4 Deskriptive Statistik

INHALTSVORSCHAU

In diesem Abschnitt stellen wir Ihnen Methoden der deskriptiven Statistik vor. Hierzu gehören Kennwerte, mit denen die Lage und die Breite (Streuung) von Verteilungen beschrieben werden können. Anschließend stellen wir Ihnen Methoden zur graphischen Darstellung von Verteilungen vor.

4.1 Die Mittelwerte (Lagemaße)

Die Aufgabe der deskriptiven (beschreibenden) Statistik ist die Datenaufbereitung, d. h. Zusammenfassung. Hierzu werden die Werte durch statistische Kennwerte und graphische Darstellungen beschrieben.

Zusammenfassung von Daten soll Übersichtlichkeit verbessern

Dabei muss beachtet werden, dass durch Zusammenfassen von Daten einerseits mehr Übersichtlichkeit gewonnen wird, andererseits aber damit auch ein teilweise erheblicher Informationsverlust verbunden ist. Deshalb ist die ursprüngliche Datenreihe (Urliste) von großer Bedeutung, sie muss in jedem Fall verfügbar bleiben. So geht bereits durch das Ordnen von Daten die Information über die Reihenfolge, in der die Daten gewonnen wurden, verloren. Messwerte können aber einen Trend enthalten, beispielsweise können sich die Werte in einer Messreihe ändern, wenn sich während der Durchführung der Messungen etwa die Temperatur gleichmäßig ändert (steigend oder fallend). Ein solcher Trend ist nur an der ursprünglichen Reihenfolge der einzelnen Messwerte erkennbar.

Beschreibung einer Datenreihe durch Lage- und Streuungsmaß

Lagemaße und Streuungsmaße sind die wichtigsten Kenngrößen für metrische Daten (→ Kap. 3.3.2, 4.1.4). Während ein Lagemaß die Lage des mittleren Bereichs einer Messreihe bestimmt, beschreiben die Streuungskenngrößen die Streuung einzelner Werte um das Lagemaß. Diese Kennzahlen sollen die Eigenschaften einer Messreihe möglichst gut wiedergeben.

Es können unterschiedliche Mittelwerte gebildet werden, wobei die Auswahl von den Eigenschaften und der Anzahl der Daten abhängig ist.

4.1.1 Der Modalwert

Der Modalwert ist der am häufigsten auftretende Wert.

Das einfachste Lagemaß ist der Modalwert oder Modus $\bar{x}_D$, der den in einer Datenreihe am häufigsten auftretenden Wert beschreibt. Die Bestimmung des Modalwertes ist deshalb auch nur dann sinnvoll, wenn eine relativ große Anzahl an Werten vorliegt. Für nominalskalierte Merkmale (→ Kap. 3.3.1) ist der Modalwert der einzige sinnvolle Lageparameter.

4.1.2 Der Median (Zentralwert)

Der Median bezeichnet denjenigen Wert einer Messreihe, bei dem die der Größe nach geordneten Werte in zwei gleich große Anteile geteilt werden, d. h. oberhalb und unterhalb des Medians liegt die gleiche Anzahl an Werten. Ist die Anzahl n der

Beobachtungswerte $x_1, x_2, \ldots, x_n$ ungerade, so gibt es genau einen mittleren Wert und es gilt für den Median:

$$\tilde{x} = x_{\frac{n+1}{2}}$$ Gleichung 4.1.2–1

Der Median teilt die Datenreihe in zwei gleich große Hälften.

mit x_n als Rangzahl des größten Wertes,
z. B. bei einer geordneten Messreihe von 5 Werten entspricht also der dritte Wert dem Median.

Beispiel 4.1.2–1
Messwerte in einer Urliste mit $n = 5$

Urliste ($n = 5$)	8,7	5,2	6,1	7,9	6,5
Geordnete Messwerte	5,2	6,1	6,5	7,9	8,7
$\tilde{x}$ = mittlerer Wert = 6,5					

Bei einer geraden Anzahl von Werten ist der Median der arithmetische Mittelwert (→ Kap. 4.1.4) der beiden in der Mitte der geordneten Reihe stehenden Werte:

$$\tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$$ Gleichung 4.1.2–2

Beispiel 4.1.2–2
Messwerte in einer Urliste mit $n = 6$.

Urliste ($n = 6$)	5,4	3,9	4,7	4,2	5,8	4,4
Geordnete Messwerte	3,9	4,2	4,4	4,7	5,4	5,8

$$\tilde{x} = \frac{(4{,}4 + 4{,}7)}{2} = 4{,}55 \approx 4{,}6$$

Der Median ist unabhängig von Extremwerten, die z. B. Ausreißer (→ Kap. 9.1) sein können. Er wird deshalb häufig für die Auswertung von Werten, die eine große Streuung aufweisen, wie z. B. In-vivo-Werten (Daten, die am lebenden Organismus (Tier oder Mensch) gewonnen werden), eingesetzt. Außerdem wird er auch bei kleinen Stichproben ($n \leq 4$) als Mittelwert berechnet, da diese häufig schief verteilt sind.

Ausreißer haben auf den Median keinen Einfluss.

4.1.3 Das Quantil

Quantile bezeichnen diejenigen Werte einer Messreihe, die die der Größe nach geordneten Werte nach einem bestimmten Schema unterteilen. Hierbei werden die geordneten Werte in x gleich große Anteile aufgeteilt, wobei sich $x-1$ Schnittstellen ergeben.

Das Quantil gibt an, welcher Wert von einem bestimmten Anteil der Daten nicht überschritten wird.

Quantile sind Kenngrößen, die auf Rangnummern beruhen. Sie stellen sowohl ein Maß für die Lage einer Verteilung als auch für deren Breite dar. Der Median ist ein Beispiel eines Quantils. Anstatt die geordnete Reihe in zwei gleich große Hälften zu zerlegen, kann sie aber auch in vier (Quartile), zehn (Dezile) oder hundert (Perzentile) gleich große Anteile aufgeteilt werden.

