechenintensiven statistischen Verfahren kaum noch per Hand, sondern fast ausnahmslos mit entsprechenden Computerprogrammen gerechnet werden.

Eine Übersicht über die gängigsten Computerprogramme enthält in alphabetischer Reihenfolge Tabelle 1.1.

Programm	Open-Source?
PSPP	ja
R	ja
SAS	nein
IBM-SPSS	nein
STATA	nein

Tabelle 1.1: Statistikprogramme

Was die Handhabung der Programme anbelangt, so gibt es zwei prinzipielle Möglichkeiten. Modern unter Windows und komfortabel für den Anwender ist die menügeführte Handhabung, bei der die einzelnen statistischen Analysen über entsprechend gestaltete Dialogboxen angefordert werden. Gewisse Vorteile bietet es aber auch, wenn zu diesem Zweck eine Kommandosprache zur Verfügung steht. Ideal ist eine Kombination dieser beiden Möglichkeiten, wie unter anderem von dem lizenzierten Statistikprogramm IBM-SPSS – nachgehend als SPSS bezeichnet – geboten wird, aber auch von der frei erhältlichen Open-Source Version PSPP.

Das „mächtigste“ Programm ist wohl SAS, das ebenso wie SPSS modular aufgebaut ist. SPSS ist aufgrund seiner komfortablen Handhabung eines der am meist verwendeten Programme. Angehörigen von Hochschulen sei empfohlen, sich mit dem jeweiligen Rechenzentrum in Verbindung zu setzen. Zumindest die Programme der beiden Marktführer SPSS und SAS dürften dort entweder an für Hochschulangehörige zugänglichen PCs installiert oder über günstige Endbenutzerlizenzen erhältlich sein. Weiter ist das Softwarepaket R zu nennen. Es ist open-source und findet vor allem an Hochschulen zu Forschungszwecken eine große Verbreitung. R kann zum Beispiel über folgende Adresse bezogen werden: *https://cran.r-project.org/*

An einigen Stellen im Buch wird auf das Programm SPSS verwiesen, das heißt, es wird die Lösungsmöglichkeit in SPSS beschrieben. In diesem Zusammenhang werden einige Dateien zur Verfügung gestellt, die als SPSS-Speicherdateien (Kennung .sav) vorliegen. Eine Übersicht bietet Tabelle 1.2. Diese Dateien können bei Bedarf aus dem Internet unter der Adresse *www.pearson-studium.de* heruntergeladen werden. Es bleibt zu erwähnen, dass SPSS fast alle Dateitypen einlesen kann. Dies betrifft Text-Dateien (.txt), Excel-Dateien (.xls, .xlsx, .xlsm) oder Datendateien von den in Tabelle 1.1 aufgelisteten Programmen.

SPSS-Datei	Kapitel
arbeit.sav	15
durchstr.sav	13
ee.sav	8
fkv.sav	14
gewicht.sav	7
hemmung.sav	13
iq.sav	4
jugend.sav	14
kenia.sav	14
stadt.sav	5
tpf.sav	15
welt.sav	11
ziel.sav	15

Tabelle 1.2: Beispieldateien

Die SPSS-Dateien können nur mit dem Programmsystem SPSS geöffnet werden.

Nach diesem Exkurs über Statistikprogramme wollen wir uns wieder der eigentlichen Statistik zuwenden. Eine statistische Untersuchung lässt sich in fünf Abschnitte einteilen:

1. Planung der Untersuchung
2. Datenerhebung
3. beschreibende Statistik
4. analytische Statistik
5. Interpretation und Präsentation der Ergebnisse

Diese Schritte werden anhand eines Beispiels in Kapitel 13 ausführlicher erläutert.

Deskriptive Statistik

2

ÜBERBLICK

LERNZIELE

- Begriff der Variablen und das Messen von Variablenwerten
- Skalenniveaus
- Häufigkeitstabellen
- Mittelwert
- Median
- Varianz, Standardabweichung und Standardfehler
- Quartile und Interquartilsabstand
- Grafiken

Deskriptive Statistik ist im Gegensatz zur analytischen Statistik die reine Beschreibung der Daten durch Häufigkeitstabellen, passende Kennwerte oder Grafiken. Zunächst aber sei der Begriff der *Variablen* und das *Messen* von Variablen erläutert. Ferner werden vier verschiedene *Skalenniveaus* von Variablen vorgestellt.

Statistische Analysen können unter Zugrundelegung der verschiedensten Variablen vorgenommen werden. Da gibt es auf der einen Seite die quantitativen Variablen mit stetigen Messwerten wie zum Beispiel Körpergröße oder Körpergewicht, welche im Prinzip beliebig genau gemessen werden können, und auf der anderen Seite qualitative Variablen wie zum Beispiel Schulnoten oder die Kodierung eines Merkmals wie den Familienstand in vier Kategorien. Diese qualitativen Variablen können nur diskrete Werte annehmen.

Eine genauere Einteilung der Variablen als die in qualitativ – quantitativ oder diskret – stetig ist diejenige nach vier verschiedenen Skalenniveaus (auch Messniveaus genannt). Bevor auf diese grundlegend wichtige Einteilung ausführlich eingegangen wird, soll zunächst der Begriff des Messens erläutert werden.

2.1 Das Messen

Der Begriff des „Messens“ und die verschiedenen Skalenniveaus sollen anhand einer Studie zu Rauchgewohnheiten erläutert werden. Dabei wurden unter anderem die folgenden Angaben abgefragt:

- Geschlecht
- Alter
- Familienstand
- Schulbildung
- Beruf
- Körpergewicht
- Rauchgewohnheit

Die Zuordnung der aktuellen Variablenwerte bei den einzelnen Fällen (hier: befragte Personen) erfolgt mit einem Vorgang, den man „Messen“ nennt.

