Statistik für Wirtschafts-wissenschaftler

Ideal zum Selbst-studium!

Josef Bleymüller
Rafael Weißbach
Achim Dörre

Vahlen

18. Auflage

Bleymüller/Weißbach/Dörre
Statistik für Wirtschaftswissenschaftler

Statistik für Wirtschaftswissenschaftler

von

Professor Dr. Josef Bleymüller,

Professor Dr. Rafael Weißbach

und

Dr. Achim Dörre

18., überarbeitete und erweiterte Auflage

Verlag Franz Vahlen München

Prof. Dr. Josef Bleymüller war Direktor des Instituts für Ökonometrie und Wirtschaftsstatistik der Universität Münster.
Prof. Dr. Rafael Weißbach ist Inhaber des Lehrstuhls für Statistik und Ökonometrie am Institut für Volkswirtschaftslehre der Universität Rostock, an dem auch **Dr. Achim Dörre** als wissenschaftlicher Mitarbeiter tätig ist.

ISBN Print 978 3 8006 6142 8
ISBN E-Book 978 3 8006 6143 5

Wilhelmstraße 9, 80801 München
Satz: satz&sonders GmbH
Weidenstraße 17, 48249 Dülmen
Druck und Bindung: Beltz Grafische Betriebe GmbH
Am Fliegerhorst 8, 99947 Bad Langensalza
Umschlaggestaltung: Ralph Zimmermann – Bureau Parapluie
Bildnachweis: © MicroStockHub – istockphoto.com

www.vahlen.de

Gedruckt auf säurefreiem, alterungsbeständigem Papier
(hergestellt aus chlorfrei gebleichtem Zellstoff)

Vorwort

Die 18. Auflage zeichnet sich neben der Ergänzung der Autorengemeinschaft durch drei neue Abschnitte aus. In Abschnitt 2.5 legen wir nun schon früh die Grundlagen der beschreibenden Statistik für die in Kapitel 19 behandelten Verteilungstests. Aber auch für die Wahrscheinlichkeitsrechnung werden hier einige Konzepte in ihrer beschreibenden Form schon einmal erläutert. Abschnitt 2.5 ermöglicht, dass im neuen Abschnitt 4.7 der Zusammenhang von zwei Merkmalen als Streuungsmaß eines bivariaten Merkmals eingeführt wird. Den neuen Abschnitt 24.6 verwenden wir auf die logistische Regression, die ihrerseits das Verständnis des Zusammenhangs zweier Merkmale erweitert. Wir schließen insofern eine Lücke, als dass in Kapitel 18 mit der ANOVA die Abhängigkeit eines metrischen Merkmals von einem nominalen modelliert wird, aber der umgekehrte Fall der Abhängigkeit eines nominalen von einem metrischen bislang nicht. Wir beschränken uns aber auf ein dichotomes abhängiges Merkmal und haben nötige methodische Grundlagen in Abschnitt 15.5 ergänzt. Über den Text verteilt haben wir auch einige Rechenvereinfachungen eingefügt, die sich in der Lehrpraxis als hilfreich erwiesen haben.

Ziel der Verfasser bleibt, in einem den Studierenden der Wirtschafts- aber auch der Sozialwissenschaften zumutbaren Umfang und Schwierigkeitsgrad diejenigen statistischen Methoden zu behandeln, die als Grundlage für das wissenschaftliche Studium dringlich benötigt werden.

Bei der Darstellung wird besonderer Wert auf gute Verständlichkeit gelegt. Auf die theoretischen Grundlagen wird insoweit eingegangen, wie es für eine korrekte Anwendung der behandelten statistischen Methoden notwendig erscheint. An mathematischen Vorkenntnissen wird beim Leser nicht viel mehr vorausgesetzt als elementare Grundkenntnisse in Infinitesimal- und Matrizenrechnung sowie das Rechnen mit dem Summen- und Produktzeichen.

Zur Ergänzung erscheinen im gleichen Verlag zwei Taschenbücher:

„Übungen zur Statistik für Wirtschaftswissenschaftler“
„Statistische Formeln und Tabellen“

Das Erste soll den zahlreichen Nachfragen von Studierenden Rechnung tragen, durch vielfältige Übungen die behandelten statistischen Methoden zu erlernen. Das Zweite stellt die wichtigsten Formeln und Tabellen zusammen. Dieser handliche Band eignet sich besonders als Nachschlagewerk bei der Anwendung der Methoden und macht damit auch bei Klausuren die Herausgabe gesonderter Hilfsblätter mit Formeln und Tabellen überflüssig.

Herrn Dennis Brunotte sind wir für die langjährige Zusammenarbeit mit dem Vahlen-Verlag dankbar.

Münster und Rostock, im Januar 2020

Josef Bleymüller *Rafael Weißbach* *Achim Dörre*

Inhaltsverzeichnis

1 Einführung

Statistik untersucht Massenerscheinungen. Ob Erhebung, Stichprobe oder Grundgesamtheit, ob metrisches oder nominales Merkmal – erst ein Konsens über die Begriffe ermöglicht das Studium.

1.1 Begriff und Aufgaben der Statistik

Das Wort **Statistik** wurde gegen Ende des 17. Jahrhunderts geprägt und bedeutete lange Zeit ganz allgemein die verbale oder numerische Beschreibung eines bestimmten Staates oder – um eine Definition *Achenwalls* aus dem 18. Jahrhundert zu gebrauchen – den Inbegriff der *Staatsmerkwürdigkeiten eines Landes und Volkes*.

Heute wird das Wort „Statistik" im doppelten Sinne gebraucht: Einmal versteht man darunter *quantitative Informationen* über bestimmte Tatbestände schlechthin, wie z. B. die „Bevölkerungsstatistik" oder die „Umsatzstatistik", zum anderen aber eine *formale Wissenschaft*, die sich mit den *Methoden der Erhebung, Aufbereitung und Analyse numerischer Daten* beschäftigt.

Statistische Methoden gehören zum unentbehrlichen Instrumentarium vieler Fachwissenschaften wie – um nur einige zu nennen – der Physik, der Biologie, der Medizin, der Geographie, der Psychologie und natürlich auch der *Wirtschaftswissenschaften*. Es tut der Bedeutung der Statistik in der Volkswirtschaftslehre und der Betriebswirtschaftslehre keinen Abbruch, dass in diesen Disziplinen die **Ökonometrie** und die **Unternehmensforschung** (*Operations Research*) als neue spezielle Hilfswissenschaften hinzugetreten sind. Vielmehr sollte erwähnt werden, dass sowohl in der Ökonometrie als auch in der Unternehmensforschung in nicht unbeträchtlichem Ausmaße wiederum auf statistische Methoden zurückgegriffen wird.

In der **deskriptiven Statistik** werden vor allem methodisch einfachere Probleme wie die Darstellung von Daten in Tabellen und Schaubildern, die Berechnung von Mittelwerten und Streuungsmaßen, die Indexberechnung und die Konzentrationsmessung behandelt. Der Schwerpunkt der statistischen Forschung liegt heute allerdings auf der wahrscheinlichkeitstheoretisch fundierten sogenannten **induktiven Statistik**, die Problemkreise wie statistisches Schätzen, statistische Tests, statistische Entscheidungstheorie und multivariate statistische Methoden umfasst.

Während die Statistik früher vor allem eine **deskriptive** (*beschreibende*) **Funktion** hatte, rückt ihre **operationale Funktion**, d. h. ihre Anwendung bei der *Entscheidungsfindung*, immer mehr in den Vordergrund. Das gilt sowohl für den volkswirtschaftlichen als auch für den betriebswirtschaftlichen Anwendungsbereich. So sind die Ausgaben eines Staates für die amtlichen statistischen Dienste oder die Ausgaben eines Unternehmens für die laufenden statistischen Aktivitäten keineswegs Selbstzweck. Sie werden vielmehr vor allem deswegen getätigt, weil die Verantwortlichen bei ihren Entscheidungen auf quantitative Informationen nicht verzichten können.

Oft wird vom Statistiker nicht nur die Vorlage von Ist-Zahlen der Vergangenheit, sondern auch deren Extrapolation in die Zukunft hinein verlangt. Er muss sich dann auf das recht schwierige Gebiet der **statistischen Prognosen** begeben. Da der Grad ihrer Übereinstimmung mit der zukünftig tatsächlich eintretenden Entwicklung entscheidend von den im Zeitpunkt der Prognoseerstellung getroffenen *Annahmen* abhängt, sollten diese *stets sorgfältig präzisiert* werden. Als Beispiele wirtschaftswissenschaftlich wichtiger Prognosen seien nur Bevölkerungsprognosen, Energieprognosen, Steuervorausschätzungen und Absatzprognosen genannt.

1.2 Träger der Wirtschaftsstatistik und ihre Veröffentlichungen

Wenn man als Volks- oder Betriebswirt statistisches Zahlenmaterial benötigt, wird man nur in ganz seltenen Fällen eigene Erhebungen durchführen können. Man wird vielmehr bereits vorliegende Veröffentlichungen der verschiedenen **Träger der Wirtschaftsstatistik** zu Rate ziehen müssen.

Wohl die wichtigste und für den Benutzer auch besonders kostengünstige Quelle wirtschaftsstatistischer Daten ist die **amtliche Statistik**. Sie umfasst einmal die *statistischen Ämter*; in der Bundesrepublik Deutschland sind dies das Statistische Bundesamt (Wiesbaden), die Statistischen Landesämter und die Städtestatistischen Ämter. Von zentraler Bedeutung für die Datenfindung sind die vom Statistischen Bundesamt herausgegebenen Veröffentlichungen *Das Statistische Jahrbuch für die Bundesrepublik Deutschland* (kostenloser Download über „Publikationsservice" des Statistischen Bundesamts) und die monatlich erscheinende Zeitschrift *Wirtschaft und Statistik*. Sie enthalten nicht nur eine Fülle von Wirtschaftsdaten, sondern erschließen durch zahlreiche zusätzliche Angaben zu den einzelnen Statistiken und viele Verweisungen den Zugang zu dem außerordentlich weit gestreuten nationalen und internationalen Veröffentlichungsmaterial. – Im *Internet* ist das *Statistische Bundesamt* über die Seite **www.destatis.de** zu erreichen und die von ihm gemeinsam mit den Statistischen Landesämtern betriebene Seite: *Statistische Ämter des Bundes und der Länder* über **www.statistikportal.de**.

Zur amtlichen Statistik gehört neben den statistischen Ämtern auch die sogenannte *Ressortstatistik*, wie sie in den Bundes- und Landesministerien sowie den ihnen nachgeordneten Behörden betrieben wird. Ein wichtiges Beispiel ist die *Arbeitsmarktstatistik*, die im wesentlichen von der zum Bereich des Bundesministeriums für Arbeit und Sozialordnung gehörenden Bundesagentur für Arbeit in Nürnberg durchgeführt wird.

Neben der amtlichen Statistik steht die **nichtamtliche Statistik**; deren Träger sind u. a. *Wirtschaftsverbände, Arbeitgeber- und Arbeitnehmerorganisationen, Industrie- und Handelskammern, Markt- und Meinungsforschungsinstitute, wirtschaftswissenschaftliche Forschungsinstitute* (teilweise auch solche an Hochschulen) *und größere Unternehmen*. Ein großer Teil des im Bereich der nichtamtlichen Statistik erhobenen, aufbereiteten und/oder analysierten statistischen Materials wird veröffentlicht und ist so jedem Interessenten zugänglich. Eine Ausnahme machen hier jedoch die kommerziell betriebenen Markt- und Meinungsforschungsinstitute, welche die meisten Ergebnisse nur *gegen Entgelt* zur Verfügung stellen; eine Reihe von Umfragen wird von ihnen überhaupt erst auf einen entsprechenden Kundenauftrag hin vorgenommen. Andere wirtschaftsstatistische Informationen bleiben der breiten Öffentlichkeit *vorenthalten*; so gibt es Beispiele für Verbandsstatistiken, die ausschließlich Verbandsmitgliedern zur Verfügung stehen.

Auch **internationale Organisationen** sind auf statistischem Gebiet aktiv; es seien beispielhaft folgende genannt: *United Nations (UN)* mit wichtigen Veröffentlichungen wie

Statistical Yearbook, Yearbook of International Trade Statistics und Demographic Yearbook;

Food and Agriculture Organization (FAQ) mit ihrem Production Yearbook etc.;

Europäische Union (EU) mit ihrem umfangreichen Veröffentlichungsprogramm.

Ein detaillierter Überblick ist dem Anhang „*Internationale Übersichten*“ des Statistischen Jahrbuchs der Bundesrepublik Deutschland zu entnehmen.

1.3 Vorgehensweise bei statistischen Untersuchungen

Bei einer statistischen Untersuchung sind folgende fünf Schritte zu unterscheiden, deren Gewicht von Fall zu Fall stark variieren kann:

(1) Planung
Hierunter fallen vor allem die exakte Formulierung des Untersuchungszieles, die Festlegung des Erhebungsprogramms sowie die Klärung organisatorischer Fragen.