Das 1. Quartil ($x_{0,25}$) trennt das untere Viertel von den oberen drei Vierteln der geordneten Daten ab. Das 2. Quartil ($x_{0,5}$) ist identisch mit dem Median: $x_{0,5} = \tilde{x}$. Teilt man eine geordnete Datenreihe nicht in vier, sondern in zehn gleiche Teile, so erhält man als Trennpunkte die Dezile. Es gibt demnach neun Dezile: $x_{0,1}$, $x_{0,2}$,. . ., $x_{0,9}$.

Beispiel 4.1.3–1

Messwerte in geordneter Reihenfolge mit $n = 10$

Messwerte (geordnet)	32	34	35	37	38	39	41	42	43	46
Rang-nummer	1	2	3	4	5	6	7	8	9	10

$\tilde{x} = 38{,}5$; $x_{0,25} = 35$; $x_{0,75} = \underline{\underline{42}}$

Eine allgemeine Regel zur Bestimmung der p-Quantile x_p von geordneten Messreihen metrischer Merkmale des Umfanges n lautet:

Regel zur Bestimmung eines Quantils

Ist das Produkt $n \cdot p$ nicht ganzzahlig, so wird die größte ganze Zahl bestimmt, die kleiner oder gleich $n \cdot p$ ist, zu dieser wird 1 addiert. Die erhaltene Summe ist die Rangnummer desjenigen Messwertes, der gleich x_p ist.

Im obigen Beispiel ($n = 10$, $p = 0{,}25$) ist $n \cdot p = 10 \cdot 20 = 2{,}5$. Die Rangnummer ist $2 + 1 = 3$, somit ist $x_{0,25} = 35$. Entsprechend erhält man $x_{0,75} = 42$, denn es ist $10 \cdot 0{,}75 = 7{,}5$ und $7 + 1 = 8$.

Ist das Produkt $n \cdot p$ ganzzahlig, so ist x_p vereinfacht gleich dem arithmetischen Mittel der beiden Messwerte mit den Rangnummern $n \cdot p$ und $n \cdot p + 1$.

Das 9. Dezil für das obige Beispiel berechnet sich wie folgt:

$n = 10$, $p = 0{,}9$ $\qquad n \cdot p + 1 = 10$ $\qquad x_{0,9} = \dfrac{(43+46)}{2} = \underline{\underline{44{,}5}}$

Der arithmetische Mittelwert 4.1.4

Das am häufigsten verwendete Lagemaß

Das bekannteste und am häufigsten eingesetzte Lagemaß ist der arithmetische Mittelwert. Zu seiner Ermittlung werden die Einzelwerte addiert und die erhaltene Summe durch die Anzahl der Werte dividiert.

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \ldots\ldots\ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Gleichung 4.1.4–1

Da alle Werte in die Berechnung eingehen, wird der arithmetische Mittelwert auch von Extremwerten beeinflusst. Die Bestimmung des arithmetischen Mittels ist nur sinnvoll für metrische Daten (→ Kap. 3.3.2).

Ungeeignet bei schiefen und mehrgipfeligen Verteilungen

Der arithmetische Mittelwert ist bei eingipfeligen, angenähert symmetrischen Verteilungen ein geeignetes Lagemaß, sogar das effizienteste, wenn die Daten normalverteilt (→ Kap. 6) sind. Bei ausgeprägt schiefen Verteilungen (→ Kap. 6.3.3, 6.3.4) oder mehrgipfeligen Verteilungen ist das arithmetische Mittel für die Beschreibung der „durchschnittlichen Lage" einer Verteilung dagegen ungeeignet. Ein häufig zitiertes Beispiel für eine schiefe Verteilung ist die Häufigkeitsverteilung der Einkommen der Bevölkerung in einem Land. Schiefe eingipfelige Verteilungen sind dadurch charakterisiert, dass der größte Teil der Werte auf der einen Seite vom Mittelwert liegt, während eine geringe Anzahl von Werten weit auseinander liegend über die andere Seite verteilt ist. So hatten in Deutschland etwa 82 % der Erwerbstätigen ein Brutto-Jahreseinkommen von bis 50000 €, während der restliche Teil der Bevölkerung ein Einkommen bis zu 5000000 € und mehr hatte (Angaben für 2001, Quelle: Statistisches Bundesamt). Der mittels des arithmetischen Mittelwertes berechnete Durchschnittsverdienst liegt zu hoch. Ein realistisches Bild gibt in diesem Fall der Median. Da die meisten Arbeitnehmer ein „unterdurchschnittliches" Einkommen aufweisen, ist das „Medianeinkommen" kleiner als das arithmetische Mittel der Einkommen.

Beispiel für eine schiefe Verteilung: durchschnittliches Jahreseinkommen

Beschreibt $\bar{x}$ das arithmetische Mittel, $\tilde{x}$ den Median und $\bar{x}_D$ den Modus einer eingipfeligen Häufigkeitsverteilung, so wird diese wie folgt bezeichnet:

Rechtsschief oder linkssteil, wenn:	$\bar{x} > \tilde{x} > \bar{x}_D$
Linksschief oder rechtssteil, wenn:	$\bar{x} < \tilde{x} < \bar{x}_D$
Symmetrisch, wenn:	$\bar{x} = \tilde{x} = \bar{x}_D$

Der geometrische Mittelwert 4.1.5

Ein weiteres, allerdings weniger häufig als der arithmetische Mittelwert eingesetztes Lagemaß ist der geometrische Mittelwert. Hierzu wird aus dem Produkt von n Werten die n-te Wurzel gezogen.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots\ldots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Gleichung 4.1.5–1

Es kann auch die Summe der Logarithmen der Einzelwerte durch die Anzahl der Werte dividiert werden. Das gesuchte geometrische Mittel wird nach Entlogarithmieren des hierbei berechneten Wertes erhalten.

$$\lg \bar{x}_G = \frac{1}{n}\left(\lg x_1 + \lg x_2 + \lg x_3 + \ldots\ldots + \lg x_n\right)$$

Gleichung 4.1.5–2

Mittelwert für relative Änderungen, z. B. Wachstumsprozesse

Der geometrische Mittelwert ist dann ein geeignetes Lagemaß, wenn Merkmalsausprägungen relative bzw. proportionale Änderungen darstellen, z. B. Wachstumsprozesse: Zellzahl im Laufe der Vermehrung von Bakterien, mittlere Zuwachsraten, mittlere Produktionssteigerung, durchschnittliche Zunahme der Bevölkerung in der Zeit.