Betrachtet man etwa die Variable „Körpergewicht“, so ist klar, wie diese zu messen ist: Man benutzt eine Waage, wobei in der Regel eine Messgenauigkeit von 1 kg ausreichend ist.

Etwas anders liegt der Fall bei der Variablen „Alter“. Dieses misst man nicht mithilfe einer technischen Apparatur; man muss es erfragen oder aus der Geburtsurkunde oder dem Personalausweis erschließen. Trotzdem kann man auch hier von „Messen“ reden, wenn man die Definition des Messens wie folgt fasst:

Das Messen einer Variablen ist die Zuordnung von Zahlen zu den einzelnen Fällen.

Mit dieser Definition lassen sich auch Variablen wie das Geschlecht, der Familienstand oder die Rauchgewohnheit „messen“. Beim Geschlecht ordnet man zum Beispiel den Männern die Zahl 1 und den Frauen die Zahl 2 zu; beim Familienstand vergibt man für die gegebenen vier Kategorien die Zahlen 1 bis 4. Ebenso verfährt man bei der Rauchgewohnheit:

Geschlecht:	1 = männlich
	2 = weiblich

Familienstand:	1 = ledig
	2 = verheiratet
	3 = verwitwet
	4 = geschieden

Rauchgewohnheit:	1 = Nichtraucher
	2 = mäßig
	3 = stark
	4 = sehr stark

Bei diesen Variablen erfolgt das „Messen“ per Augenschein (Geschlecht) oder durch eine entsprechende Befragung. Die Zuordnung („Kodierung“) von Zahlen zu solchen „kategorialen“ Variablen ist spätestens dann notwendig, wenn die statistische Analyse nicht per Hand, sondern unter Einsatz eines entsprechenden Statistik-Programmsystems mithilfe eines Computers erfolgen soll.

2.2 Skalenniveaus

Von entscheidender Bedeutung für die Auswahl eines korrekten statistischen Verfahrens ist die Feststellung des so genannten *Skalenniveaus* (auch *Messniveaus*) der beteiligten Variablen. So macht es beispielsweise sehr wohl Sinn, ein Durchschnittsalter von befragten Personen zu berechnen, jedoch sicherlich nicht das durchschnittliche Geschlecht. Man unterscheidet das Nominal-, Ordinal-, Intervall- und Verhältnisniveau. Dabei werden diese Skalenniveaus gemäß Tabelle 2.1 unterschieden.

In den folgenden Kapiteln werden die einzelnen Skalenniveaus näher erläutert.

Skalenniveau	Unterscheidbar?	Ränge bildbar?	Differenzen interpretierbar?	Verhältnisse interpretierbar?
Nominal	ja	nein	nein	nein
Ordinal	ja	ja	nein	nein
Intervall	ja	ja	ja	nein
Verhältnis	ja	ja	ja	ja

Tabelle 2.1: Skalenniveaus

2.2.1 Nominalskala

Betrachten wir zunächst das Geschlecht, so stellen wir fest, dass die Zuordnung der beiden Ziffern 1 und 2 willkürlich ist; man hätte sie auch anders herum oder mit anderen Ziffern vornehmen können.

Keinesfalls soll schließlich damit ausgedrückt werden, dass Frauen nach den Männern einzustufen sind; auch soll andererseits nicht suggeriert werden, dass Frauen mehr wert seien als Männer.

Eine nominalskalierte Variable ist auch der Familienstand; auch hier hat die Zuordnung der Ziffern zu den Kategorien des Familienstands keinerlei empirische Relevanz. Im Gegensatz zum Geschlecht ist die Variable aber nicht dichotom; sie beinhaltet vier statt zwei Kategorien.

Bei einer nominalskalierten Variablen bilden die Antwortmöglichkeiten eine Liste, deren Reihenfolge frei gewählt werden kann. Ein typisches Beispiel ist die Angabe des Berufs. Hier könnte etwa folgende Kodierung gewählt werden, die sich beim besten Willen nicht in eine sinnvolle Reihenfolge bringen lässt:

1 = Angestellter
2 = Beamter
3 = Arbeiter
4 = Selbstständiger
5 = Hausfrau
6 = Auszubildender
7 = Rentner

Nominalskalierte Variablen sind in ihrer Auswertungsmöglichkeit sehr eingeschränkt. Genau genommen können sie nur einer Häufigkeitsauszählung unterzogen werden. Die Berechnung etwa eines Mittelwerts ist sinnlos.

Eine gewisse Ausnahme bilden dichotome nominalskalierte Variablen. Das sind Variablen, welche nur zwei Ausprägungen haben. Dichotome Skalierungen sind häufig von der Art

1 = ja	1 = richtig	1 = trifft zu	1 = stimme zu
2 = nein	2 = falsch	2 = trifft nicht zu	2 = stimme nicht zu

Bei dichotomen nominalskalierten Variablen kann man stets von einer gegebenen Ordnungsrelation sprechen. So bedeutet etwa im Fall des letzten Beispiels eine niedrige Kodierung Zustimmung, eine hohe Kodierung Ablehnung.

2.2.2 Ordinalskala

Betrachten wir etwa die Rauchgewohnheit, so kommt den vergebenen Kodezahlen insofern eine empirische Bedeutung zu, als sie eine Ordnungsrelation wiedergeben. Die Variable Rauchgewohnheit ist schließlich nach ihrer Wertigkeit aufsteigend geordnet: Ein mäßiger Raucher raucht mehr als ein Nichtraucher, ein starker Raucher mehr als ein mäßiger Raucher und ein sehr starker Raucher mehr als ein starker Raucher. Solche Variablen, bei denen den verwendeten Kodezahlen eine empirische Bedeutung hinsichtlich ihrer Ordnung zukommt, nennt man ordinalskaliert.