(2) Erhebung
Die Erhebung dient der Gewinnung des statistischen Datenmaterials. Man unterscheidet zwischen *primärstatistischen* und *sekundärstatistischen* Untersuchungen. Bei der primärstatistischen Untersuchung müssen die Daten eigens für den Untersuchungszweck erhoben werden. Die sekundärstatistische Untersuchung kann auf schon vorhandene Daten zurückgreifen, die etwa auch für andere Zwecke bereits gesammelt worden sind. – Bei der primärstatistischen Untersuchung lassen sich die folgenden Erhebungsarten unterscheiden:

(a) *Schriftliche Befragung:*
Der Vorteil des Fragebogens liegt vor allem in den geringen Kosten. Falls jedoch kein Auskunftszwang besteht, kommt unter Umständen nur ein kleiner Teil der Fragebogen zurück, worunter die Repräsentativität der Ergebnisse leiden kann. Nachteilig ist auch der relativ lange Erhebungszeitraum.

(b) *Mündliche Befragung:*
Das Interview ist eine relativ teure Erhebungsart, wird jedoch bei intensiver Schulung der Interviewer und sorgfältiger Abfassung des Fragebogens zu guten Ergebnissen führen.

(c) *Beobachtung:*
Diese Erhebungsart bringt exakte Ergebnisse, ist jedoch in den Wirtschaftswissenschaften relativ selten anwendbar.

(d) *Experiment:*
Auch diese Erhebungsart findet vor allem in den Naturwissenschaften und in der Psychologie Verwendung. Ein Anwendungsfall in den Wirtschaftswissenschaften ist der sogenannte Produkttest, bei dem auf experimenteller Basis die subjektiven Wirkungen der zu untersuchenden Waren auf bestimmte Testpersonen festgestellt werden.

(e) *Automatische Erfassung:*
Die Erhebung erfolgt im Augenblick der Entstehung der Daten; beispielsweise werden die Verkaufsdaten in einem computergestützten Warenwirtschaftssystem durch *Scannen* der Waren an der Kasse automatisch erfasst. Weiterhin wäre etwa an die Messung der tageszeitlichen Auslastung eines Telefonnetzes oder eines städtischen Elektrizitätswerkes zu denken.

(3) Aufbereitung
Hierunter versteht man die *Verdichtung* des Urmaterials bis hin zu Tabellen und Schaubildern. Je nach Umfang des Urmaterials wird man sich manueller oder maschineller Verfahren bedienen.

(4) Analyse
Bereits die verdichtete Darstellung der Daten in Tabellen und Schaubildern kann als eine elementare Analyse angesehen werden. Die eigentliche Analyse bedient sich jedoch *mathematisch-statistischer Methoden*, wie sie in diesem Kurs behandelt werden sollen.

(5) Interpretation
In diesem letzten Schritt werden die erhaltenen Ergebnisse interpretiert und in Aussagen zusammengefasst.

1.4 Statistische Einheiten und statistische Gesamtheiten

Das Interesse der Statistik richtet sich nie auf ein *einzelnes, elementares Objekt (statistische Einheit, Element)*, sondern stets auf *Mengen von Elementen*, die als **statistische Gesamtheiten** oder **statistische Massen** bezeichnet werden. In einer statistischen Gesamtheit sollten sinnvollerweise nur solche Elemente zusammengefasst werden, die *vom Untersuchungsziel her als gleichartig* angesehen werden. Der Klarheit wegen muss jede zu untersuchende Gesamtheit *zeitlich*, *räumlich* und *sachlich* eindeutig *abgegrenzt* werden. *Beispiele* für statistische Gesamtheiten, die aus materiellen und immateriellen Objekten wie Gegenständen, Personen, Ereignissen usw. bestehen können, sind etwa:

(a) Erwerbstätige in der Bundesrepublik am 12. Februar 20..,
(b) Rechnungen des Unternehmens A im Monat April 19.. und
(c) Tödliche Verkehrsunfälle in der Bundesrepublik im Jahre 20...

Die Forderung nach exakter Abgrenzung erfordert oft zusätzliche Überlegungen; so entsteht im Beispiel (a) die Frage, wie die Teilzeitbeschäftigten oder auch die deutschen Angehörigen deutscher Firmen im Ausland zu erfassen sind.

Neben **realen statistischen Gesamtheiten**, wie sie eben betrachtet wurden, gibt es auch **hypothetische Gesamtheiten**, wie z. B. die Menge der Ergebnisse eines theoretisch fortlaufend ausgespielten Würfels. Ihrem Umfang nach wäre diese letztgenannte Gesamtheit – wenigstens dem Modell nach – keine **endliche**, sondern eine **unendliche Gesamtheit**. Eine Unterscheidung der statistischen Gesamtheiten, die oft getroffen wird, besteht in der Aufteilung in *Bestandsmassen (Streckenmassen)* und *Bewegungsmassen (Punktmassen)*.

Bei den **Bestandsmassen** kann den einzelnen Elementen eine „Lebensdauer“ *(Zeitstrecke)* zugeordnet werden; da mehrere Elemente gleichzeitig nebeneinander existieren können, werden die Bestandsmassen zu gewissen Zeitpunkten erfasst. Als Beispiele für Bestandsmassen seien genannt:

(a) Einwohner der Bundesrepublik am 1.1.20..,
(b) Positionen eines Lagers am 30.6.20.. und
(c) Kassenbestand eines Warenhauses am 31.12.20...

Jedem einzelnen Element einer Bewegungsmasse kann nur ein Zeitpunkt zugeordnet werden. Die Elemente werden auch als „Ereignisse“ bezeichnet und können, da sie zeitlich aufeinanderfolgen, nur innerhalb bestimmter Zeitspannen erfasst werden. Beispiele für Bewegungsmassen sind:

(a) Geburten in der Bundesrepublik im Jahre 20..,
(b) Baufertigstellungen in Nordrhein-Westfalen im Oktober 20.. und
(c) Bei einer Bank im Monat April 20.. eingegangene Schecks.

Bestands- und Bewegungsmassen können durch die Fortschreibungsformel zueinander in Beziehung gesetzt werden. Da für jedes Element einer Bestandsmasse sowohl der Beginn als auch das Ende der Existenz ein Ereignis darstellt, gilt:

Anfangsbestand + Zugang − Abgang = Endbestand
(Bestandsmasse) (Bewegungsmassen) (Bestandsmasse)

Eine im Rahmen der modernen Statistik besonders wichtige Unterscheidung ist die von **Grundgesamtheit** und **Stichprobe** *(Teilgesamtheit)*. Beispiele für Grundgesamtheiten sind etwa „Sämtliche Haushalte in der BRD“ oder „Alle im Werk 3 hergestellten Leuchtstoffröhren vom Typ 134“. Soll nun z. B. ermittelt werden, wie groß der Anteil der Haushalte ist, die ein bestimmtes Waschmittel verwenden, bzw. wie groß die durchschnittliche Brenndauer der Leuchtröhren ist, so wird der Statistiker in keinem der beiden Fälle sämtliche Elemente der Grundgesamtheit erfassen können, und zwar im ersten Fall aus Kostengründen und im zweiten Fall des vernichtenden Charakters der Qualitätskontrolle wegen. Er wird deshalb nur Stichproben untersuchen, deren Elemente zweckmäßigerweise nach gewissen Zufallsprinzipien aus den Grundgesamtheiten ausgewählt werden. Anhand der Stichprobenergebnisse lassen sich dann mit einer bestimmten Wahrscheinlichkeit der Anteil der Verwender des Waschmittels oder die durchschnittliche Brenndauer der Leuchtstoffröhren in der Grundgesamtheit schätzen. Die der induktiven Statistik zugehörigen Schätzmethoden werden weiter unten noch ausführlich behandelt.

1.5 Merkmale, Merkmalsausprägungen und Skalen

Bei statistischen Untersuchungen interessieren an jeder statistischen Einheit ein einziges **Merkmal** *(charakteristische Eigenschaft)* oder auch mehrere. So können etwa an einer Person die folgenden Merkmale von Interesse sein: Alter, Geschlecht, Größe, Einkommen usw. Jedes Merkmal hat zwei oder mehr Merkmalsausprägungen, die nach Art des betrachteten Merkmals anhand verschiedener **Skalen** gemessen werden.

Ihre Unterscheidung ist deshalb von Bedeutung, weil sie den Kreis der anwendbaren statistischen Methoden bestimmen. Man unterscheidet vier Skalentypen:

(1) Nominalskala
Diese Skala findet bei Merkmalen Anwendung, bei denen die Ausprägungen *keine natürliche Reihenfolge* bilden, sondern *gleichberechtigt nebeneinanderstehen*. Beispiele sind:

(a) Religion,
(b) Geschlecht,
(c) Farbe und
(d) Autokennzeichen.

Jeder einzelnen Merkmalsausprägung kann eine Zahl zugeordnet werden *(Verschlüsselung)*, diese Zahlen dienen aber nur der Identifikation der einzelnen Gruppen.

(2) Ordinalskala
Hier besteht zwischen den einzelnen Merkmalsausprägungen eine natürliche Rangordnung. Es lässt sich zwischen den Merkmalsausprägungen eine „*größer als*“-*Beziehung* aufstellen; allerdings sind die Abstände zwischen den Merkmalsausprägungen nicht quantifizierbar. Beispiele sind:

(a) Examensnoten,
(b) Güteklassen bei Lebensmitteln und
(c) Rangplätze einer Fußballiga.

Eine Ordinalskala mit ausschließlich *ganzzahligen Ordnungsziffern (Rängen, Rangziffern)*, die mit 1 beginnend in ununterbrochener Reihenfolge hintereinander stehen, wie z. B. die Rangplätze $1, 2, \ldots$ der Bundesliga, heißt *Rangskala*.

(3) Intervallskala
Neben die Rangordnung tritt hier noch die Möglichkeit, die *Abstände zwischen den einzelnen Merkmalsausprägungen anzugeben*. Dafür ist es notwendig, dass die Ausprägungen als Vielfaches einer elementaren Maßeinheit angegeben werden können. Der 0-Punkt kann willkürlich festgelegt werden. Beispiele für Intervallskalen sind:

(a) Temperaturmessung in °C und
(b) Kalenderzeitrechnung.

Bei intervallskalierten Merkmalen dürfen keine Quotienten gebildet werden; so ist z. B. die Aussage „20 °C ist doppelt so warm wie 10 °C“ sinnlos.

(4) Verhältnisskala
Zusätzlich zu den Eigenschaften der Intervallskala hat die Verhältnisskala noch einen *absoluten Nullpunkt*. Dadurch wird der Quotient zweier Ausprägungen unabhängig von der gewählten Maßeinheit. Beispiele sind:

(a) Körpergröße,
(b) Alter und
(c) Einkommen.

Da Intervall- und Verhältnisskalen ein Maßsystem zugrunde liegt, werden sie auch vielfach als **metrische Skalen** bezeichnet; die Merkmalsausprägungen bezeichnet man hier auch als **Merkmalswerte**.

Jede Menge von Merkmalsausprägungen, die an den Elementen einer statistischen Gesamtheit gemessen werden, beinhaltet ein ganz *bestimmtes Ausmaß an Informationen* eben über diese Gesamtheit. Dieses Ausmaß an Informationen ist von der benutzten Skala abhängig und nimmt – wie man sich an Beispielen leicht klar machen kann – des hierarchischen Aufbaus der vier betrachteten Skalen wegen von 1 bis 4 zu. Mit jedem Skalentyp ist also ein *eindeutig festgelegtes Informationsniveau* verbunden. – Aus methodischen oder anderen Gründen ist es nun oft notwendig, *Merkmalsausprägungen* zu *transformieren*; so werden zwei bestimmte Artikel eines Versandkataloges bei dessen Neuauflage beispielsweise von 113 und 114 in 100 113 und 100 114 „umgeschlüsselt“. - Selbstverständlich wird man nur solche Transformationen vornehmen wollen, bei denen die ursprünglich enthaltenen Informationen unverändert erhalten bleiben. Man führt deshalb folgende Definition ein:

Eine Transformation von Skalenwerten ist auf einem bestimmten Skalenniveau nur dann *zulässig (informationserhaltend)*, wenn die in den Skalenwerten enthaltenen Informationen dabei nicht verändert werden. Bei jedem Skalentyp sind nun ganz bestimmte Transformationen zulässig:

1. *Nominalskala:* Zulässig sind *symmetrische Transformationen*, bei denen lediglich die Klassenbezeichnungen geändert werden, wie etwa beim oben angegebenen Beispiel.

2. *Ordinalskala:* Zulässig sind *streng monotone Transformationen*, bei denen der neue Skalenwert $x^\star$ aus dem alten Skalenwert x als $x^\star = f(x)$ so gebildet wird, dass für zwei Skalenwerte $x_1 < x_2$ nach der Transformation $x_1^\star < x_2^\star$ gilt.
3. *Intervallskala:* Zulässig sind *lineare Transformationen* der Art $x^\star = ax + b \quad (a > 0)$.
4. *Verhältnisskala:* Zulässig sind *Ähnlichkeitstransformationen* des Typs $x^\star = ax \quad (a > 0)$.

Es bleibt dem Leser überlassen, diese vier Lehrsätze (Theoreme) anhand selbst gewählter Beispiele zu überprüfen.