Beispiel 4.1.5–1

Wachstum von Bakterien

Platte, Nr.	Koloniebildende Einheiten nach 2 Tagen
1	30
2	16
3	64
4	32
5	26
6	54

In diesem Beispiel beträgt der geometrische Mittelwert 33,4, d. h. nach 2 Tagen liegen durchschnittlich 33 koloniebildende Einheiten vor.

4.1.6 Der harmonische Mittelwert

Mittelwert für Verhältniszahlen

Der harmonische Mittelwert wird angewandt, wenn der relevante Parameter der zu mittelnden Größe im Nenner steht, z. B. bei der Bestimmung der mittleren Dichte im Gesamtraum aus einzelnen Dichten von Flüssigkeiten in Teilräumen, bei Frequenzmessungen (Frequenz als Kehrwert der Zeit) oder der Bestimmung einer Durchschnittsgeschwindigkeit aus Geschwindigkeiten für Teilstrecken. Zur Berechnung wird der Quotient aus der Anzahl der Werte und der Summe der reziproken Werte der Einzelwerte gebildet (Gleichung 4.1.6–1).

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots\ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Gleichung 4.1.6–1

Beispiel 4.1.6–1

1 Kilometer wurde mit 30 km/h gefahren, ein weiterer Kilometer mit 60 km/h. Wie groß ist die Durchschnittgeschwindigkeit?

Durchschnittsgeschwindigkeit für die 2 km = $\dfrac{2}{\frac{1}{30}+\frac{1}{60}} = 40 \text{ km/h}$

Das arithmetische Mittel zur Bestimmung von Durchschnittsgeschwindigkeiten führt dann zum richtigen Ergebnis, wenn die gegebenen Geschwindigkeiten sich nicht auf Teilstrecken, sondern auf Teilzeiträume beziehen (Angaben wie Stunden pro Kilometer, anstatt Kilometer pro Stunde).

Die Streuungsmaße 4.2

Streuung der Einzelwerte in einer Verteilung

Mittelwerte sind zwar geeignet, Verteilungen hinsichtlich ihrer Lage zu vergleichen, zeigen aber nicht, wie sich die Werte bzw. deren Häufigkeiten um einen Mittelwert verteilen. Diesem Zweck dienen die Streuungsmaße einer Verteilung.

Beispiel 4.2–1

Es liegen folgende Beobachtungsreihen mit jeweils 3 Werten vor:
a) 499, 500, 501 b) 400, 500, 600 c) 5, 500, 995

In allen drei Fällen beträgt das arithmetische Mittel $\bar{x} = 500$. Die Verteilungen sind dennoch unterschiedlich, da die Werte bei c) sehr viel weiter auseinander liegen als bei b) und diese weiter auseinander liegen als bei a). Die Charakterisierung einer Datenreihe allein durch den Mittelwert ist deshalb nicht ausreichend, zusätzlich muss die Streuung der Werte berücksichtigt werden.
Die Streuung von Beobachtungswerten kann durch unterschiedliche Kenngrößen beschrieben werden.

Die Spannweite 4.2.1

Das einfachste Streuungsmaß ist die Spannweite R. Unter der Spannweite wird die Differenz zwischen dem größten und dem kleinsten Beobachtungswert verstanden.

$$R = x_{i_{max}} - x_{i_{min}} \qquad \text{Gleichung 4.2.1–1}$$

Ein einfach zu bestimmendes, aber wenig aussagekräftiges Streuungsmaß

Die Spannweite ist ein sehr einfach zu bestimmendes, aber wenig aussagekräftiges Streuungsmaß. Sie berücksichtigt nur den größten und kleinsten Wert der Verteilung. Eine Aussage darüber, wie die Werte dazwischen streuen, ist mit der Spannweite nicht möglich. Allerdings ist die Spannweite bei kleinen Stichproben ($n < 10$) ein sehr sinnvolles und häufig eingesetztes Streuungsmaß (→ Kap. 6.5.5).

4.2.2 Die mittlere absolute Abweichung

Die mittlere absolute Abweichung ist ein Streuungsmaß, welches alle Werte einer Verteilung berücksichtigt.

Mittlere absolute Abweichung vom Mittelwert

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \qquad \text{Gleichung 4.2.2–1}$$

Mittlere absolute Abweichung vom Median

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}\right| \qquad \text{Gleichung 4.2.2–2}$$

4.2.3 Die Varianz und die Standardabweichung

Das am häufigsten verwendete Streuungsmaß ist die Varianz s^2 bzw. die Quadratwurzel der Varianz, die Standardabweichung s. Sie stellt die Summe der Quadrate der Abweichungen der Einzelwerte vom Mittelwert, dividiert durch die Zahl der Freiheitsgrade (→ Kap. 6.3.3, 7.8), dar.

Varianz

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \qquad \text{Gleichung 4.2.3–1}$$

Standardabweichung

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} \qquad \text{Gleichung 4.2.3–2}$$

Maß für die Schwankungen der Werte in einer Verteilung

Die Standardabweichung ist ein sehr wichtiges Maß für die Präzision. Ein direkter Vergleich von Standardabweichungen zur Beurteilung der Präzision ist jedoch nicht möglich, da bei größeren Werten in der Regel auch größere Standardabweichungen erhalten werden. Es darf deshalb nicht allein aus einer höheren Standardabweichung auf eine höhere Variabilität der Werte geschlossen werden. Für einen solchen direkten Vergleich muss der Variationskoeffizient berechnet werden.