Die empirische Relevanz dieser Kodierung bezieht sich aber nicht auf die Differenz zweier Kodezahlen. So ist zwar die Differenz zweier Kodezahlen zwischen einem Nichtraucher und einem mäßigen Raucher einerseits und zwischen einem mäßigen Raucher und einem starken Raucher andererseits jeweils 1. Man wird aber nicht sagen können, dass der tatsächliche Unterschied zwischen einem Nichtraucher und einem mäßigen Raucher einerseits und einem mäßigen Raucher und einem starken Raucher andererseits gleich ist; dafür sind die Begriffe zu vage. Die semantische, d. h. die bedeutungsmäßige Differenz, ist aber kaum gleich groß.

Ein weiteres Beispiel einer ordinalskalierten Variablen ist die Schulbildung, wenn sie etwa in der folgenden Kodierung vorliegt:

1 = Hauptschule
2 = Berufsschule
3 = Mittlere Reife
4 = Abitur
5 = Hochschule

Ein typisches Beispiel einer ordinalskalierten Variablen ist die Vorgabe einer Altersklasseneinteilung in einem Fragebogen:

1 = bis 30 Jahre
2 = 31 bis 50 Jahre
3 = über 50 Jahre

Ein solches Vorgehen ist eigentlich nicht empfehlenswert. Da jeder sein eigenes Alter sicherlich ohne Mühe exakt (in Jahren) angeben kann, sollte man dies auch so erfassen. Werden hingegen Variablen wie das Einkommen von Personen erfragt, so erhöht sich allenfalls die Bereitschaft der Befragten zu antworten, wenn das Einkommen klassiert erfragt wird.

Bei allen bisher genannten Beispielen liegt die ordinale Skalierung unmittelbar auf der Hand. In vielen anderen Fällen kann man eine solche nach etwas Nachdenken erkennen bzw. durch geschickte Kodierung erreichen.

2.2.3 Intervallskala

Bei Variablen, deren entsprechende Werte nicht nur bezüglich der Rangordnung eine Bedeutung haben, sondern auch bezüglich der Differenz (nicht aber des Verhältnisses), spricht man von intervallskalierten Variablen. Dies betrifft Variablen, welche keinen natürlichen Nullpunkt haben und/oder negative Werte aufweisen können. Ein gutes Beispiel an dieser Stelle liefert die Grad-Celsius-Skala, mit welcher Temperaturen gemessen werden. Der Nullpunkt dieser Skala wurde als der Gefrierpunkt von Wasser definiert, der Siedepunkt mit 100 °C festgesetzt. Dieser Nullpunkt macht jedoch keinen physikalischen Sinn, was daraus deutlich wird, dass auch negative Werte auftreten können. 30 °C sind somit nicht dreimal so warm wie 10 °C, sehr wohl aber um 20 °C wärmer als die erwähnten 10 °C.

Man sieht also, dass es hier keinen Sinn macht, Verhältnisse zu berechnen, sehr wohl jedoch aber Differenzen. Solche Variablen, bei denen der Differenz der Werte eine Bedeutung zukommt, sind intervallskaliert. Ihre Bearbeitung unterliegt somit auch fast keinen Einschränkungen; so ist zum Beispiel der Mittelwert ein sinnvoller statistischer Kennwert zur Beschreibung dieser Variablen. Für das geometrische Mittel trifft dies jedoch nicht zu, da die einzelnen Werte zur Berechnung multipliziert werden müssen, wie im Laufe dieses Kapitels noch klar werden wird.

2.2.4 Verhältnisniveau

Bei allen diesen Variablen kommt nicht nur der Differenz zweier Werte, sondern auch dem Verhältnis zweier Werte empirische Bedeutung zu. Ist etwa Emil 20 Jahre und Fritz 40 Jahre alt, so wird man sagen können, dass Fritz doppelt so alt ist wie Emil. Solche Variablen nennt man verhältnisskaliert. Es sind dies alle intervallskalierten Variablen, die den Wert Null annehmen können, wobei dieser gleichzeitig der niedrigste denkbare Wert ist. Beispiele, bei denen dies nicht der Fall ist, sind etwa die in Grad Celsius gemessene Temperatur (wegen der möglichen Werte kleiner als Null) und der Intelligenzquotient (wegen des nicht möglichen Werts von Null). Bei den in diesem Buch behandelten statistischen Verfahren kommt der Unterscheidung zwischen intervall- und verhältnisskalierten Variablen in der Regel keine Bedeutung zu; es gibt nämlich mit einer Ausnahme (geometrisches Mittel) darunter keine Verfahren, die Verhältnisniveau voraussetzen.

Abschließend ist zu erwähnen, dass nominal- und ordinalskalierte Variablen kategoriale Daten, intervall- und verhältnisskalierte Daten Messwerte darstellen.

Zusammenfassend teilt sich die empirische Relevanz der einzelnen Skalenniveaus in vier Dimensionen ein: Unterscheidbarkeit (sind die einzelnen Beobachtungen unterscheidbar?), ob Ränge bildbar sind, ob Differenzen und schließlich ob Verhältnisse interpretierbar sind. Es ist sehr wichtig zu erwähnen, dass der Einsatz möglicher statistischer Verfahren direkt vom Messniveau abhängt.

2.3 Häufigkeitstabellen

Als einfachstes statistisches Verfahren gilt das Zählen. Im Falle von nominalskalierten Variablen ist dies auch die einzig mögliche statistische Operation.

In einer Bevölkerungsumfrage wurde unter anderem nach dem Familienstand der interviewten Personen gefragt. Die Auszählung ergab die Häufigkeiten der Tabelle 2.2.