Früher war es üblich, nach **qualitativen** und **quantitativen Merkmalen** zu differenzieren, wobei es sich bei den qualitativen Merkmalen um nominalskalierte und bei den quantitativen Merkmalen um metrisch skalierte Merkmale handelt; eine Einordnung der ordinalskalierten Merkmale macht hier Schwierigkeiten.

Bei den metrisch skalierten Merkmalen unterscheidet man zwischen **diskreten** und **stetigen (kontinuierlichen) Merkmalen**. Ein Merkmal wird dann als diskret bezeichnet, wenn es auf einer metrischen Skala nur bestimmte Werte annehmen kann. Kann es dagegen – zumindest in einem bestimmten Intervall – jeden beliebigen Wert annehmen, dann spricht man von einem stetigen Merkmal. *Diskret* sind beispielsweise die folgenden Merkmale:

(a) Zahl der Studenten in einem Hörsaal,
(b) Zahl der Beschäftigten eines Betriebs und
(c) Geldeinkommen.

Als stetig sind etwa die folgenden Merkmale anzusehen:

(a) Lebensalter,
(b) Länge eines Werkstücks und
(c) Füllgewicht.

Stetige Merkmale lassen sich allerdings in der Praxis wegen der Grenzen der Messgenauigkeit *nur diskret erfassen*. So ist das *Füllgewicht* zwar ein stetiges Merkmal, es kann aber nicht feiner gemessen werden, als es die kleinste auf der Waage angegebene Skaleneinheit zulässt.

1.6 Ausgewählte Literatur

(a) Statistische Lehrbücher

Deutschsprachig:

Bamberg, Günter, Franz Baur, Michael Krapp, Statistik (17., überarb. Aufl.). München 2012.

Bohley, Peter, Statistik – Einführendes Lehrbuch für Wirtschafts- und Sozialwissenschaftler (7., gründl. überarb. u. akt. Aufl.). München 2000.

Bosch, Karl, Elementare Einführung in die angewandte Statistik (9., erw. Aufl.). Wiesbaden 2010.

Bosch, Karl, Elementare Einführung in die Wahrscheinlichkeitsrechnung (9., durchges. Aufl.). Wiesbaden 2006.

Hartung, Joachim, Bärbel Elpelt, Karl-Heinz Klösener, Statistik – Lehr- und Handbuch der angewandten Statistik (14., unwesentl. veränd. Aufl.). München 2005.

Hochstädter, Dieter, Statistische Methodenlehre (8., verb. Aufl.). Frankfurt a. M. 1996.

Kreyszig, Erwin, Statistische Methoden und ihre Anwendungen (7. Aufl., 5., unveränd. Nachdruck). Göttingen 1999.

Pfanzagl, Johann, Allgemeine Methodenlehre der Statistik, Teil 1 (6., verb. Aufl.) und Teil 2 (5., verb. Aufl.). Berlin, New York 1983 und 1978.

Schaich, Eberhard, Dieter Köhle, Walter Schweitzer, Fritz Wegner, Statistik für Volkswirte, Betriebswirte und Soziologen, Teil I (4., überarb. Aufl.) und Teil II (3., überarb. Aufl.). München 1993 und 1990.

Schlittgen, Rainer, Einführung in die Statistik (12. korr. Aufl.). München 2012.

Ven, Ad van der, Einführung in die Skalierung, übersetzt und herausgegeben von Jo Goebel. Aachen, Bern, Stuttgart, Wien 1980.

Vogel, Friedrich, Beschreibende und schließende Statistik (13., korr. u. erw. Aufl.). München 2005.

Englischsprachig:

DeGroot, Morris H., Mark J. Schervish, Probability and Statistics (4th rev. ed.) Reading (Mass.), Menlo Park (Cal.), Don Mills (Ont.) usw. 2010.

Harnett, Donald L., James L. Murphy, Introductory Statistical Analysis (3rd ed.). Reading (Mass.), Menlo Park (Cal.), London usw. 1982.

Kahler, Heinz, Essentials of Statistics. Glenview (III.), London, Boston 1988.

Pfaffenberger, Roger C., James H. Patterson, Statistical Methods for Business and Economics (4th ed.). Homewood (III.) 1991.

Siegel, Andrew F., Charles J. Morgan, Statistics and Data Analysis – An Introduction (2nd ed.). New York, Chichester, Brisbane usw. 1998.

Walpole, Ronald E., Introduction to Statistics (3rd ed.). New York, London 1982.

Wonnacott, Thomas H., Ronald J. Wonnacott, Introductory Statistics for Business and Economics (5th ed.). New York, Chichester, Brisbane usw. 1990.

(b) Lehrbücher der Wirtschaftsstatistik

Abels, Heiner, Horst Degen, Wirtschafts- und Bevölkerungsstatistik (3., vollst. überarb. u. erw. Aufl.). Wiesbaden 1992.

Kunz, Dietrich, Praktische Wirtschaftsstatistik. Stuttgart, Berlin, Köln, Mainz 1987.

Lippe, Peter Michael von der, Wirtschaftsstatistik (5., völl. neubearb. u. erw. Aufl.). Stuttgart 1996.

Rinne, Horst, Wirtschafts- und Bevölkerungsstatistik; Erläuterungen – Erhebungen – Ergebnisse (2., überarb. u. erw. Aufl.). München 1996.

Schaich, Eberhard, Walter Schweitzer, Ausgewählte Methoden der Wirtschaftsstatistik. München 1995.

Ungerer, Albrecht, Siegfried Hauser, Wirtschaftsstatistik als Entscheidungshilfe. Freiburg i. Brsg. 1986.

Wagenführ, Rolf, Wirtschafts- und Sozialstatistik gezeigt am Beispiel der BRD, Teil 1 und Teil 2. Freiburg i. Brsg. 1970 und 1973.

Zwer, Reiner, Einführung in die Wirtschafts- und Sozialstatistik (2., überarb. u. erw. Aufl.). München 1994.

Aufgaben zu Kapitel 1

Die Lösung zu diesen Aufgaben finden Sie am Ende des Buches.

1.1 Ermitteln Sie anhand des *Statistischen Jahrbuchs 2011 für die Bundesrepublik Deutschland* folgende statistische Angaben:
(a) Prozentanteil der Bevölkerung der Bundesrepublik Deutschland, die 2009 65 Jahre und älter war.
(b) Umsatz im Baugewerbe in der Bundesrepublik 2009.
(c) Prozentuale Veränderung des Verbrauchs an Fleisch und Fleischerzeugnissen je Einwohner und Jahr (in kg) zwischen 2001 und 2009.
(d) Produktion von Bier aus Malz (ohne alkoholfreies Bier) und Zahl der produzierten Zigaretten in der Bundesrepublik 2010.
(e) Personenkraftwagen insgesamt und je 1000 Einwohner in der Bundesrepublik Deutschland 2008.

1.2 Handelt es sich bei den folgenden statistischen Gesamtheiten um Bestands- oder Bewegungsmassen?
(a) Studierende einer Universität,
(b) Todesfälle in einer Gemeinde,
(c) Personenkraftwagen einer Behörde,
(d) Maschinenausfälle in einer Werkstatt,
(e) Anmeldungen in einem Einwohnermeldeamt und
(f) Wartende Postkunden vor einem Schalter.

1.3 Der Bestand eines bestimmten Halbfabrikats betrug am Wochenanfang 6318 und am Wochenende 7480 Stück. Für diesen Zeitraum wurde ein Lagerzugang von 3620 Stück festgestellt. Wie hoch war der Lagerabgang in dieser Woche?

1.4 Sind die folgenden Merkmale diskret oder stetig?
(a) Rechnungsbetrag,
(b) Wahlergebnis einer Partei,
(c) Kraftstoffverbrauch eines Personenkraftwagens auf 100 km,
(d) Zeitspanne, die zur Verrichtung einer bestimmten Arbeit benötigt wird,
(e) Zahl der pro Stunde in einem Geschäft eintreffenden Kunden und
(f) Grundstücksgröße.

1.5 Auf welche Skalen sind die folgenden Transformationen – ohne Informationsverlust – anwendbar $(a, b > 0)$?
(a) $x^* = bx^2$, $(x > 0)$,
(b) $x^* = a + bx$ und
(c) $x^* = bx$.

1.6 Welches Skalenniveau besitzt das Merkmal „Jahresumsatz“. Auf welches Skalenniveau gelangt man, wenn man die Rangordnung von Unternehmen einer Branche nach Jahresumsätzen angibt? Welche Information wird dabei aufgegeben?

2

Empirische Verteilungen

Dass Menschen und andere Merkmalsträger unterschiedlich sind, ist zweifellos. Aber ob es in den Unterschieden Muster gibt und ob diese Muster als mathematische Funktionen darstellbar sind, das soll der Begriff der Verteilung analysierbar machen: Wie verteilen sich die Merkmalsträger (der statistischen Masse) auf die möglichen Merkmalsausprägungen?

2.1 Häufigkeitsverteilung

Betrachtet man bei einer statistischen Gesamtheit mit N Elementen (Merkmalsträgern) ein *einziges metrisch skaliertes Merkmal*, so wird dieses in der Regel bei den einzelnen Elementen in unterschiedlichen Ausprägungen auftreten. Durch Aneinanderreihung dieser beobachteten Merkmalsausprägungen erhält man eine Beobachtungsreihe oder **Urliste**. Eine elementare statistische Tätigkeit besteht nun darin, die auf eine bestimmte Merkmalsausprägung entfallende Anzahl von Elementen auszuzählen. Hat das Merkmal etwa k Merkmalsausprägungen, $x_1, \ldots, x_k$, so ist h_i $(i = 1, \ldots, k)$ die Anzahl der Elemente, welche die Merkmalsausprägung x_i besitzen; man bezeichnet h_i als die **absolute Häufigkeit** der Ausprägung x_i. Dividiert man die absoluten Häufigkeiten h_i durch die Gesamtzahl der Elemente N, so erhält man die **relativen Häufigkeiten** f_i:

$$f_i = \frac{h_i}{N} \qquad (i = 1, \ldots, k).$$

Für die absoluten Häufigkeiten h_i gilt

$$0 \le h_i \le N \qquad (i = 1, \ldots, k)$$

und

$$h_1 + h_2 + \ldots + h_k = \sum_{i=1}^{k} h_i = N\,.$$

Damit ergibt sich für die relativen Häufigkeiten f_i

$$0 \le f_i \le 1 \qquad (i = 1, \ldots, k)$$

und

$$\sum_{i=1}^{k} f_i = 1\,.$$

Die Darstellung der Merkmalsausprägungen x_i mit den dazugehörenden Häufigkeiten h_i bzw. f_i in *tabellarischer* oder *grafischer Form* bezeichnet man als **Häufigkeitsverteilung**.

Zur Erläuterung diene folgendes *Beispiel:* Ein Zeitungskioskinhaber notiert 200 Tage lang täglich die Zahl der verkauften Exemplare einer bestimmten Zeitung; die Ergebnisse der so entstandenen **Urliste** sind in Tabelle 2.1 ausschnittsweise wiedergegeben.

Laufende Nummer des Beobachtungstags	Anzahl der verkauften Zeitungen
1	3
2	1
3	0
4	2
⋮	⋮
199	2
200	5

Tabelle 2.1: Urliste

Die statistischen Einheiten sind hier die Beobachtungstage, das untersuchte Merkmal ist die Anzahl der an einem Tag verkauften Zeitungen. Die Ermittlung der Häufigkeiten h_i erfolgt über die in Tabelle 2.2 dargestellte **Strichliste**.

Nr. i	Anzahl der verkauften Zeitungen x_i	Anzahl der Tage mit x_i verkauften Zeitungen
1	0	𝍸 𝍸 𝍸 𝍸 I
2	1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I
3	2	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 IIII
4	3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
5	4	𝍸 𝍸 𝍸 𝍸 𝍸 IIII
6	5	𝍸 𝍸
7	6	𝍸
8	7 und mehr	

Tabelle 2.2: Strichliste

Durch Auszählung ergibt sich aus der Strichliste die in Tabelle 2.3 dargestellte Häufigkeitsverteilung, in der sowohl die absoluten Häufigkeiten (h_i) als auch die relativen Häufigkeiten (f_i) sowie die Prozentanteile $(100f_i)$ wiedergegeben sind.

Nr. i	Anzahl der verkauften Zeitungen x_i	Anzahl der Tage h_i	Anteil der Tage f_i	Prozentanteil der Tage $100f_i$
1	0	21	0,105	10,5
2	1	46	0,230	23,0
3	2	54	0,270	27,0
4	3	40	0,200	20,0
5	4	24	0,120	12,0
6	5	10	0,050	5,0
7	6	5	0,025	2,5
	$\sum$	200	1,000	100

Tabelle 2.3: Häufigkeitsverteilung

Bei der grafischen Darstellung der Häufigkeitsverteilung kann man, da hier ein diskretes Merkmal vorliegt, zwischen einer *höhenproportionalen* (Stabdiagramm) und einer *flächenproportionalen Darstellung* (Histogramm) wählen. Beim **Stabdiagramm** (vgl. Abbildung 2.1) werden die Häufigkeiten durch Strecken, beim **Histogramm** (vgl. Abbildung 2.2) durch Flächen von Säulen beschrieben.