Der Variationskoeffizient (relative Standardabweichung) 4.2.4

In vielen Fällen ist weniger die Streuung von Messwerten als ihre Relation zum arithmetischen Mittelwert von Interesse. Dieses Verhältnis wird durch den Variationskoeffizienten *CV* (relative Standardabweichung) gemessen, der häufig in Prozentzahlen angegeben wird.

$$CV = \frac{s}{\bar{x}}$$ Gleichung 4.2.4–1

$$CV(\%) = \frac{s}{\bar{x}} \cdot 100$$ Gleichung 4.2.4–2

Der Variationskoeffizient ist somit ein relatives, dimensionsloses Streuungsmaß, das insbesondere zum Vergleich der Streuung von zwei oder mehreren Messreihen eingesetzt wird.

Der Variationskoeffizient ist geeignet Standardabweichungen zu vergleichen.

Der Standardfehler des Mittelwertes 4.2.5

Ein weiteres Streuungsmaß ist der Standardfehler des Mittelwertes $s_{\bar{x}}$:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$ Gleichung 4.2.5–1

s: Stichproben-Standardabweichung von n Einzelwerten, *n:* Stichprobenumfang.
Während die Standardabweichung die in der Grundgesamtheit zu erwartende Streuung der Einzelwerte beschreibt, gibt der Standardfehler des Mittelwertes die Variabilität der Mittelwerte (→ Gleichung 6.6.1–4 bis 6.6.1–6) an. Würden aus einer Grundgesamtheit wiederholt Zufallsstichproben des Umfangs *n* gezogen werden und jeweils der arithmetische Mittelwert berechnet werden, so würde eine Serie von Mittelwerten $\bar{x}_1$, $\bar{x}_2$, … erhalten werden. Haben Einzelwerte x_i aus einer normalverteilten Grundgesamtheit die Standardabweichung σ, so besitzt die Verteilung der Mittelwerte $\bar{x}$ die Standardabweichung $\sigma/\sqrt{n}$.
Der Standardfehler des Mittelwertes ist somit die Standardabweichung der Mittelwerte-Verteilung, von der der beobachtete Mittelwert $\bar{x}$ ein einzelnes Element ist. Er beschreibt die Präzision des geschätzten Mittelwertes und dient hauptsächlich zur Berechnung des Vertrauensbereiches des berechneten Mittelwertes.

Streuung der Stichproben-Mittelwerte um den Mittelwert der Grundgesamtheit

Aus Gleichung 4.2.5–1 ergibt sich eine für die Praxis wichtige Erkenntnis:
Die Präzision einer Schätzung ist umgekehrt proportional zur Quadratwurzel des Stichprobenumfangs. Um z. B. durch Mehrfachmessungen eine doppelte Präzision erhalten zu können, muss der Stichprobenumfang vervierfacht werden.

Der Variationskoeffizient des Mittelwertes (Relativer Standardfehler des arithmetischen Mittelwertes)

Um die Präzision des Mittelwertes, gemessen durch $s_{\bar{x}}$, zu vergleichen, wird, wie beim Variationskoeffizienten, s_x in Beziehung zu $\bar{x}$ gesetzt.

$$CV_{s_{\bar{x}}} = \frac{s_{\bar{x}}}{\bar{x}}$$ Gleichung 4.2.5–2

Oft wird $s_{\bar{x}}$ auch als Prozentanteil von $\bar{x}$ angegeben und als prozentualer Fehler des Mittelwertes bezeichnet.

$$CV_{s_{\bar{x}}}(\%) = \frac{s_{\bar{x}}}{\bar{x}} \cdot 100$$ Gleichung 4.2.5–3

4.2.6 Die Quantilsabstände

Quartilsabstand ist unabhängig von Extremwerten

Als ein weiteres Streuungsmaß ist der Quartilsabstand zu nennen, die Differenz zwischen dem 3.($x_{0,75}$) und 1.($x_{0,25}$) Quartil der geordneten Messwert-Reihe (→ Kap. 4.1.3). Innerhalb des Quartilsabstands liegen 50 % („zentrale 50 %“) der geordneten Messwerte, da unterhalb von $x_{0,75}$ drei Viertel der geordneten Messwert-Reihe und unterhalb von $x_{0,25}$ ein Viertel liegen. Der Quartilsabstand Q wird deshalb auch als Hälftespielraum bezeichnet.

Verallgemeinerungen des Quartilsabstandes ergeben sich, wenn anstelle des ersten und dritten Quartils beliebige Quantile verwendet werden. So liegen (bei umfangreichen Messwert-Reihen) die zentralen 80 % der geordneten Messwerte zwischen dem 10 %- und dem 90 %-Quantil (80 %-Spielraum). Die zentralen 90 % werden dagegen von dem 5 %- und dem 95 %-Quantil eingeschlossen (90 %-Spielraum).

Die ☐ Tab. 4.2.6–1 gibt eine Übersicht über die für die verschiedenen Skalenniveaus geeigneten Lage- und Streuungsmaße.

☐ **Tab. 4.2.6–1** Skalenniveau und zulässige Lage- und Streuungsmaße

Skalenniveau	Lagemaße	Streuungsmaße
Nominalskala	Modalwert	
Ordinalskala	Modalwert Median Quantile	Quartilsabstand sonstige Quantilsabstände
Metrische Skala	Modalwert Median Quantile arithmetischer Mittelwert geometrischer Mittelwert harmonischer Mittelwert	Quartilsabstand sonstige Quantilsabstände Spannweite Varianz Standardabweichung

Graphische Darstellungen von Häufigkeitsverteilungen 4.3

Visualisierung von Daten

Die Zusammenfassung und Darstellung von Daten können in unterschiedlicher Weise erfolgen. Neben der Darstellung in Tabellenform und der zahlenmäßigen Charakterisierung durch statistische Kenngrößen (→ Kap. 4.1) können Daten durch eine graphische Darstellung aufbereitet werden. Hierzu gibt es eine Vielzahl von Möglichkeiten, von denen nur einige wenige, häufig verwendete Darstellungsformen hier vorgestellt werden können.