Familienstand	Häufigkeit
ledig	777
verheiratet	1761
verwitwet	373
geschieden	141

Tabelle 2.2: Beobachtete Häufigkeiten

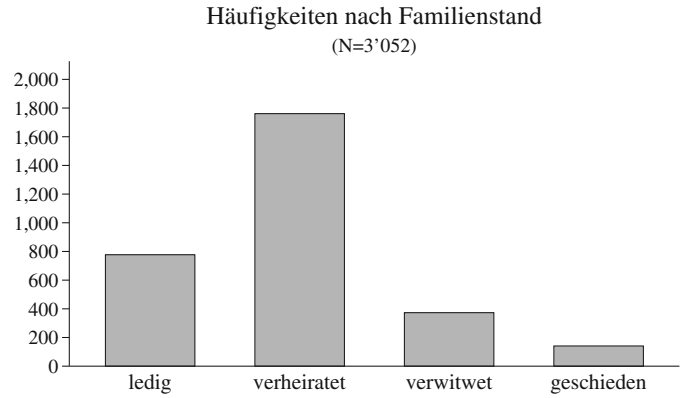

Abbildung 2.1: Balkendiagramm zur Darstellung der Häufigkeit des Familienstandes

2.3.1 Beobachtete und prozentuale Häufigkeiten

Bei nominalskalierten Variablen ist es sinnvoll, die komplette Häufigkeitstabelle anzugeben und zusätzlich zu den *beobachteten* Häufigkeiten die *prozentualen* Häufigkeiten anzugeben.

Bezeichnet man die Anzahl der Kategorien mit k und die beobachteten Häufigkeiten innerhalb von Kategorie j mit f_j, so ist die Gesamtsumme der Häufigkeiten

$$n = \sum_{j=1}^{k} f_j$$

Daraus berechnen sich die prozentualen Häufigkeiten zu

$$p_j = \frac{f_j}{n} \cdot 100 \quad j = 1, \ldots, k$$

Diese prozentualen Häufigkeiten sind in Tabelle 2.3 mit eingetragen.

Familienstand	Häufigkeit	Prozent
ledig	777	25,5 %
verheiratet	1761	57,7 %
verwitwet	373	12,2 %
geschieden	141	4,6 %
Summe	3052	100 %

Tabelle 2.3: Prozentuale Häufigkeiten

Klasse j	Kirchgang	Häufigkeit f_j	Prozent p_j	kumulierte Häufigkeit F_j	kumulierte Prozente P_j
1	mindestens zweimal pro Woche	73	2,6 %	73	2,6 %
2	einmal pro Woche	360	13,0 %	433	15,7 %
3	ein- bis dreimal pro Monat	331	12,0 %	764	27,2 %
4	mehrmals im Jahr	660	23,9 %	1424	51,6 %
5	seltener	935	33,9 %	2359	85,5 %
6	nie	402	14,6 %	2761	100,0 %

Tabelle 2.4: Kumulierte Häufigkeiten

2.3.2 Kumulierte Häufigkeiten

Bei ordinalskalierten Variablen empfiehlt sich die Angabe der beobachteten und prozentualen Häufigkeiten. Sinnvoll ist dann ebenfalls die Bestimmung der kumulierten Häufigkeiten F_j und der kumulierten prozentualen Häufigkeiten P_j. Erstere sind dabei die bis zur betreffenden Kategorie aufsummierten beobachteten Häufigkeiten, die dann wieder auf der Basis der Gesamtsumme der Häufigkeiten in Prozenten ausgedrückt werden können.

In derselben Bevölkerungsumfrage wurde auch die Frage gestellt „Wie oft gehen Sie in die Kirche?“. Alle anfallenden Häufigkeiten sind in Tabelle 2.4 zusammengestellt.

Den kumulierten prozentualen Häufigkeiten kann man zum Beispiel entnehmen, dass über die Hälfte der Befragten, nämlich 51,6 %, zumindest mehrmals im Jahr in die Kirche gehen.

Alter	Anzahl n	Alter	Anzahl n	Alter	Anzahl n	Alter	Anzahl n
18	60	36	49	54	33	72	31
19	48	37	46	55	42	73	40
20	49	38	61	56	51	74	26
21	64	39	52	57	31	75	30
22	76	40	55	58	47	76	27
23	55	41	43	59	42	77	24
24	85	42	35	60	56	78	22
25	88	43	38	61	39	79	21
26	72	44	42	62	44	80	18
27	71	45	41	63	40	81	13
28	70	46	52	64	45	82	14
29	59	47	47	65	45	83	13
30	58	48	51	66	56	84	5
31	55	49	55	67	38	85	8
32	56	50	49	68	47	86	3
33	58	51	49	69	27	87	3
34	60	52	56	70	27	89	3
35	56	53	39	71	39	92	2

Tabelle 2.5: Beobachtete Häufigkeiten von Altersangaben

2.3.3 Klassenbildung

Bei intervall- oder verhältnisskalierten Variablen liegen meist viele verschiedene Werte vor, so dass eine Häufigkeitstabelle recht unübersichtlich wird. In diesem Fall bietet es sich an, mehrere benachbarte Werte zu *Klassen* zusammenzufassen.

Als Beispiel sei eine Häufigkeitstabelle von Altersangaben betrachtet, die einer Fragebogenaktion entnommen wurde (Tabelle 2.5).

Vor einer Klassenzusammenfassung sind zwei Entscheidungen zu treffen, nämlich über die *Klassenbreite* und über den Beginn der ersten Klasse. Was die Wahl der Klassenbreite anbelangt, so gibt es hierfür keine verbindliche Regel. Eine geringe Klassenbreite bedingt eine große Klassenzahl und Unübersichtlichkeit, eine große Klassenbreite hingegen kann typische Verteilungsformen verwischen.

Man sollte etwa zehn bis zwanzig Klassen wählen, und zwar so, dass in der Mitte alle Klassen besetzt sind. Am linken und rechten Verteilungsrand können nach unten bzw. nach oben *offene Klassen* verwendet werden.