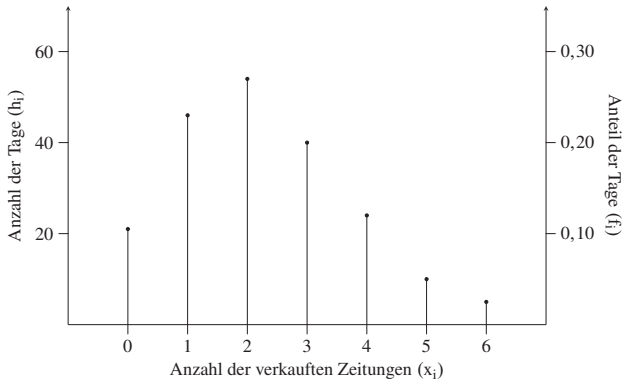

Abb. 2.1: Stabdiagramm

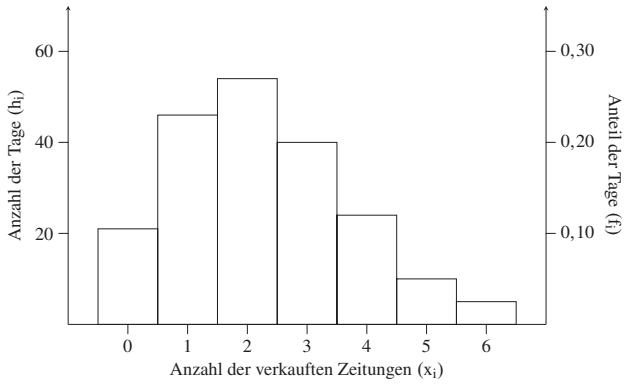

Abb. 2.2: Histogramm

Da absolute und relative Häufigkeiten zueinander proportional sind (im Beispiel ist $f_i/h_i = 1/200$), kann das im Schaubild durch einen doppelten Maßstab berücksichtigt werden (vgl. Abbildung 2.1 und 2.2).

2.2 Summenhäufigkeitsfunktion

Durch fortlaufende Summierung *(Kumulierung)* lassen sich aus den absoluten Häufigkeiten die **absoluten Summenhäufigkeiten** H_i wie folgt ermitteln:

$$H_i = h_1 + \ldots + h_i = \sum_{j=1}^{i} h_j \qquad (i = 1, \ldots, k)$$

H_i gibt die Anzahl der Elemente an, die einen Merkmalswert besitzen, der *höchstens* x_i beträgt. In entsprechender Weise lassen sich die **relativen Summenhäufigkeiten** F_i berechnen:

$$F_i = f_1 + \ldots + f_i = \sum_{j=1}^{i} f_j \qquad (i = 1, \ldots, k)$$

oder

$$F_i = \frac{H_i}{N} \qquad (i = 1, \ldots, k)$$

Nr. i	Anzahl der verkauften Zeitungen x_i	Anzahl der Tage, an denen höchstens x_i Zeitungen verkauft wurden H_i	Anteil F_i
1	0	21	0,105
2	1	67	0,335
3	2	121	0,605
4	3	161	0,805
5	4	185	0,925
6	5	195	0,975
7	6	200	1,000

Tabelle 2.4: Absolute und relative Summenhäufigkeiten

Mithilfe der relativen Summenhäufigkeiten lässt sich die **Summenhäufigkeitsfunktion** *(empirische Verteilungsfunktion)* F(x) definieren. F(x) gibt den Anteil der Elemente mit einem Merkmalswert kleiner oder gleich x an:

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ F_i & \text{für } x_i \le x < x_{i+1} \\ 1 & \text{für } x \ge x_k \end{cases} \qquad (i = 1, \ldots, k-1)$$

Wie in Abbildung 2.3 gezeigt wird, hat die Summenhäufigkeitsfunktion das Bild einer *Treppenfunktion*; in jedem Merkmalswert x_i springt die Summenhäufigkeitsfunktion auf den entsprechenden Wert F_i.

In umgekehrter Weise können aus den Summenhäufigkeiten wieder die einzelnen Häufigkeiten ermittelt werden. Es gilt hier nämlich:

$$h_i = H_i - H_{i-1} \qquad (i = 1, \ldots, k)$$

bzw.

$$f_i = F_i - F_{i-1} \qquad (i = 1, \ldots, k)$$

mit

$$H_0 = F_0 = 0\,.$$

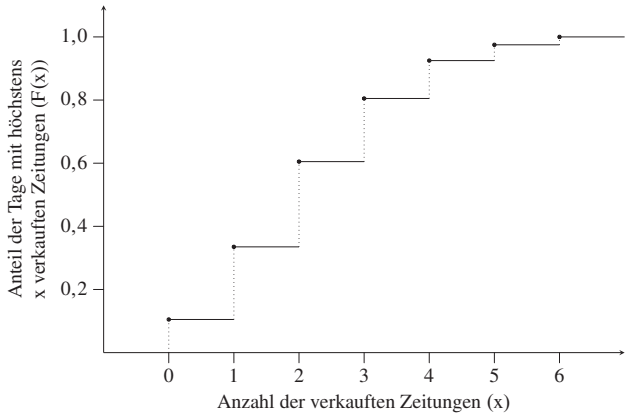

Abb. 2.3: Summenhäufigkeitsfunktion

2.3 Häufigkeitsverteilung klassifizierter Daten

Liegt entweder ein *diskretes Merkmal mit sehr vielen unterschiedlichen Merkmalsausprägungen* vor, oder handelt es sich um ein *stetiges Merkmal*, so ist es zweckmäßig, die Häufigkeiten nicht mehr jeder einzelnen Merkmalsausprägung zuzuordnen, sondern die Merkmalsausprägungen in **Klassen** zusammenzufassen und die Anzahl der Elemente zu bestimmen, deren Merkmalswerte in die einzelnen Klassen fallen. Jede Klasse i ist charakterisiert durch die *untere Klassengrenze* x_i^u und die *obere Klassengrenze* x_i^o. Da die einzelnen Klassen aneinanderstoßen, ist – bei k Klassen –

$$x_i^o = x_{i+1}^u \qquad (i = 1, \ldots, k-1).$$

Die *Klassenbreite* Δx_i ergibt sich zu:

$$\Delta x_i = x_i^o - x_i^u \qquad (i = 1, \ldots, k).$$

Als repräsentativen Merkmalswert der Klasse i wählt man die *Klassenmitte* x_i' mit

$$x_i' = \frac{1}{2}(x_i^u + x_i^o) \qquad (i = 1, \ldots, k).$$

Bei der **Klasseneinteilung** wird das Ziel verfolgt, die Struktur der untersuchten Gesamtheit möglichst deutlich herauszuarbeiten. Wie viele Klassen dabei gebildet werden sollen, lässt sich nicht generell angeben. Legt man zu viele Klassen zugrunde, dann wird die Verteilung unübersichtlich, weil viele Klassen zu gering oder gar nicht besetzt sind; bei zu wenigen Klassen wird u. U. die charakteristische Form der Verteilung nicht zum Ausdruck kommen. Selbst bei umfangreichem Datenmaterial sollte die Zahl der Klassen 20 nicht übersteigen. Im allgemeinen wird man bestrebt sein, Klassen mit *konstanter Klassenbreite* $\Delta x_i = \Delta x = \text{const.}$ zu bilden; bei einem großen Variationsbereich des Datenmaterials kann es jedoch sinnvoll sein, *unterschiedliche Klassenbreiten* zu verwenden.

Weiterhin ist festzulegen, wie ein Merkmalswert zu behandeln ist, der genau auf eine Klassengrenze fällt. Man kann ihn der niedrigeren oder der höheren oder jeder der beiden Klassen zur Hälfte anrechnen. Darüber hinaus besteht die Möglichkeit, die Klassengrenzen so zu wählen, dass aus messtechnischen Gründen kein Merkmalswert auf sie fallen kann. Will man beispielsweise die Häufigkeitsverteilung von Gebrauchtwagenpreisen ermitteln, und sind die Gebrauchtwagenpreise in vollen €-Beträgen angegeben, so könnte man hier als Klassengrenzen etwa die Werte 500,50 €, 1 000,50 €, 1 500,50 € usw. wählen.

Die Anzahl der Elemente, deren Merkmalswert in die Klasse i (i = 1, . . . , k) fällt, bezeichnet man als *absolute Klassenhäufigkeit* h_i. Die *relative Klassenhäufigkeit* f_i ergibt sich bei k Klassen zu

$$f_i = \frac{h_i}{N} \quad (i = 1, \ldots, k) \quad \text{mit} \quad N = \sum_{i=1}^{k} h_i \,.$$

Bei der grafischen Darstellung der Häufigkeitsverteilung wird das *Histogramm* verwendet. Die absoluten bzw. relativen Klassenhäufigkeiten sind hier den Flächeninhalten der einzelnen Säulen proportional. Bezeichnet man die Höhe der Säule der Klasse i mit h_i^* und mit Δx_i die Säulenbreite, dann gilt – mit a als Proportionalitätsfaktor – die Beziehung:

$$h_i = a \cdot h_i^* \cdot \Delta x_i \qquad (i = 1, \ldots, k).$$

Damit ergibt sich für die Säulenhöhe h_i^*:

$$h_i^* = \frac{h_i}{a \cdot \Delta x_i} \qquad (i = 1, \ldots, k).$$

Besitzen sämtliche Klassen die gleiche Klassenbreite Δx, dann sind die Säulenhöhen den Klassenhäufigkeiten proportional.

Beispiel: Bei der Untersuchung der monatlichen Bruttoverdienste von N = 250 Beschäftigten eines Betriebes werden k = 10 Klassen gebildet, die alle eine konstante Breite von 300 € besitzen. Die Untergrenze der ersten Klasse ist 500 €. Die obere Klassengrenze soll immer zur unteren Klasse gehören. Wären einzelne Beschäftigte mit einem Bruttomonatsverdienst von weit über 3 500 € vorhanden, dann würde man noch eine weitere Klasse als *offene Randklasse* (über 3 500 €) anfügen.

Die Auswertung der Unterlagen der Buchhaltung liefere dann die in Tabelle 2.5 dargestellten Werte.

Bei der grafischen Darstellung der Häufigkeitsverteilung können, da es sich hier um Klassen mit konstanter Breite handelt, in dem Histogramm die Klassenhäufigkeiten als Säulenhöhen verwendet werden (vgl. Abbildung 2.4).

Klasse Nr. i	Bruttomonatsverdienst in €	Klassenbreite in € Δx_i	Beschäftigte Anzahl h_i	Beschäftigte Anteil f_i
1	500 bis 800	300	6	0,024
2	über 800 bis 1 100	300	13	0,052
3	über 1 100 bis 1 400	300	22	0,088
4	über 1 400 bis 1 700	300	32	0,128
5	über 1 700 bis 2 000	300	40	0,160
6	über 2 000 bis 2 300	300	42	0,168
7	über 2 300 bis 2 600	300	39	0,156
8	über 2 600 bis 2 900	300	31	0,124
9	über 2 900 bis 3 200	300	20	0,080
10	über 3 200 bis 3 500	300	5	0,020
$\sum$	.	.	250	1 000

Tabelle 2.5: Häufigkeitsverteilung

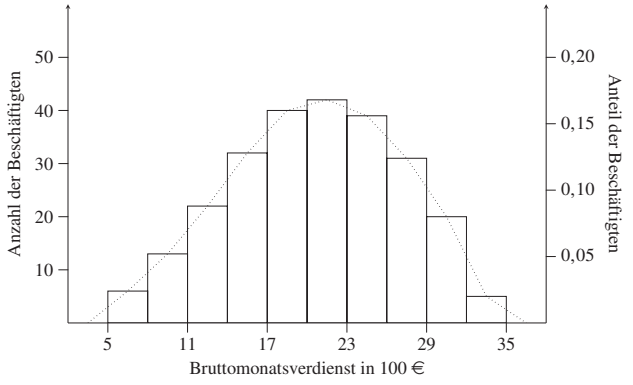

Abb. 2.4: Histogramm mit Klassen konstanter Breite sowie das entsprechende Häufigkeitspolygon (gestrichelter Linienzug)

Bei klassifizierten Daten – wie sie in unserem Beispiel vorlagen – ist eine weitere grafische Darstellung der Häufigkeitsverteilung, nämlich das sogenannte **Häufigkeitspolygon**, möglich. Es entsteht, indem man die Mittelpunkte der oberen Säulenseiten miteinander verbindet (vgl. Abbildung 2.4).

Verwendet man bei der Klassifizierung der Daten Klassen mit unterschiedlichen Klassenbreiten, dann sind die Säulenhöhen des Histogramms nicht mehr den Klassenhäufigkeiten proportional. Man wählt dann bei der Darstellung des Histogramms eine Klassenbreite aus, auf die sich der Maßstab der Ordinate beziehen soll. Zweckmäßigerweise wird man dazu diejenige Klassenbreite heranziehen, die am häufigsten auftritt.