Bei der Aufbereitung umfangreicher Beobachtungsreihen werden zunächst Klassen gebildet, in denen gleiche oder ähnliche Merkmalsausprägungen zusammengefasst werden. Die Anzahl der Werte in einer einzelnen Klasse wird als Klassenhäufigkeit, Besetzungszahl oder absolute Häufigkeit bezeichnet. Wenn unterschiedliche Beobachtungsreihen miteinander verglichen werden sollen, ist die relative Häufigkeit zu berechnen.

Die Auswahl einer korrekten und geeigneten graphischen Darstellung ist vom Skalenniveau (→ Kap. 3.3.1) der Daten abhängig.

Graphische Darstellung von qualitativen und diskret quantitativen Merkmalen 4.3.1

Das Stabdiagramm

Bei Auftragung relativer Häufigkeiten zum Vergleich mehrerer Beobachtungsreihen sehr gut geeignet

Bei nominalen Daten (→ Kap. 3.3.1) wird die Zuordnung von Merkmalsausprägungen zu verschiedenen Klassen als Klassifikation bezeichnet. Als graphische Darstellung wird häufig ein Stabdiagramm gewählt, wobei die Höhe der Stäbe proportional zu den absoluten bzw. relativen Häufigkeiten in den einzelnen Klassen ist. Die Breite und der Abstand der Stäbe sind frei wählbar. Aus optischen Gründen sollten auch bei nominalskalierten Daten die Abstände gleich gewählt werden, obwohl bei nominalen Daten die Abstände zwischen den Klassen ohne Bedeutung sind. Eine graphische Variante des Stabdiagramm ist das Säulendiagramm, bei dem die Stäbe lediglich durch Rechtecke ersetzt werden, die mittig über die Ausprägungen gezeichnet werden, wobei die Rechteckflächen nicht aneinander stoßen. Eine weitere Variante stellt das Balkendiagramm dar, bei dem Merkmalsausprägungen auf der vertikalen Achse (Ordinate) und die Häufigkeiten auf der horizontalen Achse abgetragen werden.

Beispiel 4.3.1–1

Häufigkeit der Blutgruppen

Blutgruppe	%
A	40
0	40
B	15
AB	5

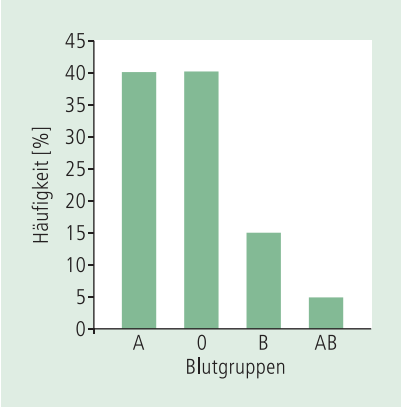

Abb. 4.3.1–1 Relative Häufigkeit der Blutgruppen A, B, 0, AB, dargestellt als Säulendiagramm, Werte aus Beispiel 4.3.1–1

Tab. 4.3.1–1 Schulnoten im Fach Chemie, Physik und Mathematik (Ergebnis einer Umfrage bei Studierenden der Pharmazie im 1. Semester)

Note	Absolute Häufigkeit	Relative Häufigkeit	Absolute Summenhäufigkeit	Relative Summenhäufigkeit
Chemie				
Sehr gut	16	0,18	16	0,18
Gut	54	0,62	70	0,80
Befriedigend	16	0,18	86	0,98
Ausreichend	1	0,02	87	1,00
Physik				
Sehr gut	7	0,09	7	0,09
Gut	37	0,46	44	0,55
Befriedigend	30	0,38	74	0,93
Ausreichend	5	0,06	79	0,99
Ungenügend	1	0,01	80	1,00
Mathematik				
Sehr gut	23	0,24	23	0,24
Gut	35	0,37	58	0,61
Befriedigend	33	0,35	91	0,96
Ausreichend	2	0,02	93	0,98
Ungenügend	2	0,02	95	1,00

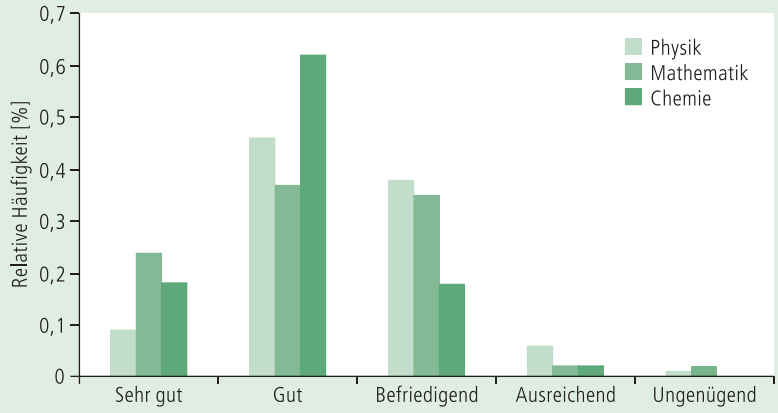

Abb. 4.3.1–2 Relative Häufigkeit von Schulnoten in den Fächern Physik, Chemie und Mathematik, dargestellt als Säulendiagramm (Werte aus Tab. 4.3.1–1)

Bei ordinalen Daten (→ Kap. 3.3.1) ist die Anordnung der Klassen nicht mehr frei wählbar, da die Merkmalsausprägungen einer Rangordnung unterliegen (s. Tab. 4.3.1–1). Die Stablänge gibt die absolute bzw. relative Häufigkeit der Merkmalsausprägungen an.

Die Häufigkeitssummenverteilung

Durch Addition der absoluten oder relativen Häufigkeiten der einzelnen Klassen wird die Häufigkeitssummenverteilung erhalten, aus der unmittelbar der Anteil, der höchstens gleich (kleiner gleich) einem bestimmten Wert ist, abgelesen werden kann, z. B. kann ausgesagt werden, dass 55 % der befragten Studenten (n = 80) in Physik die Note gut oder sehr gut erhalten haben (s. Tab. 4.3.1–1).