Wir wollen uns im gegebenen Beispiel mit acht Klassen begnügen, wobei die erste und die achte Klasse offene Klassen sind (Tabelle 2.6). Abbildung 2.2 zeigt die Häufigkeiten der Variable Alter über eine Klassenbreite von 10 Jahren. Eine Darstellung der absoluten oder prozentualen Häufigkeiten über die Klassenbreite von intervall- oder verhältnisskalierten Variablen nennt man Histogramm. Auf das Histogramm kommen wir bei den grafischen Auswertungen wieder zu sprechen.

Klasse	Häufigkeit	Prozent	kumulierte Prozente
bis 20 Jahre	157	5,1 %	5,1 %
21 bis 30 Jahre	698	22,9 %	28,0 %
31 bis 40 Jahre	548	18,0 %	46,0 %
41 bis 50 Jahre	453	14,8 %	60,8 %
51 bis 60 Jahre	446	14,6 %	75,4 %
61 bis 70 Jahre	408	13,4 %	88,8 %
71 bis 80 Jahre	278	9,1 %	97,9 %
über 80 Jahre	64	2,1 %	100,0 %

Tabelle 2.6: Klassenhäufigkeiten

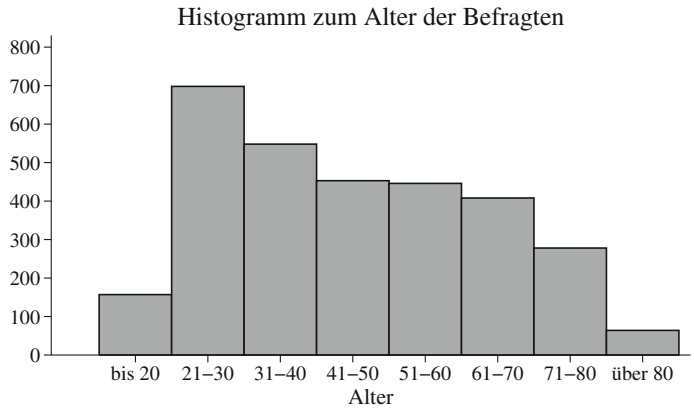

Abbildung 2.2: Histogramm zur Darstellung der Häufigkeit des Alters

Bei der ersten, nach unten offenen Klasse ist zu bedenken, dass sie nur drei Jahrgänge umfasst, so dass von vornherein eine geringere Klassenhäufigkeit zu erwarten ist. Davon abgesehen handelt es sich um eine linksgipflige Verteilung. Dies wird besonders deutlich, wenn man die gegebene Verteilung der Häufigkeiten in Form eines *Histogramms* darstellt. Auf die Bedeutung von linksgipfligen oder rechtsgipfligen Verteilungen kommen wir später zu sprechen.

2.4 Lokalisationsparameter

Lokalisationsparameter beschreiben die Lage einer Verteilung bzw. ihre zentrale Tendenz.

2.4.1 Modus

Der Modus ist der am häufigsten vorkommende Wert. In Tabelle 2.2 zum Familienstand von 3052 befragten Personen wurde die Kategorie „verheiratet“ mit 1761 Nennungen am häufigsten genannt. Der Modus des Familienstandes ist somit 2. Der Modus ist der einfachste Lokalisationsparameter. Er kann bei nominalskalierten Variablen verwendet werden. Seine Aussagekraft ist beschränkt, da in die Berechnung des Modus nur ein Wert einfließt und die anderen unberücksichtigt bleiben.

2.4.2 Der Mittelwert

Der Mittelwert ist der passende Lokalisationsparameter für intervallskalierte und normalverteilte Variablen (die Normalverteilung lernen wir im Kapitel 4 näher kennen). Er ist weniger geeignet für nicht normalverteilte oder ordinalskalierte Variablen und unsinnig für nominalskalierte Variablen.

Wir besprechen drei Varianten des Mittelwertes: das arithmetische, das geometrische und das harmonische Mittel. Am gebräuchlichsten ist das arithmetische Mittel.

Arithmetisches Mittel

Das arithmetische Mittel von n Werten x_i ist die Summe dieser Werte, geteilt durch ihre Anzahl. Das arithmetische Mittel wird umgangssprachlich auch mit Durchschnitt bezeichnet.

Definition des arithmetischen Mittels:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Als Beispiel seien die Altersangaben von $n = 12$ Personen betrachtet.

Wert x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Alter	40	37	67	23	45	39	29	51	56	24	42	38

Die Summe dieser Werte ist 491; damit ergibt sich

$$\overline{x} = \frac{491}{12} = 40{,}917$$

Die Personen sind also im Mittel 40,9 Jahre alt.

Oft liegen Werte in gehäufter Form vor wie zum Beispiel die Noten einer Klassenarbeit in Tabelle 2.7.

Note (x_j)	Klassenhäufigkeit (f_j)	$f_j \cdot x_j$
1	3	3
2	5	10
3	9	27
4	6	24
5	5	25
Summe	28	89

Tabelle 2.7: Noten einer Klassenarbeit

In diesem Fall kann man für die Berechnung des Mittelwerts eine modifizierte Formel anwenden und direkt die Häufigkeiten in den Klassen $1 \ldots k$ benutzen:

$$\overline{x} = \frac{\sum_{j=1}^{k} f_j \cdot x_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} f_j \cdot x_j}{n}$$

Im gegebenen Beispiel ergibt sich als mittlere Note

$$\overline{x} = \frac{89}{28} = 3{,}179$$

Es sei hier erneut erwähnt, dass k die Anzahl Klassen bezeichnet und n die Anzahl vorkommender Werte. Obige Formel mag auf den ersten Blick etwas befremdend erscheinen. Das Vorgehen liefert aber dasselbe Resultat, als hätte man die Daten in Listenform zur Verfügung und man dann den Mittelwert wie gewohnt berechnet.