Beispiel: Bei der Inventur eines Gebrauchtwagenlagers, das 70 Personenkraftwagen umfasst, ergeben sich die in Tabelle 2.6 dargestellten Werte.

Klasse Nr. i	Wert in 1 000 €	Klassenbreite in 1 000 € Δx_i	Gebrauchtwagen Anzahl h_i	Gebrauchtwagen Anteil f_i	Gebrauchtwagen Anzahl bez. auf 1 000 € h_i^*
1	1 bis 2	1	8	0,114	8
2	über 2 bis 3	1	10	0,143	10
3	über 3 bis 4	1	16	0,229	16
4	über 4 bis 5	1	15	0,214	15
5	über 5 bis 7	2	10	0,143	5
6	über 7 bis 9	2	8	0,114	4
7	über 9 bis 15	6	3	0,043	0,5
$\sum$	.	.	70	1 000	.

Tabelle 2.6: Häufigkeitsverteilung

Bei der grafischen Darstellung wird man hier für den Maßstab der Ordinate die Klassenbreite $\Delta x = 1000$ € zugrunde legen. Da die ersten vier Klassen eine Klassenbreite von $\Delta x = 1000$ € aufweisen, können hier als Säulenhöhen h_i^* direkt die Klassenhäufigkeiten h_i verwendet werden. Bei der Klasse $i = 5$ entfallen auf eine Klassenbreite von $\Delta x_5 = 2\,000$ € $h_5 = 10$ Elemente, d. h. auf die dem Maßstab zugrundeliegende halb so große Klassenbreite von 1 000 € entfallen $h_i^* = 10 : 2 = 5$ Elemente. Besitzt allgemein eine Klasse i die Klassenbreite $\Delta x_i = c_i \cdot \Delta x$, dann ergibt sich für die zugehörige Säulenhöhe h_i^*:

$$h_i^* = \frac{h_i}{c_i} .$$

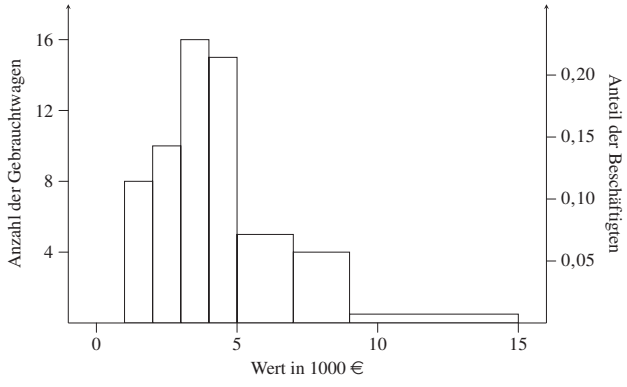

Abb. 2.5: Histogramm mit Klassen unterschiedlicher Breite

2.4 Summenhäufigkeitsfunktion klassifizierter Daten

Wie bei nicht klassifizierten Daten können auch bei klassifizierten Daten durch Kumulierung der Klassenhäufigkeiten h_i die absoluten Summenhäufigkeiten H_i und die relativen Summenhäufigkeiten F_i gebildet werden. Da unbekannt ist, welche

Merkmalswerte die einzelnen Elemente innerhalb einer Klasse besitzen, kann die Summenhäufigkeit H_i bzw. F_i immer *nur der oberen Klassengrenze* x_i^o *zugeordnet* werden.

Aus den relativen Summenhäufigkeiten F_i lässt sich die *Summenhäufigkeitsfunktion* (empirische Verteilungsfunktion) $F(x)$ ableiten. Ihr grafisches Bild erhält man dadurch, dass man in einem Koordinatensystem den oberen Klassengrenzen x_i^o die relativen Summenhäufigkeiten F_i zuordnet. Diese einzelnen Punkte können unter der Annahme, dass die Elemente innerhalb der Klassen gleichmäßig über die gesamte Klassenbreite streuen, linear miteinander verbunden werden. Dieser stetige Linienzug, das grafische Bild der Summenhäufigkeitsfunktion $F(x)$, wird auch als **Summenpolygon** bezeichnet. $F(x)$ gibt wiederum den Anteil der Elemente an, die einen Merkmalswert kleiner oder gleich x besitzen.

Der Wert der Summenhäufigkeitsfunktion $F(x)$ an einer beliebigen Stelle x lässt sich mathematisch bestimmen, indem zunächst die Klasse bestimmt wird, zu der x gehört. So ergeben sich die untere Klassengrenze x_i^u und obere Klassengrenze x_i^o sowie die kumulierten relativen Häufigkeiten $F(x_i^u)$ und $F(x_i^o)$ an den Klassengrenzen. Anschließend wird der Wert von $F(x)$ mittels

$$F(x) = F(x_i^u) + \frac{x - x_i^u}{x_i^o - x_i^u}[F(x_i^o) - F(x_i^u)]$$

berechnet. Diese Formel basiert ebenso wie die grafische Darstellung auf der Annahme, dass die Elemente innerhalb der Klassen gleichmäßig über die gesamte Klassenbreite verteilt sind.

Für das *Beispiel* der Verteilung der Bruttomonatsverdienste erhält man:

Klasse Nr. i	Bruttomonats-verdienst in €	Beschäftigte Kumulierte Anzahl H_i	Beschäftigte Kumulierter Anteil F_i
1	bis 800	6	0,024
2	bis 1 100	19	0,076
3	bis 1 400	41	0,164
4	bis 1 700	73	0,292
5	bis 2 000	113	0,452
6	bis 2 300	155	0,620
7	bis 2 600	194	0,776
8	bis 2 900	225	0,900
9	bis 3 200	245	0,980
10	bis 3 500	250	1,000

Tabelle 2.7: Absolute und relative Summenhäufigkeiten

Die Summenhäufigkeitsfunktion besitzt die in Abbildung 2.6 dargestellte Form.

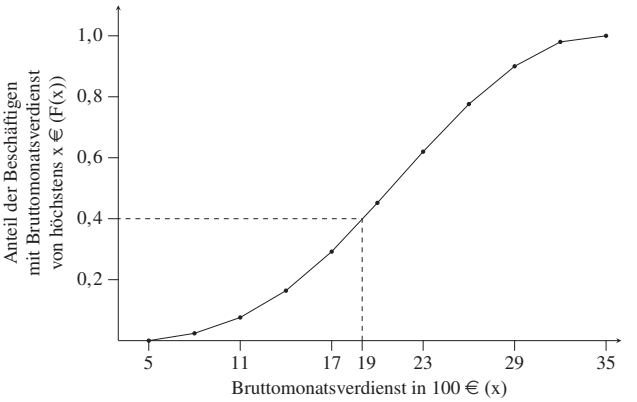

Abb. 2.6: Summenhäufigkeitsfunktion

Interessiert etwa der Anteil der Empfänger, die einen Bruttomonatsverdienst von höchstens x = 1 900 € beziehen, so findet man mithilfe der Summenhäufigkeitsfunktion diesen gesuchten Anteil zu 0,40. Man beachte hierzu, dass x = 1 900 zu Klasse 4 gehört wegen $1700 \leq x = 1900 \leq 2000$. Somit sind $x_i^u = 1700$, $x_i^o = 2000$, $F(x_i^u) = 0{,}292$ und $F(x_i^o) = 0{,}452$ nach Tabelle 2.7. Somit gilt

$$\begin{aligned} F(1900) &= 0{,}292 + \frac{1900 - 1700}{2000 - 1700}(0{,}452 - 0{,}292) \\ &\approx 0{,}292 + 0{,}67 \cdot 0{,}16 \approx 0{,}40 \end{aligned}$$

Für das Beispiel der Verteilung der Gebrauchtwagenpreise erhält man die folgenden Summenhäufigkeiten:

Klasse Nr. i	Wert in 1 000 €	Gebrauchtwagen Kumulierte Anzahl H_i	Gebrauchtwagen Kumulierter Anteil F_i
1	bis 2	8	0,114
2	bis 3	18	0,257
3	bis 4	34	0,486
4	bis 5	49	0,700
5	bis 7	59	0,843
6	bis 9	67	0,957
7	bis 15	70	1,000

Tabelle 2.8: Absolute und relative Summenhäufigkeiten

Das entsprechende Summenpolygon hat dann das folgende Aussehen:

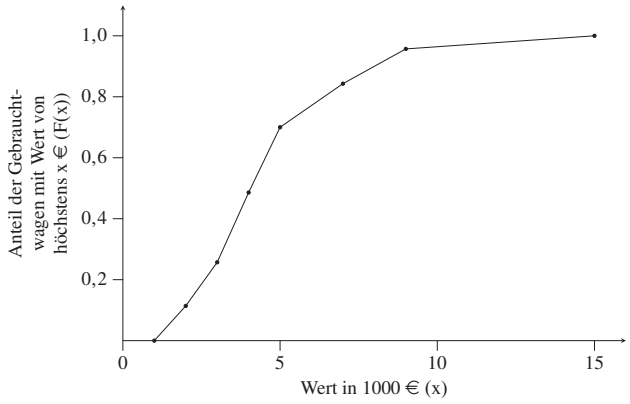

Abb. 2.7: Summenhäufigkeitsfunktion

2.5 Häufigkeitsverteilung zweier Merkmale

Ein Merkmal X habe wieder Ausprägungen $x_1, \ldots, x_r$, hinzu komme nun ein zweites Merkmal Y desselben Merkmalsträgers mit Ausprägungen $y_1, \ldots, y_s$. Die statistische Masse umfasse die N Einheiten $a_l \in \{x_1, \ldots, x_r\}$ und $b_l \in \{y_1, \ldots, y_s\}$ und ist in Tabelle 2.9 dargestellt.

Statistische Einheit l	Merkmals-ausprägung (X)	Merkmals-ausprägung (Y)
1	a_1	b_1
2	a_2	b_2
3	a_3	b_3
⋮	⋮	⋮

Tabelle 2.9: Urliste zweier Merkmale

Man fragt sich nun, wie sich die Merkmalsträger auf die Ausprägungskombinationen der Merkmale verteilen. Beispielsweise wurde im Sommersemester 2019 von zehn Studierenden einer Vorlesung ihre Sportbegeisterung als nominal-skaliertes Merkmal X mit x_1 als „ja“ und x_2 als „nein“ festgestellt. Das Merkmal Y war ihr Geschlecht. Zähle nun h_{ij} das *gemeinsame* Auftreten der Ausprägungen x_i und y_j. Formal ist eine sogenannte Indikatorfunktion $I_{\{a_l=x_i, b_l=y_j\}}$ nur eins, wenn die Bedingung im Index erfüllt ist und es berechnet sich $h_{ij} = \sum_{l=1}^{N} I_{\{a_l=x_i, b_l=y_j\}}$. Zusammen mit den **Randhäufigkeiten**

$$h_{i.} = \sum_{j=1}^{s} h_{ij} \qquad (i = 1, \ldots, r) \quad \text{und}$$

$$h_{.j} = \sum_{i=1}^{r} h_{ij} \qquad (j = 1, \ldots, s)$$

sind die **gemeinsamen Häufigkeiten** in Tabelle 2.10 dargestellt.

Auspräg. von Merkmal Y / Auspräg. von Merkmal X	y_1	…	y_j	…	y_s	$\sum$
x_1	h_{11}	…	h_{1j}	…	h_{1s}	$h_{1.}$
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
x_i	h_{i1}	…	h_{ij}	…	h_{is}	$h_{i.}$
.	.		.		.	.
.	.		.		.	.
.	.		.		.	.
x_r	h_{r1}	…	h_{rj}	…	h_{rs}	$h_{r.}$
$\sum$	$h_{.1}$	…	$h_{.j}$	…	$h_{.s}$	N

Tabelle 2.10: Kontingenztabelle

Die **gemeinsamen relativen Häufigkeiten** sind $f_{ij} = h_{ij}/N$ und lassen sich in einem zweidimensionalen Balkendiagramm wie in Abbildung 2.8 darstellen.

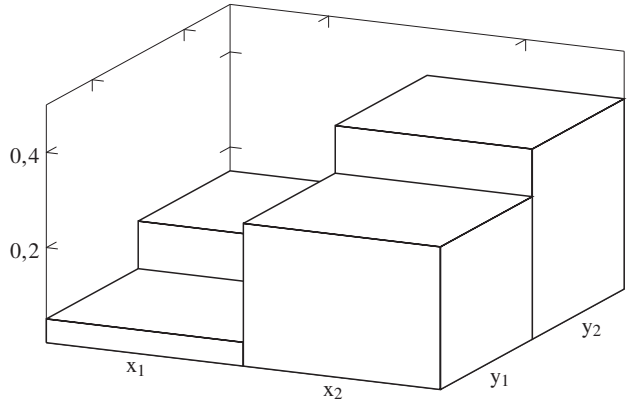

Abb. 2.8: Balkendiagramm für r = s = 2

Die relativen Randhäufigkeiten $f_{i.} = h_{i.}/N$ bzw. $f_{.j} = h_{.j}/N$ entsprechen den Häufigkeiten des jeweiligen Merkmals.