Das Kreisdiagramm (Sektordiagramm, Tortendiagramm)

Vergleich relativer Häufigkeiten

Eine weitere Möglichkeit, die Häufigkeit qualitativer (nominale und ordinale) und diskret quantitativer Merkmale graphisch darzustellen, ist ein Kreisdiagramm. Jeder Merkmalsausprägung wird ein Kreissektor zugeordnet. Die Größe eines Kreissektors ist proportional zu den absoluten bzw. relativen Häufigkeiten. In einem Kreisdiagramm wird die Reihenfolge der einzelnen Merkmalsausprägungen nicht wiedergegeben. Bei ordinalen Daten ist deshalb ein Säulen- oder Balkendiagramm dem Kreisdiagramm vorzuziehen. Es eignet sich besonders zur Darstellung von relativen Zahlenverhältnissen.

Der Vollkreis (360°) wird gleich 100 % gesetzt, die prozentuale Winkel-Einteilung (α) erfolgt anhand folgender Gleichung:

$$\alpha = \frac{360° \cdot f_i(\%)}{100\%} = \underline{\underline{3{,}6 \cdot f_i\,(\%)}}$$

Gleichung 4.3.1–1

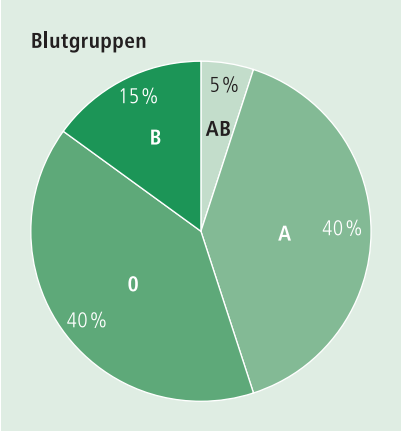

Abb. 4.3.1–3 Relative Häufigkeit der Blutgruppen A, B, 0, AB, dargestellt als Kreisdiagramm (Sektordiagramm), Werte aus Beispiel 4.3.1–1

Werte aus Beispiel 4.3.1–1

Blutgruppe	f_i (%)	α°
A	40	144
0	40	144
B	15	54
AB	5	18

Nachteilig ist, dass beim Vergleich Unterschiede zwischen ähnlich großen Sektoren schwieriger zu erkennen sind als beim Stabdiagramm.

4.3.2 Graphische Darstellung von metrischen Merkmalen

Das Histogramm

Während sich bei nominalen und ordinalen Daten (→ Kap. 3.3.1) die Klassenbildung oft auf Grund natürlicher Unterschiede (z. B. Geschlecht, Blutgruppen) bzw. vorgegebener Abstufungen (z. B. Schulnoten) von selbst ergibt, müssen bei metrischen Daten (→ Kap. 3.3.2) die Klassen künstlich festgelegt werden. Durch die Zuordnung zu Klassen werden Daten zusammengefasst. Dieser Vorgang wird als Klassierung bezeichnet.

Bei der graphischen Darstellung tritt an die Stelle des Stabdiagramms das Histogramm. Beim Histogramm (griech. *histon*: Säule) ist neben der Anordnung und Höhe der Säulen auch deren Breite von Bedeutung, denn der Flächeninhalt repräsentiert graphisch die Klassenhäufigkeit.

Gewinn an Übersichtlichkeit ist mit Informationsverlust verbunden.

Es muss beachtet werden, dass durch das Klassieren von Werten einerseits mehr Übersichtlichkeit gewonnen wird, andererseits damit aber auch ein Informationsverlust verbunden ist. So ist die Reihenfolge, in der die Urdaten gewonnen wurden, nicht mehr erkennbar. Weiterhin ist aus der Häufigkeitsverteilung nicht zu erkennen, wie die einzelnen Werte innerhalb einer Klasse verteilt sind. Für eine Auswertung eines Histogramms muss angenommen werden, dass die Werte zwi-

schen den Klassengrenzen gleichmäßig verteilt sind und durch die Klassenmitte repräsentiert werden. Die Verfälschung der Urdaten ist umso größer, je breiter die Klassen sind.

Beispiel 4.3.2–1

Urliste einer Stichprobe von $n = 100$ Tablettenmassen (mg)

117,0	117,5	118,5	118,6	118,9
120,2	120,4	119,4	119,1	119,4
120,1	120,9	119,7	119,4	119,2
120,9	121,3	121,0	120,1	120,6
121,1	123,9	121,3	121,2	120,8
121,4	125,1	119,9	121,0	120,2
121,5	122,5	120,1	121,5	119,5
120,6	123,5	122,9	123,8	123,7
119,5	122,1	123,4	123,5	121,0
118,6	122,4	122,5	124,5	122,5
121,6	122,6	124,3	119,5	123,2
123,6	122,8	123,6	120,0	122,1
120,5	121,7	121,8	122,1	120,3
120,1	121,6	122,1	122,6	120,4
119,6	121,7	120,5	121,9	121,7
119,8	121,3	120,9	123,6	122,0
121,4	120,7	119,6	120,6	120,5
119,6	124,0	123,4	122,6	120,6
119,4	123,3	123,5	121,7	120,8
121,4	122,4	122,5	122,5	121,6

Festlegung der Klassenbreite

Es muss festgelegt werden, ob Messwerte, die genau auf eine Klassengrenze fallen, der benachbarten unteren oder oberen Klasse zugeordnet werden sollen. Die Wahl ist beliebig, muss dann aber für alle Klassen einheitlich sein. Schlägt man solche Werte jeweils der unteren Klasse zu, so erhält man bei einer Klassenbreite $w = 1$ mg Klassen z. B. 117,6 bis 118,5; man kann dafür eine symbolische Schreibweise verwenden, z. B. (117,5–118,5], (118,5–119,5]. Die runde bzw. eckige Klammer soll bedeuten, dass die daneben stehende Klassengrenze aus- bzw. eingeschlossen wird.