Wert x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	…	x_{28}
Note	1	1	1	2	2	2	2	2	…	5

Werden nun aus obiger Tabelle alle Noten x_i addiert und durch die Anzahl Werte $n = 28$ geteilt, so erhält man:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{89}{28} = 3{,}179$$

Zwei Eigenschaften des Mittelwerts seien erwähnt:

- Die Summe der Differenzen aller Werte von ihrem Mittelwert ist null.
- Die Summe der Quadrate der Differenzen aller Werte von ihrem Mittelwert ist kleiner als die Summe der Quadrate der Differenzen aller Werte zu irgendeinem anderen Wert.

Bei Abweichungen von der Normalverteilung, was insbesondere bei Auftreten von Ausreißern der Fall ist, ist die Berechnung des Mittelwerts oft nicht sinnvoll, wie das folgende extreme Beispiel zeigt.

Vier Berufstätige wurden nach ihrem monatlichen Einkommen gefragt:

4000 €
7000 €
5000 €
100 000 €

Als mittleres Einkommen ergibt sich

$$\overline{x} = \frac{116\,000\,€}{4} = 29\,000\,€$$

Das mittlere Einkommen der befragten Personen beträgt also 29 000 Euro. Handelt es sich um vier Einwohner eines Dorfes und möchte der Bürgermeister das durchschnittliche Einkommen seiner Bewohner wissen, da sich hiernach als konstanter Prozentsatz die Steuereinnahmen der Gemeinde berechnen, ist dieser Wert sinnvoll. Soll er aber als Maß für den „typischen Fall" eines Einkommens dienen, so wäre er eine sinnlose Größe und der Median wäre vorzuziehen. Den Median lernen wir im Verlaufe dieses Kapitels kennen.

Zuweilen tritt das Problem auf, dass bei verschiedenen Stichproben gewonnene Mittelwerte zu einem gemeinsamen Mittelwert zusammengeführt werden sollen. Angenommen, bei einer zweiten Stichprobe von nunmehr 20 Personen habe sich ein Altersmittelwert von 43,2 Jahren ergeben.

Zur Berechnung des gemeinsamen Mittelwerts mit unserer ersten Stichprobe ($\overline{x} = 40{,}9$; $n = 12$) wäre es falsch, die beiden Mittelwerte zu addieren und die Summe durch 2 zu teilen, da dann die unterschiedlichen Fallzahlen nicht berücksichtigt würden.

Richtig ist es, bei der Berechnung des gemeinsamen Mittelwerts zweier Stichproben, von denen die Mittelwerte $\overline{x}_1$ und $\overline{x}_2$ bei den Fallzahlen n_1 und n_2 vorliegen, diese Mittelwerte entsprechend zu gewichten:

$$\overline{x} = \frac{n_1 \cdot \overline{x}_1 + n_2 \cdot \overline{x}_2}{n_1 + n_2}$$

Im gegebenen Beispiel ergibt sich hiermit

$$\overline{x} = \frac{12 \cdot 40{,}9 + 20 \cdot 43{,}2}{12 + 20} = 42{,}3$$

Bei mehr als zwei zu vereinigenden Mittelwerten wird entsprechend verfahren.

Es wurde schon darauf hingewiesen, dass die Berechnung des Mittelwerts bei nominalskalierten Variablen unsinnig ist. Haben Sie etwa ein neues Medikament getestet und die auftretenden insgesamt 27 verschiedenen Nebenwirkungen mit einer Codierung von 1 bis 27 versehen, so ist die Aussage „die mittlere Nebenwirkung beträgt 12,4" sinnlos.

Geometrisches Mittel

Das geometrische Mittel dient zum Beispiel zur Ermittlung von Wachstumsraten aufeinander folgender Perioden. Seine Berechnung setzt die Verhältnisskala voraus.

Definition des geometrischen Mittels:

$$\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Tabelle 2.8 enthält die Wachstumsraten des Konsumentenpreis-Indexes (KPI) der Russischen Föderation von 2010 bis 2016.

Jahr	KPI	Wachstumsrate x_i	Inflationsrate
2010	100		
2011	108,44	1,0844	8,44 %
2012	113,94	1,0507	5,07 %
2013	121,64	1,0675	6,75 %
2014	131,15	1,0782	7,82 %
2015	151,53	1,1553	15,53 %
2016	162,19	1,0704	7,04 %

Tabelle 2.8: Wachstumsraten des Konsumentenpreis-Indexes

Der KPI misst den Preis eines durchschnittlichen Warenkorb in der jeweiligen Landeswährung in einem gegebenen Jahr. Über die Wachstumsraten dieses Preises im Vergleich zum Vorjahr lässt sich dann die Inflation ermitteln. Das geometrische Mittel kann hier benützt werden, um die durchschnittliche Inflation über einen Zeitraum zu ermitteln.

Das geometrische Mittel der Wachstumsraten ist

$$\overline{x}_G = \sqrt[6]{1{,}0844 \cdot 1{,}0507 \cdot 1{,}0675 \cdot 1{,}0782 \cdot 1{,}1553 \cdot 1{,}0704} = 1{,}0839$$

Dies ist die konstante Wachstumsrate, die zum gleichen Gesamtwachstum geführt hätte. Die durchschnittliche Inflation hat über den Zeitraum 2011–2016 somit 8,39 % betragen.

Harmonisches Mittel

Das harmonische Mittel wird zum Beispiel bei der Varianzanalyse eingesetzt, wenn ungleiche Zellenumfänge durch das harmonische Mittel ersetzt werden (siehe Abschnitt 13.4).

Definition des harmonischen Mittels:

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Ein praktisches Beispiel ist die Berechnung mittlerer Geschwindigkeiten. Angenommen, Sie fahren die Strecke von Kassel nach Marburg (100 km) mit einer durchschnittlichen Geschwindigkeit von 80 km/h und die Strecke von Marburg nach Frankfurt (ebenfalls 100 km) mit einer solchen von 120 km/h. Falls Sie nun ohne länger nachzudenken die Durchschnittsgeschwindigkeit von Kassel bis Frankfurt mit 100 km/h angeben, liegen Sie falsch.