Betrachte das Beispiel der Befragung in Rostock zum Thema Fußballbegeisterung. Die zehn Studierenden wurden konkret gefragt, ob sie Fan des Fußballbundesligisten F.C. Hansa Rostock sind. Zusammen mit dem Geschlecht der Befragten ergab sich Tabelle 2.11.

Die **bedingte (relative) Häufigkeit** gibt den Anteil $f(x_i/y_j)$ mit $i = 1, \ldots, r$ und $j = 1, \ldots, s$ der statistischen Einheiten an, welche die Merkmalsausprägung x_i (für X) besitzen an all jenen, die die Merkmalsausprägung y_j (für Y) haben. Entsprechendes

Hansafan (X) \ Geschlecht (Y)	w/y_1	m/y_2	$f_{i\cdot} = f_X(x_i)$
ja/x_1	0,1	0,2	0,3
nein/x_2	0,3	0,4	0,7
$f_{\cdot j} = f_Y(y_j)$	0,4	0,6	1

Tabelle 2.11: Kontingenztabelle relativer Häufigkeiten

gilt für $f(y_j/x_i)$. Klarerweise ist

$$f(x_i/y_j) = \frac{h_{ij}}{h_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}.$$

Die Berechnung stellt eine Einschränkung der Datenmenge auf eine Merkmalsausprägung dar und ist nicht mit der Randverteilung zu verwechseln. Die absoluten Häufigkeiten gleichen den gemeinsamen Häufigkeiten, lediglich die Gesamtanzahl erhält ein anderes Symbol, d. h. $N \neq h_{\cdot j}$ bzw. $N \neq h_{i\cdot}$.

Mit den Notationen $f_{ij} = f(x_i, y_j)$, $f_{\cdot j} = f_Y(y_j)$ und $f_{i\cdot} = f_X(x_i)$ gilt der **Multiplikationssatz der beschreibenden Statistik**, nämlich, dass die gemeinsame relative Häufigkeit sich darstellen lässt als Produkt von relativer bedingter und Randhäufigkeit, d. h.

$$f(x_i, y_j) = f(x_i/y_j) \cdot f_Y(y_j) = f(y_j/x_i) \cdot f_X(x_i).$$

Klarerweise ist $f(x_i/y_i) = f_{ij}/f_{\cdot j} = f(x_i, y_j)/f_Y(y_j) \Leftrightarrow f(x_i, y_j) = f(x_i/y_j) \cdot f_Y(y_j)$. Im Beispiel ergibt sich etwa

$$\begin{aligned} f(x_1, y_1) &= f(x_1/y_1) \cdot f_Y(y_1) \\ f(\text{ja}, \text{w}) &= f(\text{ja}/\text{w}) \cdot f_Y(\text{w}) \\ 0{,}1 &= 0{,}25 \cdot 0{,}4. \end{aligned}$$

Anschaulich ist die Kenntnis von y_j nur dann informativ für das Merkmal X, wenn sich die relativen Häufigkeiten $f(x_i/y_j)$ für verschiedene j unterscheiden. Gilt also Gleichheit für alle j, und damit auch Gleichheit mit der relativen Randhäufigkeit, d. h. $f(x_i/y_j) = f_X(x_i)$, scheint X von y_j unabhängig. Gilt das für alle i, und aus Gründen der Symmetrie auch umgekehrt für Y abhängig von X, spricht man von Unabhängigkeit. Genauer gilt die **Unabhängigkeit im beschreibenden Sinne** von zwei Merkmalen, wenn sich jede gemeinsame relative Häufigkeit als Produkt der relativen Randhäufigkeiten darstellen lässt, d. h.

$$f(x_i, y_j) = f_X(x_i) \cdot f_Y(y_j) = f_{i\cdot} \cdot f_{\cdot j}.$$

Man könnte auch formulieren, dass man die Bildung der gemeinsamen relativen Häufigkeiten als Produkt *erwartet*, wenn man Unabhängigkeit unterstellt oder vermutet. Im Beispiel ergibt sich etwa

$$\begin{aligned} f(x_1, y_1) &= f_X(x_1) \cdot f_Y(y_1) \\ f(\text{ja}, \text{w}) &= f_X(\text{ja}) \cdot f_Y(\text{w}) \\ 0{,}1 &\neq 0{,}3 \cdot 0{,}4 = 0{,}12. \end{aligned}$$

Allerdings ist aus rein numerischen Gründen zu vermuten, dass es kaum Zahlenkonstellationen gibt, bei denen eine Gleichheit vorliegt. Wann ein Unterschied zum Anlass für eine Entscheidung *gegen* die Abhängigkeit genommen werden kann, wird Gegenstand von Abschnitt 19.2 sein. Grob kann natürlich das Verhältnis

$$\frac{f(x_1/y_1)}{f(x_1/y_2)}$$

über die Stärke der Abhängigkeit (der Assoziation) Auskunft geben. Im Beispiel scheint das Verhältnis der Hansafans bei Frauen und Männern von 0,25/0,33 = 0,75 durchaus erheblich verschieden vom Wert eins, dem Wert bei Gleichheit. Hätten wir allerdings die Zeilen in Tabelle 2.11 vertauscht, wäre das Verhältnis 0, 75/0, 66 = 1, 13 und näher an eins. Um nicht durch die Wahl der Ausprägung x_1 Willkür zu vermitteln, relativiert man mitunter Zähler und Nenner durch den Anteil bei der anderen Ausprägung von X zu einer (bis auf Kehrwertbildung) eindeutigen Charakterisierung

$$\frac{\frac{f(x_1/y_1)}{f(x_2/y_1)}}{\frac{f(x_1/y_2)}{f(x_2/y_2)}} = \frac{\frac{f(x_1/y_1)}{1-f(x_1/y_1)}}{\frac{f(x_1/y_2)}{1-f(x_1/y_2)}}.$$

Im Beispiel ergibt sich ein Wert von 0,66. Wieder spricht ein Wert von eins für Unabhängigkeit.

Für die *absoluten* Häufigkeiten bei N Beobachtungen schlägt sich die Erwartung derart nieder, dass sich h^e_{ij} ergeben aus dem Produkt der Randdichten, multipliziert mit der Anzahl der Elemente N. Somit ist

$$h^e_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{N}.$$

Um erwartete und beobachtete Häufigkeiten deutlich zu unterscheiden, wird für die beobachteten („observed“) auch h^o_{ij} geschrieben.

Wir wollen den Zusammenhang noch in einen anderen Maß darstellen, dass leichter auf mehr Ausprägungen der beiden Merkmals auszuweiten ist. Betrachte hierzu erneut das Beispiel der Beliebtheit eines Fußballclubs (siehe Tabelle 2.1). Zur

		Geschlecht		
Hansafan	ja	1	2	3
	nein	3	4	7
		4	6	10

Tabelle 2.12: Arbeitstabelle

Berechnung des Zusammenhangs werden die *erwarteten Häufigkeiten* den beobachteten gemeinsamen Häufigkeit h^o_{ij} gegenüber gestellt (siehe Tabelle 2.13), die aus den beobachteten

Randhäufigkeiten $h^o_{i\cdot}$ und $h^o_{\cdot j}$ gebildet wird. Der Begriff *Erwartung* bezieht sich hierbei auf die Annahme, dass kein Zusammenhang zwischen den Merkmalen besteht. Je größer die beobachteten von den erwarteten Häufigkeiten abweichen, desto stärker wird Abhängigkeit signalisiert. Ein χ^2 genannter Hilfswert wird anschließend durch die Formel

		Geschlecht		
Hansafan	ja	1,2	1,8	3
	nein	2,8	4,2	7
		4	6	10

Tabelle 2.13: Arbeitstabelle

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}}$$

bestimmt, bei der jede beobachtete Häufigkeit h^o_{ij} zur zugehörigen erwarteten Häufigkeit h^e_{ij} ins Verhältnis gesetzt wird. Bei einer Vierfeldertafel, d. h. r = s = 2, ist vereinfachend

$$\chi^2 = \frac{(h^o_{11}h^o_{22} - h^o_{12}h^o_{21})^2}{h^o_{1\cdot}\,h^o_{2\cdot}\,h^o_{\cdot 1}\,h^o_{\cdot 2}}.$$

Schließlich ist der Kontingenzkoeffizient

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

ein grundlegendes Zusammenhangsmaß für nominale Merkmale. Die Interpretation dieses Zusammenhangsmaßes wird dadurch etwas erschwert, dass sein Wertebereich nicht genau zwischen 0 und 1 liegt, sondern oberhalb beschränkt ist. Durch die rechnerische Anpassung

$$C_{korr} = C \cdot \sqrt{\frac{\min(r, s)}{\min(r, s) - 1}}$$

wird dieser Umstand behoben, sodass der korrigierte Kontingenzkoeffizient C_{korr} tatsächlich den Wertebereich $0 \leq C_{korr} \leq 1$ besitzt. Der Wert 0 drückt die Abwesenheit eines Zusammenhangs und der Wert 1 perfekte Abhängigkeit aus. Im gegebenen Beispiel ergibt sich der Wert

$$\begin{aligned}\chi^2 &= \frac{(1{,}0 - 1{,}2)^2}{1{,}2} + \frac{(2{,}0 - 1{,}8)^2}{1{,}8} \\ &\quad + \frac{(3{,}0 - 2{,}8)^2}{2{,}8} + \frac{(4{,}0 - 4{,}2)^2}{4{,}2} \\ &\approx 0{,}079\end{aligned}$$

und daraus

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \approx 0{,}089$$

Da r = 2 Zeilen und s = 2 Spalten vorliegen, ist min(r, s) = 2. Somit ergibt sich schließlich der **korrigierte Kontingenzkoeffizient**

$$C_{korr} = C \cdot \sqrt{\frac{2}{2 - 1}} = C \cdot \sqrt{2} \approx 0{,}125$$

Der Wert des korrigierten Kontingenzkoeffizienten zeigt, dass kein (oder ein sehr schwacher) Zusammenhang vorliegt.

2.6 Ausgewählte Literatur

Abels, Heiner, Horst Degen, Handbuch des statistischen Schaubilds. Herne, Berlin 1981.

Bertin, Jacques, Graphische Darstellung. Berlin 1982.

Hartung, Joachim, Bärbel Elpelt, Karl-Heinz Klösener, Statistik – Lehr- und Handbuch der angewandten Statistik (14., unwesentl. veränd. Aufl.). München 2005.

Kohler, Heinz, Essentials of Statistics. Glenview (Ill.), London, Boston 1988.

Nagel, M., A. Benner, R. Ostermann, K. Henschke, Grafische Datenanalyse. Stuttgart, Jena, New York 1996.

Schmid, Calvin F., Statistical Graphics – Design, Principles and Practices. New York, Chichester, Brisbane usw. 1983 (Reprint 1992).

Schmid, Calvin F., Stanton E. Schmid, Handbook of Graphic Presentation (2nd ed.). New York, Chichester, Brisbane 1979.

Stange, Kurt, Angewandte Statistik, Teil 1 (2. Aufl.) Berlin, Heidelberg, New York 2001.

Tufte, Edward R., The Visual Display of Quantitative Information (2nd ed.). Cheshire (Connecticut) 2001.

Weißbach, Rafael, Michael Herzog, Gisela Menzel, Regionaler Anteil kariesfreier Vorschulkinder – eine cluster-randomisierte Studie in Südhessen –, AStA Wirtsch Sozialstat Arch 9, 2015, S. 27—39.

Aufgaben zu Kapitel 2

2.1 An einem Bankschalter werden Kundenankünfte (Anzahl der pro 10-Minuten-Zeitintervall ankommenden Kunden) beobachtet. Für 40 derartige Zeitintervalle erhält man folgende Ergebnisse:

0, 0, 1, 3, 4, 1, 2, 2, 1, 1,
1, 2, 3, 0, 2, 0, 1, 3, 1, 2,
2, 0, 1, 1, 6, 1, 0, 2, 3, 1,
1, 4, 2, 3, 2, 0, 3, 0, 1, 2,

Ermitteln Sie absolute und relative Häufigkeiten der Kundenankünfte und stellen Sie Häufigkeitsverteilung und Summenhäufigkeitsfunktion grafisch dar.

2.2 1 000 Motoren eines bestimmten Typs weisen folgende Lebensdauerverteilung auf:

Lebensdauer in Jahren	Anzahl der Motoren
bis 2	33
über 2 bis 4	276
über 4 bis 6	404
über 6 bis 8	237
über 8 bis 10	50

Tabelle 2.14: Lebensdauerverteilung

Stellen Sie diese Häufigkeitsverteilung und die dazugehörende Summenhäufigkeitsfunktion grafisch dar und bestimmen Sie den Anteil der Motoren mit einer Lebensdauer von mehr als 5 Jahren.

2.3 An der Lebensmittelkasse eines Kaufhauses werden die Rechnungsbeträge von 100 Kunden erfasst; es ergibt sich folgende Häufigkeitsverteilung:

Rechnungsbetrag in €	Anzahl der Rechnungen
bis 10	16
über 10 bis 20	48
über 20 bis 40	27
über 40 bis 80	9

Tabelle 2.15: Häufigkeitsverteilung

Häufigkeitsverteilung und Summenhäufigkeitsfunktion des Rechnungsbetrages sind grafisch darzustellen.