Die graphische Darstellung einer Häufigkeitsverteilung wird wesentlich von der Wahl der Klassenbreiten bestimmt. Um dies zu zeigen, werden für das Beispiel der Tablettenmassen drei unterschiedliche Klassenbreiten gewählt: 0,5 mg (□ Tab. 4.3.2–1), 1 mg (□ Tab. 4.3.2–2) und 2 mg (□ Tab. 4.3.2–3).

Die Klassenbreite bestimmt wesentlich das Aussehen eines Histogramms.

Der Einfluss der Klassenbreite auf das Aussehen des Histogramms ist deutlich zu erkennen. Bei geringer Klassenbreite entfallen auf eine einzelne Klasse oft nur wenige Beobachtungen. Die Folge kann ein unausgewogenes, lückenhaftes Histogramm sein (□ Tab. 4.3.2–1, ○ Abb. 4.3.2–1a). Breitere Klassen führen zu größeren Besetzungszahlen, aber es muss damit auch ein höherer Informationsverlust hingenommen werden. Werden die in der Urliste der Stichprobe von 100 Tabletten aufgeführten Tablettenmassen lediglich in 5 Klassen zusammengefasst, so ergibt

Tab. 4.3.2–1 Stichprobe von $n = 100$ Tablettenmassen (mg, s. Beispiel 4.3.2–1) mit Klasseneinteilung in $k = 18$ Klassen der gleichen Breite von $w = 0{,}5$ mg

Klassen-nummer	Klasse	Absolute Häufigkeit	Relative Häufigkeit	Absolute Summenhäufig-keit	Relative Summen-häufigkeit
1	(116,5–117,0]	1	0,01	1	0,01
2	(117,0–117,5]	1	0,01	2	0,02
3	(117,5–118,0]	0	0,0	2	0,02
4	(118,0–118,5]	1	0,01	3	0,03
5	(118,5–119,0]	3	0,03	6	0,06
6	(119,0–119,5]	9	0,09	15	0,15
7	(119,5–120,0]	7	0,07	22	0,22
8	(120,0–120,5]	12	0,12	34	0,34
9	(120,5–121,0]	13	0,13	47	0,47
10	(121,0–121,5]	10	0,10	57	0,57
11	(121,5–122,0]	10	0,10	67	0,67
12	(122,0–122,5]	11	0,11	78	0,78
13	(122,5–123,0]	5	0,05	83	0,83
14	(123,0–123,5]	7	0,07	90	0,90
15	(123,5–124,0]	7	0,07	97	0,97
16	(124,0–124,5]	2	0,02	99	0,99
17	(124,5–125,0]	0	0,0	99	0,99
18	(125,0–125,5]	1	0,01	100	1,00

sich das in Abbildung 4.3.2–1c dargestellte Histogramm. Für die Anzahl der Klassen und damit für die Wahl der Klassenbreite existieren einige Empfehlungen, z. B. $k = \sqrt{n}$, $k = 2\sqrt{n}$, $k = 10 \log n$. Allerdings sollte auch der subjektive Eindruck, den das Histogramm vermittelt, bei der Auswahl der Klassenbreite mit berücksichtigt werden. So ist in dem dargestellten Beispiel der Tablettenmassen das in Abb. 4.3.2–1b dargestellte mit neun Klassen ($w = 1$ mg) das geeignete.

Häufigkeitssummenverteilung unempfindlicher gegenüber der Wahl der Klassenbreiten

Eine weitere Möglichkeit der graphischen Darstellung von Daten ist die Häufigkeitssummenverteilung (kumulierte Häufigkeitsverteilung), die durch schrittweises Aufsummieren der Besetzungszahlen erhalten wird. Aus den absoluten Häufigkeiten (z. B. Tab. 4.3.2–2) werden schrittweise die absoluten Häufigkeitssummen gebildet. Die relative Häufigkeitssumme wird entsprechend durch Addieren der relativen Häufigkeiten erhalten. Aus der Berechnung bzw. Darstellung der relativen Häufigkeitssumme lässt sich unmittelbar der Anteil derjenigen Werte ermitteln, der kleiner oder gleich einem interessierenden Wert ist, so lässt sich der

Tab. 4.3.2–2 Stichprobe von $n = 100$ Tablettenmassen (mg, s. Beispiel 4.3.2–1) mit Klasseneinteilung in $k = 9$ Klassen der gleichen Breite von $w = 1$ mg

Klassennummer	Klasse	Absolute Häufigkeit	Relative Häufigkeit	Absolute Summenhäufigkeit	Relative Summenhäufigkeit
1	(116,5–117,5]	2	0,02	2	0,02
2	(117,5–118,5]	1	0,01	3	0,03
3	(118,5–119,5]	12	0,12	15	0,15
4	(119,5–120,5]	19	0,19	34	0,34
5	(120,5–121,5]	23	0,23	57	0,57
6	(121,5–122,5]	21	0,21	78	0,78
7	(122,5–123,5]	12	0,12	90	0,90
8	(123,5–124,5]	9	0,09	99	0,99
9	(124,5–125,5]	1	0,01	100	1,00

Tab. 4.3.2–3 Stichprobe von $n = 100$ Tablettenmassen (mg, s. Beispiel 4.3.2–1) mit Klasseneinteilung in $k = 5$ Klassen der gleichen Breite von $w = 2$ mg

Klassen-Nummer	Klasse	Absolute Häufigkeit	Relative Häufigkeit	Absolute Summenhäufigkeit	Relative Summenhäufigkeit
1	(116,5–118,5]	3	0,03	3	0,03
2	(118,5–120,5]	31	0,34	34	0,34
3	(120,5–122,5]	44	0,44	78	0,78
4	(122,5–124,5]	21	0,21	99	0,99
5	(124,5–126,5]	1	0,01	100	1,00

Anteil der Tablettenmassen angeben, der gleich oder kleiner einem vorgegebenen Wert sind, z. B. 47 der untersuchten Tabletten besitzen eine Masse gleich oder kleiner als 121 mg (s. Tab. 4.3.2–1).