Hier ist nicht das arithmetische, sondern das harmonische Mittel einzusetzen:

$$\overline{x}_H = \frac{2}{\frac{1}{80} + \frac{1}{120}} = 96$$

Die durchschnittliche Geschwindigkeit für die Gesamtstrecke beträgt also 96 km/h.

2.4.3 Der Median

Der Median wird bei ordinalskalierten bzw. intervallskalierten, aber nicht normalverteilten Variablen berechnet.

Definition des Medians:
Der Median ist derjenige Wert, unterhalb und oberhalb dessen jeweils die Hälfte der Messwerte liegen.

Dabei gibt es zwei verschiedene Arten der Berechnung, je nachdem, ob die Messwerte einzeln oder in klassierter Form vorliegen.

Wir betrachten zunächst ein Beispiel zur erstgenannten Möglichkeit. Bei elf Probanden seien zur Lösung einer Aufgabe die folgenden Zeiten (in Sekunden) gemessen worden:

489 113 141 120 217 109 675 218 96 225 132

Es treten zwei Ausreißerwerte auf (489, 675), so dass es sinnvoll erscheint, anstelle des Mittelwerts den gegenüber Ausreißern unempfindlichen Median zu berechnen.

Zu diesem Zweck schreibt man zunächst die Werte der Größe nach sortiert auf:

96 109 113 120 132 141 217 218 225 489 675

Bei einer solch ungeraden Anzahl von Werten ist der Median ein tatsächlich auftretender Wert, nämlich der mittlere Wert der in aufsteigender Reihenfolge sortierten Wertereihe. Im gegebenen Beispiel mit elf Messwerten ist dies der sechste Wert, also der Wert 141:

$$Median = 141$$

Links und rechts von diesem Wert liegen dann gleich viele Werte, nämlich fünf.

Wir wollen der aufsteigend notierten Wertereihe noch einen Wert anfügen:

96 109 113 120 132 141 217 218 225 489 675 690

In diesem Fall einer geraden Anzahl von Werten ist der Median der Mittelwert aus den beiden mittleren Werten der Liste, hier also

$$Median = \frac{141 + 217}{2} = 179$$

Es ist offensichtlich, dass der Median gänzlich unempfindlich gegen Ausreißerwerte ist. So ist es zum Beispiel völlig gleichgültig, welchen Wert der größte Messwert annimmt, da der Wert des Medians hiervon unberührt bleibt.

Häufig wird der Median auch bei ordinalskalierten Variablen bestimmt, wobei die Angaben in Form einer Häufigkeitstabelle vorliegen. In einem Fragebogen zur Krankheitsverarbeitung sollten 160 Patienten auf einer Fünferskala angeben, inwieweit sie aktive Anstrengungen zur Lösung ihrer gesundheitlichen Probleme unternehmen. Die entsprechenden Häufigkeiten sind in Tabelle 2.9 wiedergegeben.

aktiv	Skalenwert x_j	Häufigkeit f_j	kumulierte Häufigkeit F_j	prozentuale Häufigkeit p_j	kumulierte prozentuale Häufigkeit P_j
gar nicht	1	12	12	7,5 %	7,5 %
wenig	2	25	37	15,6 %	23,1 %
mittelmäßig	3	23	60	14,4 %	37,5 %
ziemlich	4	53	113	33,1 %	70,6 %
sehr stark	5	47	160	29,4 %	100,0 %

Tabelle 2.9: Beobachtete und kumulierte Häufigkeiten

Zusätzlich zu den Häufigkeiten ist jeweils die kumulierte Häufigkeit aufgeführt, sowohl in absoluten Werten als auch in prozentualen Werten.

Nach der erläuterten Regel zur Bestimmung des Medians ergibt sich hierfür der Wert 4. Bei insgesamt 160 Werten liegt der Median nämlich, wenn man die Werte aufsteigend sortiert, zwischen dem 80. und 81. Wert. Die kumulierte Häufigkeit zeigt an, dass sowohl der 80. als auch der 81. Wert den Wert 4 haben, womit auch der Median diesen Wert annimmt.

Es dürfte aber unmittelbar klar sein, dass dies ein recht unbrauchbarer, da zu ungenauer Wert ist. Im Falle von solchen gehäuften respektive klassierten Daten benutzt man zur genaueren Bestimmung des Medians eine verfeinerte Formel.

Berechnung des Medians bei klassierten Daten:

$$\textit{Median} = x_u + \frac{(50\,\% - P_u)}{(P_o - P_u)}(x_o - x_u)$$

Dabei gilt:

x_u, x_o	untere und obere Grenze der Klasse, die den Median enthält
P_u, P_o	untere und obere kumulierte prozentuale Häufigkeiten

Im gegebenen Beispiel sind die folgenden Werte gegeben:

$$x_u = 3{,}5 \quad x_o = 4{,}5 \quad P_u = 37{,}5\,\% \quad P_o = 70{,}6\,\%$$

Damit ergibt sich für den Median

$$\textit{Median} = 3{,}5 + \frac{(50\,\% - 37{,}5\,\%)}{(70{,}6\,\% - 37{,}5\,\%)}(4{,}5 - 3{,}5) = 3{,}877$$

Bezeichnet man die Anzahl der Kategorien mit k, so würde sich der Mittelwert aller Werte nach folgender Formel errechnen:

$$\overline{x} = \frac{\sum_{j=1}^{k} f_j \cdot x_j}{n} = \sum_{j=1}^{k} p_j \cdot x_j$$

Dies ergibt im gegebenen Beispiel

$$\overline{x} = 0{,}075 \cdot 1 + 0{,}156 \cdot 2 + 0{,}144 \cdot 3 + 0{,}331 \cdot 4 + 0{,}294 \cdot 5 = 3{,}613$$

Der Mittelwert ist also kleiner als der Median, was bei einer rechtsgipfligen Verteilung wie im gegebenen Beispiel stets der Fall ist. Bei einer linksgipfligen Verteilung ist der Mittelwert größer als der Median, in diesen Fällen spricht man von einer schiefen Verteilung. Die Schiefe einer Verteilung wird in Abschnitt 2.5.3 eingeführt.