2.4 Im Jahr 2009 wurde in Groß-Gerau bei Kindergartenkindern neben ihrem Alter erfasst, ob ihr Gebiss frei von (versorgtem wie unversorgtem) Karies ist (Weißbach et al., 2015). Es ergab sich folgende Häufigkeitsverteilung:

Alter (X) \ kariesfrei (Y)	nein (y_1)	ja (y_2)
3	50	439
4	125	479
5	185	460

Tabelle 2.16: Kontingenztabelle absoluter Häufigkeiten

Es sind die drei Maße aus Abschnitt 2.5 für die Abhängigkeit des Kariesbefalls vom Alter, eingeschränkt auf die 3- und 4-Jährigen, zu berechnen.

3

Mittelwerte

„Wo liegt das Merkmal ungefähr?“ ist eine wichtige Frage bei der prägnanten Darstellung von großen Datenmengen. Diese Frage abhängig vom Merkmalstyp richtig zu beantworten, erfordert es, den Begriff des Mittelwertes genau zu verstehen.

3.1 Einführung

Die im vorangegangenen Kapitel 2 behandelten Häufigkeitsverteilungen stellen eine Zusammenfassung oder Verdichtung der ursprünglich beobachteten Einzeldaten dar. Oft ist man auch noch daran interessiert, *solche Häufigkeitsverteilungen in knapper Form zu charakterisieren*. Dies ist durch die Berechnung **statistischer Maßzahlen** *(Kollektivmaßzahlen, Parameter)* möglich. Zu ihnen gehören die **Mittelwerte** *(Lageparameter)* mit der Aufgabe, das *Zentrum* einer Verteilung zu kennzeichnen, und die **Streuungsmaße** *(Streuungsparameter, Variabilitätsmaße, Variationsmaße)*, welche dazu dienen, die *Streuung (Variabilität, Variation)* der *Einzelwerte um das Zentrum* zu beschreiben (die Streuungsmaße werden in Kapitel 4 behandelt). Daneben gibt es noch weniger wichtige Kollektivmaßzahlen wie die Maße für die **Schiefe** und die **Wölbung**, auf die hier allerdings nicht näher eingegangen werden soll.

Berechnet man Kollektivmaßzahlen für eine **Häufigkeitsverteilung klassifizierter Daten** (vgl. Abschnitt 2.3), dann wird man für jede Klasse einen *repräsentativen Wert*, und zwar meist die entsprechende Klassenmitte x_i' zugrundelegen. Da man auf diesem Wege jedoch nur *Näherungswerte* für die wahren Maßzahlen erhalten kann, sollten diese – soweit möglich – anhand der ursprünglich beobachteten *Einzelwerte* berechnet werden.

In den nachfolgenden Abschnitten werden nur die wichtigsten **Mittelwerte**, nämlich *Arithmetisches Mittel, Median, Modus* und *Geometrisches Mittel* behandelt.

3.2 Arithmetisches Mittel

Das **arithmetische Mittel** sollte sinnvollerweise nur für *metrisch skalierte Merkmale* (vgl. Abschnitt 1.5) berechnet werden. Wenn die N Elemente einer Grundgesamtheit ein bestimmtes metrisch skaliertes Merkmal mit den – nicht notwendigerweise verschiedenen – **Einzelwerten** $a_1, a_2, \ldots, a_N$ aufweisen, dann ist das arithmetische Mittel μ als

$$\mu = \frac{1}{N}(a_1 + a_2 + \ldots + a_N) \quad \text{oder}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} a_i$$

definiert.

Beispiel: Legt ein Angestellter den Weg zwischen Wohnung und Arbeitsstätte an 5 Tagen in 12, 10, 16, 12 und 17 Minuten zurück, dann beträgt die durchschnittliche Zeit, die er für den Weg benötigt,

$$\mu = \frac{1}{5}(12 + 10 + 16 + 12 + 17) = \frac{67}{5}$$
$$= 13{,}4 \text{ Minuten.}$$

Besitzen einige der N Beobachtungswerte $a_1, a_2, \ldots, a_N$ den *gleichen* numerischen Wert, so empfiehlt es sich, sie zusammenzufassen und das arithmetische Mittel aus der entsprechenden **Häufigkeitsverteilung** zu berechnen. Bei k verschiedenen Werten $x_1, x_2, \ldots, x_k$ ergibt sich das arithmetische Mittel dann als

$$\mu = \frac{1}{N}\Big(\underbrace{x_1 + \ldots + x_1}_{h_1\text{-mal}} + \underbrace{x_2 + \ldots + x_2}_{h_2\text{-mal}} + \ldots + \underbrace{x_k + \ldots + x_k}_{h_k\text{-mal}}\Big)$$

oder

$$\mu = \frac{1}{N}(x_1 h_1 + x_2 h_2 + \ldots + x_k h_k)$$

oder schließlich

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i h_i \quad \text{mit} \quad N = \sum_{i=1}^{k} h_i\,;$$

h_i ist dabei die absolute Häufigkeit, mit der der Merkmalswert x_i auftritt.

Da die relativen Häufigkeiten f_i mit den absoluten Häufigkeiten h_i durch die Beziehung

$$f_i = \frac{h_i}{N} \qquad (i = 1, \ldots, k)$$

verknüpft sind, ergibt sich das arithmetische Mittel auch als

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i h_i = \sum_{i=1}^{k} x_i \frac{h_i}{N} \quad \text{oder}$$

$$\mu = \sum_{i=1}^{k} x_i f_i\,.$$

Weil die einzelnen Merkmalswerte x_i bei Vorliegen einer Häufigkeitsverteilung mit den absoluten bzw. relativen Häufigkeiten h_i bzw. f_i gewichtet oder *gewogen* werden, bezeichnet man μ hier als **gewogenes arithmetisches Mittel**.

Für das in Abschnitt 2.1 behandelte *Beispiel* der Häufigkeitsverteilung der Verkaufszahlen einer bestimmten Zeitung (vgl. Tabelle 2.3) ist der Rechengang zur Ermittlung der täglich im Durchschnitt verkauften Zeitungen in der folgenden Arbeitstabelle (Tabelle 3.1) angegeben; bei Verwendung eines geeigneten Taschenrechners kann auf diese Arbeitstabelle natürlich verzichtet werden.

Man erhält

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i h_i = \frac{1}{200}450 = 2{,}25 \quad \text{bzw.}$$

$$\mu = \sum_{i=1}^{k} x_i f_i = 2{,}25\,.$$

i	x_i	h_i	$x_i h_i$
1	0	21	0
2	1	46	46
3	2	54	108
4	3	40	120
5	4	24	96
6	5	10	50
7	6	5	30
$\sum$		200	450

bzw.

f_i	$x_i f_i$
0,105	0
0,230	0,230
0,270	0,540
0,200	0,600
0,120	0,480
0,050	0,250
0,025	0,150
1,000	2,250

Tabelle 3.1: Arbeitstabelle

Bei **Häufigkeitsverteilungen klassifizierter Daten** wählt man als *repräsentativen Wert* für die Einzelwerte der Klasse Nr. i in der Regel die entsprechende Klassenmitte x_i'; die Formel für eine näherungsweise Berechnung des arithmetischen Mittels lautet in diesem Fall

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i' h_i = \sum_{i=1}^{k} x_i' f_i\,,$$

wobei h_i die absolute und f_i die relative Klassenhäufigkeit der i-ten Klasse ist.

Für das bereits aus Abschnitt 2.3 bekannte *Beispiel* der Häufigkeitsverteilung der Bruttomonatsverdienste von 250 Beschäftigten eines Betriebes ergibt sich für den durchschnittlichen Bruttoverdienst (vgl. Tabelle 3.2)

$$\mu = \frac{1}{N}\sum_{i=1}^{k} x_i' h_i = \frac{1}{250}516\,200 = 2064{,}80\ €$$

bzw.

$$\mu = \sum_{i=1}^{k} x_i' f_i = 2064{,}80\ €\,.$$

i	x_i'	h_i	$x_i' h_i$
1	650	6	3 900
2	950	13	12 350
3	1 250	22	27 500
4	1 550	32	49 600
5	1 850	40	74 000
6	2 150	42	90 300
7	2 450	39	95 550
8	2 750	31	85 250
9	3 050	20	61 000
10	3 350	5	16 750
$\sum$		250	516 200

bzw.

f_i	$x_i f_i$
0,024	15,6
0,052	49,4
0,088	110,0
0,128	198,4
0,160	296,0
0,168	361,2
0,156	382,2
0,124	341,0
0,080	244,0
0,020	67,0
1,000	2 064,8

Tabelle 3.2: Arbeitstabelle

Das arithmetische Mittel besitzt **vier wichtige Eigenschaften**, auf die nachfolgend eingegangen werden soll. Für die Ableitungen wird das *ungewogene arithmetische Mittel für Einzelwerte* benutzt; die gewonnenen Aussagen gelten aber, wie man sich leicht überzeugen kann, auch für das *gewogene arithmetische Mittel*.

1. Betrachtet man N Einzelwerte a_i $(i = 1, \ldots, N)$, so ist die *Summe der Abweichungen dieser Einzelwerte von ihrem arithmetischen Mittel* (μ) *gleich Null.*

$$\sum_{i=1}^{N}(a_i - \mu) = \sum_{i=1}^{N} a_i - N\mu = N\mu - N\mu = 0\,,$$

da aus der Definition des arithmetischen Mittels

$$\sum_{i=1}^{N} a_i = N\mu$$

folgt.

2. *Die Summe der quadrierten Abweichungen der Einzelwerte von ihrem arithmetischen Mittel* μ *ist kleiner als von einem beliebigen anderen Wert* M:

$$\sum_{i=1}^{N}(a_i - \mu)^2 < \sum_{i=1}^{N}(a_i - M)^2 \qquad (M \neq \mu)$$

(vgl. Abschnitt 4.2).

3. Werden die Einzelwerte a_i einer *linearen Transformation* (vgl. Abschnitt 1.5)

$$a_i^* = \alpha + \beta a_i \qquad (i = 1, \ldots, N)$$

unterworfen, wobei α und β hier beliebige konstante reelle Zahlen sein können, so findet man für das arithmetische Mittel μ^* der transformierten Werte, wenn μ das arithmetische Mittel

der ursprünglichen Werte bezeichnet,

$$\mu^* = \frac{1}{N}\sum_{i=1}^{N} a_i^* = \frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta a_i)$$
$$= \frac{1}{N}\left(N\alpha + \beta\sum_{i=1}^{N} a_i\right) = \alpha + \frac{\beta}{N}\sum_{i=1}^{N} a_i$$
$$= \alpha + \beta\mu\,.$$

Das bedeutet, dass das *arithmetische Mittel μ der gleichen Transformation unterliegt wie die Einzelwerte* a_i.

Beispiel: Eine Autovermietung berechnet für ihre Wagen eine feste Tagesgebühr von $\alpha = 40$ € und einen Kilometersatz von $\beta = 0{,}40$ €/km; ferner sei bekannt, dass die Wagen täglich im Durchschnitt $\mu = 250$ km zurücklegen. Die durchschnittlichen täglichen Einnahmen pro Wagen (μ^*) ergeben sich dann als

$$\mu^* = 40 + 0{,}40 \cdot 250 = 140 \text{ €}\,.$$

4. Man steht oft vor der Aufgabe, das arithmetische Mittel μ für eine Grundgesamtheit vom Umfang N zu berechnen, die in *zwei oder mehr Teilgesamtheiten aufgeteilt ist, deren Umfänge und arithmetische Mittel bekannt sind*. Beschränken wir uns auf zwei Teilgesamtheiten mit den Umfängen N_1 und N_2 ($N_1 + N_2 = N$), den Einzelwerten a_{1i} ($i = 1, \ldots, N_1$) und a_{2i} ($i = 1, \ldots, N_2$) und den arithmetischen Mitteln μ_1 und μ_2, dann ergibt sich μ als

$$\mu = \frac{1}{N_1 + N_2}\left(\sum_{i=1}^{N_1} a_{1i} + \sum_{i=1}^{N_2} a_{2i}\right)$$
$$= \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}\,,$$

da aus der Definition des arithmetischen Mittels

$$\sum_{i=1}^{N_1} a_{1i} = N_1\mu_1 \quad \text{und} \quad \sum_{i=1}^{N_2} a_{2i} = N_2\mu_2$$

folgt.