2.4.4 *P*-Quantile

Der im vorangegangenen Kapitel erklärte Median stellt das 50 %-Quantil dar. Es ist jener Wert, der die Datenreihe in der Mitte teilt. Somit liegen 50 % der beobachteten Werte unterhalb des Medians und 50 % der Werte oberhalb des Medians. Analog lässt sich natürlich ein beliebiges Quantil berechnen, z. B. ein 25 %-Quantil. Das 25 % Quantil ist jener Wert, wo 25 % der beobachteten Werte unterhalb und daher 75 % der Werte oberhalb dieser Grenze liegen. Ein 75 %-Quantil würde analog die Datenreihe so aufteilen, dass 75 % der Werte unterhalb dieser Grenze und nur 25 % der Werte oberhalb liegen. Allgemein bezeichnet man Quantile als P-Quantile. Beim P-Quantil liegen P % der Werte unterhalb des Quantiles und $(100-P)$% oberhalb des Quantiles.

Wenden wir uns nochmals dem vorangehenden Beispiel zu. Für die sortierte Datenreihe der gemessenen Zeiten in Sekunden liegen $n = 12$ Werte vor.

96 109 113 120 132 141 217 218 225 489 675 690

Wenn wir das $P = 25$ %-Quantil bestimmen wollen, so ist zuerst die Lage k des Quantiles in der sortierten Datenreihe zu berechnen: $k = P \cdot N = 0{,}25 \cdot 12 = 3$. Analog zu Argumentation, die wir schon beim Median angewandt haben, ist somit der Mittelwert aus dem dritten und vierten Datenpunkt zu nehmen. Wir erhalten somit für das 25 %-Quantil: $x_{25\,\%} = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(113 + 120) = 116{,}5$.

Für das $P = 75$ %-Quantil erhalten wir für die Lage k: $k = P \cdot N = 0{,}75 \cdot 12 = 9$. Somit ist das 75 %-Quantil die Mitte vom neunten und zehnten Wert der sortierten Daten: $x_{75\,\%} = \frac{1}{2}(x_9 + x_{10}) = \frac{1}{2}(225 + 489) = 357$.

Auch im Falle von klassierte Daten, kann das Resultat verallgemeinert werden:

Berechnung des *P*-Quantils bei klassierten Daten:

$$x_P = x_u + \frac{(P - P_u)}{(P_o - P_u)}(x_o - x_u)$$

Dabei gilt:

P	Gewünschtes Quantil
x_u, x_o	untere und obere Grenze der Klasse, die den Median enthält
P_u, P_o	untere und obere kumulierte prozentuale Häufigkeiten

Das 25 %-Quantil bezeichnet den 40. Wert. Gemäß Tabelle 2.9 liegt dieser in Klasse $x_j = 3$. Daher sind die zu benutzenden Werte:

$$x_u = 2{,}5 \quad x_o = 3{,}5 \quad P_u = 0{,}231 \quad P_o = 0{,}375$$

Damit ergibt sich für das 25 %-Quantil

$$x_{25\,\%} = 2{,}5 + \frac{(25\,\% - 23{,}1\,\%)}{(37{,}5\,\% - 23{,}1\,\%)}(3{,}5 - 2{,}5) = 2{,}632$$

Das 75 %-Quantil bezeichnet den 120. Wert. Gemäß Tabelle 2.9 liegt dieser in Klasse $x_j = 5$. Daher sind die zu benutzenden Werte:

$$x_u = 4{,}5 \quad x_o = 5{,}5 \quad P_u = 0{,}706 \quad P_o = 1{,}00$$

Damit ergibt sich für das 75 %-Quantil

$$x_{75\,\%} = 4{,}5 + \frac{(75\,\% - 70{,}6\,\%)}{(100\,\% - 70{,}6\,\%)}(5{,}5 - 4{,}5) = 4{,}650$$

2.5 Dispersionsparameter

Während die Lokalisationsparameter die Lage einer Verteilung oder ihre zentrale Tendenz beschreiben, kennzeichnen die Dispersionsparameter oder Streuungsmaße die Breite einer Verteilung.

Das einfachste Streuungsmaß, *Spannweite* oder Range genannt, ist die Differenz zwischen größtem und kleinstem Wert:

$$\textit{Spannweite} = \textit{Maximum} - \textit{Minimum}$$

Der Nachteil dieses Streuungsmaßes ist, dass es lediglich auf den beiden Extremwerten basiert und somit höchst unsicher ist; es sagt zudem nichts über die dazwischen liegenden Werte aus. Daher wurden, je nach Messniveau, aussagekräftigere Streuungsmaße entwickelt.

2.5.1 Varianz, Standardabweichung und Standardfehler

Um alle Werte in die Berechnung eines geeigneten Streuungsmaßes für intervall- bzw. verhältnisskalierte Variablen einzubeziehen, könnte man versuchen, die Summe der Abweichungen der Werte zum Mittelwert zu verwenden. Diese Summe ist aber Null, da die Summe der Abweichungen der Werte oberhalb des Mittelwertes gleich groß ist wie die Summe der Abweichungen der Werte unterhalb des Mittelwertes. Dies soll mit Hilfe von untenstehender Tabelle anhand von drei Preisen eines Produktes illustriert werden. Der Mittelwert der Messwerte ist $\overline{x} = \frac{1}{3}(3\,€ + 7\,€ + 8\,€) = 6\,€$.

Index i	Wert x_i	Abweichung $x_i - \overline{x}$
1	3 €	−3 €
2	7 €	1 €
3	8 €	2 €
Summe		0 €