Beispiel: Ein Unternehmen besteht aus den beiden Betrieben A und B. Die 400 Beschäftigten von A verdienen monatlich im Durchschnitt 1920,84 € und die 300 Beschäftigten von B monatlich im Durchschnitt 2012,17 €. Der durchschnittliche Bruttomonatsverdienst sämtlicher 700 Beschäftigten von A und B zusammen beträgt dann

$$\mu = \frac{400 \cdot 1920{,}84 + 300 \cdot 2012{,}17}{400 + 300}$$
$$= 1959{,}98 \text{ €}\,.$$

Das arithmetische Mittel ist *nicht immer der aussagekräftigste Mittelwert:* Betrachtet man beispielsweise eine Gruppe von 10 Personen, von denen neun Jahreseinkommen von je 40 000 € beziehen und eine ein Jahreseinkommen von 400 000 € bezieht, so wird das arithmetische Mittel

$$\mu = \sum_{i=1}^{k} x_i f_i$$
$$= 40\,000 \cdot 0{,}9 + 400\,000 \cdot 0{,}1 = 76\,000 \text{ €}$$

zur Charakterisierung des durchschnittlichen Jahreseinkommens dieser Personengruppe wenig sinnvoll sein. Passender wäre der im folgenden Abschnitt 3.3 noch zu behandelnde Median. Man kann nämlich feststellen, dass sogenannte *„Ausreißer"*, d. h. vereinzelte Beobachtungswerte, die sehr weit vom Zentrum der Verteilung entfernt liegen, die Aussagekraft des arithmetischen Mittels erheblich einschränken können. Auch in den Fällen, in denen die Häufigkeitsverteilung kein eindeutiges Zentrum besitzt, etwa im Fall der *bimodalen (zweigipfligen) Verteilung*, kann das arithmetische Mittel nur wenig aussagen.

3.3 Median

Der **Median** *(Zentralwert)* Me ist die *Merkmalsausprägung* desjenigen Elements, das *in der Größe der nach geordneten Beobachtungsreihe in der Mitte* steht. Damit die einzelnen Elemente nach der Größe ihrer Merkmalsausprägungen geordnet werden können, muss das untersuchte Merkmal zumindest *ordinalskaliert* sein. Ordnet man die beobachteten **Einzelwerte** $a_1, \ldots, a_N$ der Größe nach, sodass

$$a_{[1]} \le a_{[2]} \le \ldots \le a_{[N]}$$

gilt, dann ist der Median bei *ungeradem* N

$$\text{Me} = a_{\left[\frac{N+1}{2}\right]}\,.$$

Für das in Abschnitt 3.2 angeführte *Beispiel* eines Angestellten, der an 5 Tagen die benötigte Zeit für den Gang von seiner Wohnung zu seiner Arbeitsstätte festhält, ergibt sich die geordnete Beobachtungsreihe

10 12 12 16 17 und

der Median

$$\text{Me} = a_{\left[\frac{5+1}{2}\right]} = a_{[3]} = 12 \text{ Minuten}\,.$$

Bei einer *geraden* Anzahl von Elementen nimmt man als Median das arithmetische Mittel der beiden mittleren Beobachtungswerte:

$$\text{Me} = \frac{1}{2}\left(a_{\left[\frac{N}{2}\right]} + a_{\left[\frac{N}{2}+1\right]}\right)\,.$$

Liegen die Daten in Form einer **Häufigkeitsverteilung** vor, dann ist der Median diejenige Merkmalsausprägung, bei der die

Summenhäufigkeitsfunktion *den Wert 0,5* überschreitet. – Für das uns bekannte *Beispiel* der Verteilung der Zeitungsverkäufe (vgl. Abschnitt 2.1) ergaben sich die in Tabelle 2.4 wiedergegebenen Summenhäufigkeiten. Man sieht, dass die relativen Summenhäufigkeiten F_i den Wert 0,5 bei 2 Zeitungen überspringen; der Median beträgt hier also

$$\text{Me} = 2 \text{ Zeitungen}.$$

Recht bequem ist dieser Wert auch aus der *grafischen Darstellung der Summenhäufigkeitsfunktion* (vgl. Abbildung 2.3) zu ermitteln. Man zieht eine Parallele zur Abszisse im Abstand 0,5; wo diese die Treppenfunktion schneidet, wird das Lot gefällt, das dann bei Me = 2 auf die Abszisse trifft.

Bei einer **Häufigkeitsverteilung klassifizierter Daten** liegt der Median in derjenigen Klasse, in der die Summenhäufigkeitsfunktion den Wert 0,5 erreicht. Für das in Abschnitt 2.3 betrachtete Beispiel der Bruttomonatsverdienste von 250 Beschäftigten eines Betriebes mit der in Abbildung 2.6 (Abschnitt 2.4) dargestellten Summenhäufigkeitsfunktion ergibt sich eine grafische Näherungslösung für den Median, wie in folgender Abbildung 3.1 dargestellt ist.

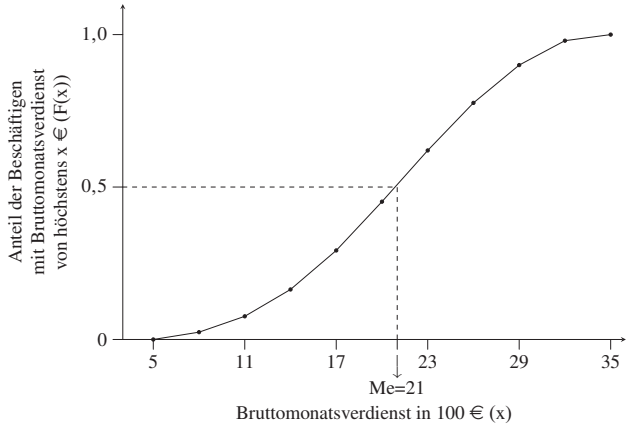

Abb. 3.1: Grafische Bestimmung des Medians bei klassifizierten Daten

Mithilfe linearer Interpolation (Abbildung 3.2) kann darüber hinaus eine *Feinberechnung des Medians* durchgeführt werden.

Es ist $\text{Me} = x_i^u + \alpha$, wenn i *die Einfallsklasse* des Medians ist. Für α findet man anhand der Ähnlichkeitsbeziehungen

$$\frac{\alpha}{x_i^o - x_i^u} = \frac{0{,}5 - F(x_i^u)}{F(x_i^o) - F(x_i^u)}.$$

Damit ist

$$\text{Me} = x_i^u + \frac{0{,}5 - F(x_i^u)}{F(x_i^o) - F(x_i^u)}(x_i^o - x_i^u).$$

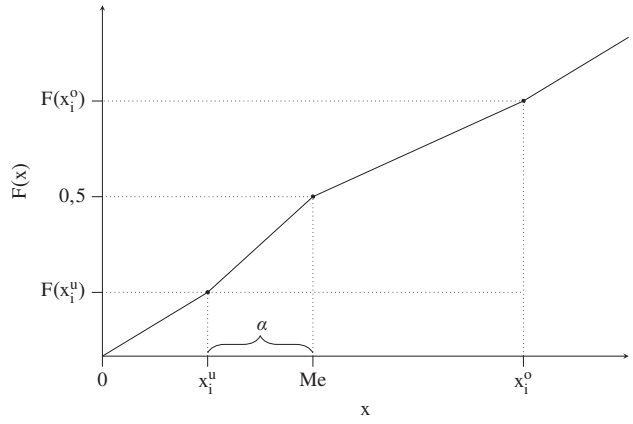

Abb. 3.2: Feinberechnung des Medians

Für das Beispiel der Bruttomonatsverdienste erhält man mit F(2000) = 0,452 und F(2300) = 0,620

$$\begin{aligned}\text{Me} &= 2000 + \frac{0{,}5 - 0{,}452}{0{,}620 - 0{,}452}(2300 - 2000)\\ &= 2000 + 85{,}71 = 2085{,}71 \text{ €}.\end{aligned}$$

3.4 Modus

Der **Modus** *(Dichtester Wert)* Mo ist ein Mittelwert, der sowohl bei *Nominalskalen* und *Ordinalskalen* als auch bei *metrischen Skalen* angewendet werden kann. Der Modus einer **Beobachtungsreihe** ist diejenige *Merkmalsausprägung, die am häufigsten vorkommt*. Für das *Beispiel* der Verteilung der Verkaufszahlen einer bestimmten Zeitung (vgl. Abschnitt 2.1) ergibt sich, da $x_i = 2$ die größte Häufigkeit besitzt, der Modus

$$\text{Mo} = 2 \text{ Zeitungen pro Tag}.$$

Sind die **Daten in Klassen** eingeteilt, dann liegt der Modus in der Klasse mit der größten Klassenhäufigkeit. Wenn man auf eine Feinberechnung des Modus verzichtet, wird man als angenäherten Wert für den Modus die *Klassenmitte* dieser Klasse wählen. Für das *Beispiel* der Bruttomonatsverdienste weist die Klasse Nr. 6 „über 2000 bis 2300 €“ die größte Klassenhäufigkeit auf, sodass der Modus

$$\text{Mo} = x_6' = 2150 \text{ €}$$

ist.

3.5 Geometrisches Mittel

Voraussetzung für die Berechnung des **geometrischen Mittels G** ist, dass das untersuchte Merkmal *verhältnisskaliert* ist. Für

N **Einzelwerte** $a_1, \ldots, a_N$ ist das geometrische Mittel G als

$$G = \sqrt[N]{a_1 \cdot a_2 \cdot \ldots \cdot a_N}$$

definiert.

Liegen die Daten in Form einer **Häufigkeitsverteilung** vor, d. h. treten die verschiedenen Merkmalswerte $x_1, \ldots, x_k$ mit den absoluten Häufigkeiten $h_1, \ldots, h_k$ auf, berechnet sich das geometrische Mittel nach der Formel

$$G = \sqrt[N]{x_1^{h_1} \cdot x_2^{h_2} \cdot \ldots \cdot x_k^{h_k}} \quad \text{mit } N = \sum_{i=1}^{k} h_i \,.$$

Logarithmiert man die Formel für Einzelwerte auf beiden Seiten, so ergibt sich

$$\log G = \frac{1}{N}(\log a_1 + \log a_2 + \ldots + \log a_N)$$

oder

$$\log G = \frac{1}{N} \sum_{i=1}^{N} \log a_i \,;$$

man erhält also das interessante Ergebnis, dass der Logarithmus des geometrischen Mittels gleich dem arithmetischen Mittel der Logarithmen der Einzelwerte ist. Für das aus der Häufigkeitsverteilung berechnete geometrische Mittel ergibt sich entsprechend

$$\begin{aligned} \log G &= \frac{1}{N}(h_1 \log x_1 + h_2 \log x_2 + \ldots + h_k \log x_k) \\ &= \frac{1}{N} \sum_{i=1}^{k} h_i \log x_i \\ &= \sum_{i=1}^{k} f_i \log x_i \,. \end{aligned}$$

Beispiel: Ein Unternehmen erzielte in den Jahren 2007–2011 die in Tabelle 3.3 angegebenen Umsätze (in Mio. €); wie groß war der *durchschnittliche relative Umsatzzuwachs (Zuwachsrate)* pro Jahr?

Nr. i	Jahr	Umsatz (in Mio €)	Zuwachsrate in Prozent des Vorjahresumsatzes p_i	Wachstumsfaktor q_i
1	2007	2,0	.	.
2	2008	2,4	+20,0000	1,200000
3	2009	2,9	+20,8333	1,208333
4	2010	2,7	− 6,8966	0,931034
5	2011	3,1	+14,8148	1,148148

Tabelle 3.3: Umsätze, Zuwachsraten, Wachstumsfaktoren

Man bestimmt zunächst die jährlichen Zuwachsraten p_i in Prozent des Vorjahresumsatzes und daraus die sogenannten jährlichen Wachstumsfaktoren $q_i = 1 + p_i/100$; sodann berechnet man aus ihnen das *geometrische Mittel*

$$\begin{aligned} G &= \sqrt[4]{1{,}200000 \cdot 1{,}208333 \cdot 0{,}931034 \cdot 1{,}148148} \\ &= 1{,}115791 \,. \end{aligned}$$

Die *durchschnittliche Zuwachsrate pro Jahr* ergibt sich zu

$$(1{,}115791 - 1) \cdot 100\,\% = 11{,}5791\,\% \approx 11{,}6\,\% \,.$$

3.6 Ausgewählte Literatur

Harnett, Donald L., James L. Murphy, Introductory Statistical Analysis (3rd ed.). Reading (Mass.), Menlo Park (Cal.), London usw. 1982.

Kreyszig, Erwin, Statistische Methoden und ihre Anwendungen (7. Aufl., 5. unveränd. Nachdruck). Göttingen 1999.

Vogel, Friedrich, Beschreibende und schließende Statistik (13., korr. u. erw. Aufl.). München 2005.

Aufgaben zu Kapitel 3

3.1 Für die Daten der Aufgabe 2.1 sind aus der Häufigkeitsverteilung (vgl. Lösung der Aufgabe 2.1) arithmetisches Mittel, Modus und Median der Kundenankünfte zu ermitteln.

3.2 Welche Werte besitzen Modus und feinberechneter Median der Lebensdauerverteilung aus Aufgabe 2.2?

3.3 Die jährliche Zuwachsrate der Produktion eines bestimmten Haushaltsgerätes entwickelte sich in 5 Jahren wie folgt:

Jahr Nr.	Zuwachsrate in Prozent
1	10
2	20
3	5
4	8
5	15

Tabelle 3.4: Jährliche Produktionszuwachsraten

Wie groß ist die durchschnittliche jährliche Zuwachsrate für den gesamten Zeitraum?

3.4 Man zeige, dass folgendes gilt:

$$\sum_{i=1}^{k} (x_i - \mu) f_i = 0 \,.$$