Eine andere Darstellungsform der Häufigkeitsverteilung ist der Polygonzug (griech.: *polys*: viel, *gonia*: Winkel). Die Ecken eines Polygonzuges markieren die Häufigkeiten in den einzelnen Klassen, wobei der Kurvenzug jeweils durch die Klassenmitte gelegt wird.

Aus dieser Abbildung lässt sich abschätzen, dass z. B. 50 % aller in der Stichprobe untersuchten Tabletten eine Tablettenmasse gleich oder kleiner als 120,7 mg aufweisen, allerdings unter der Voraussetzung, dass die Tablettenmassen innerhalb der Klasse gleichmäßig verteilt sind (s. Abb. 4.3.2–3).

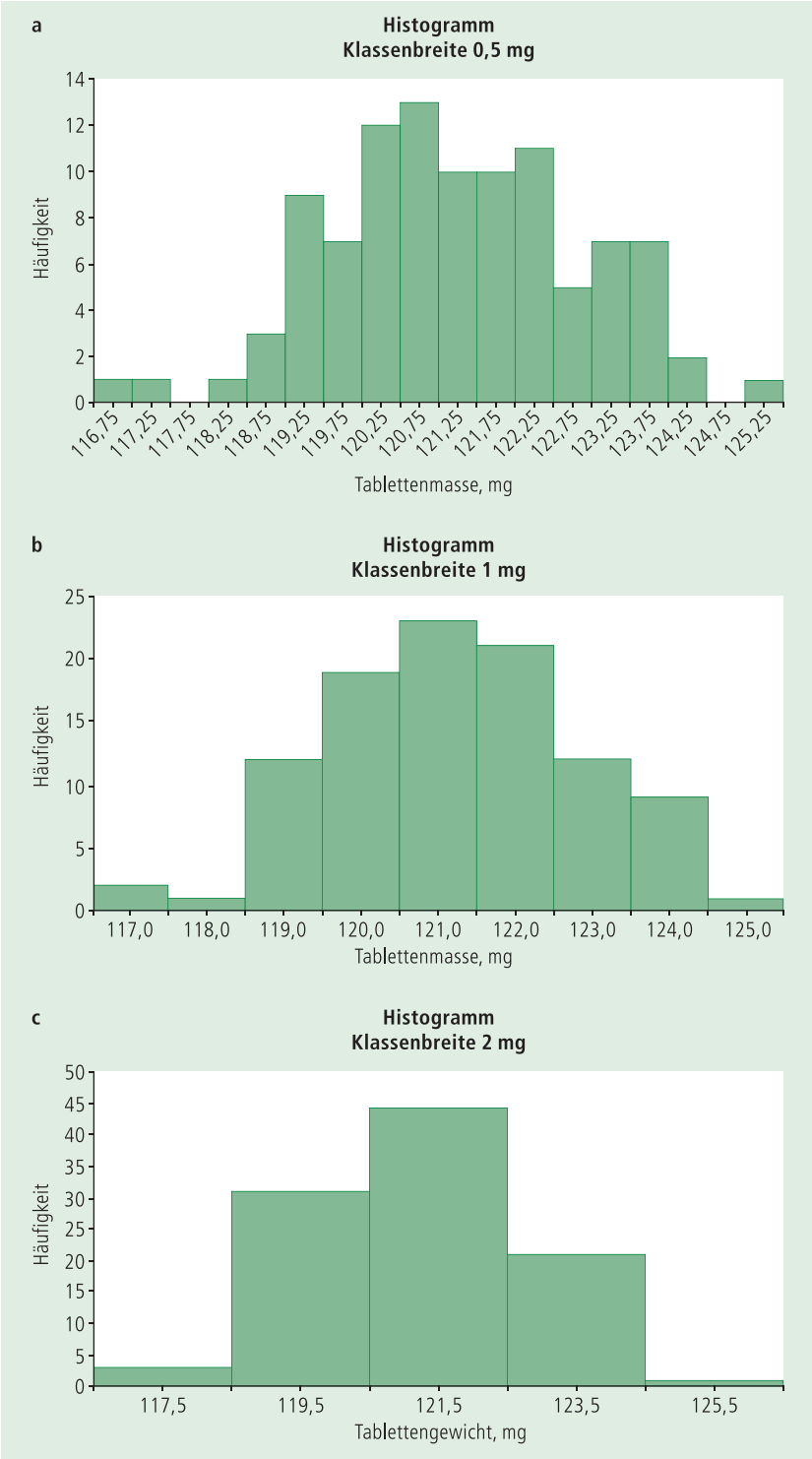

Abb. 4.3.2–1 Histogramm der Tablettenmassen mit unterschiedlichen Klassenbreiten (Beispiel 4.3.2–1, Tab. 4.3.2–1 bis 4.3.2–3)

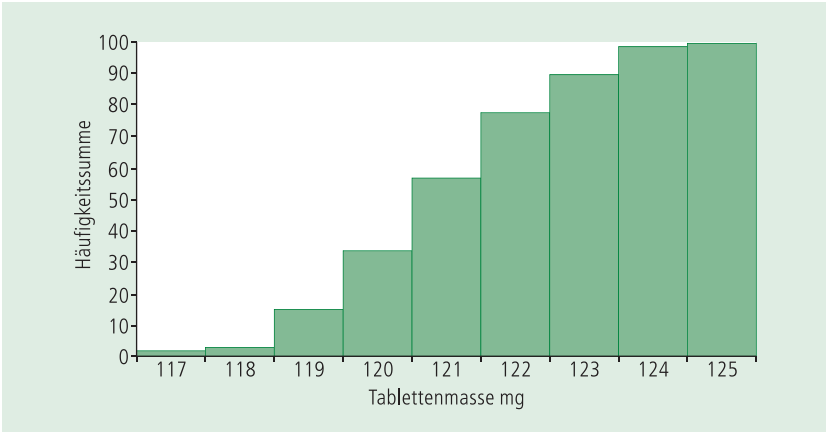

Abb. 4.3.2–2 Häufigkeitssummenverteilung von 100 Tablettenmassen (s. Tab. 4.3.2–2)

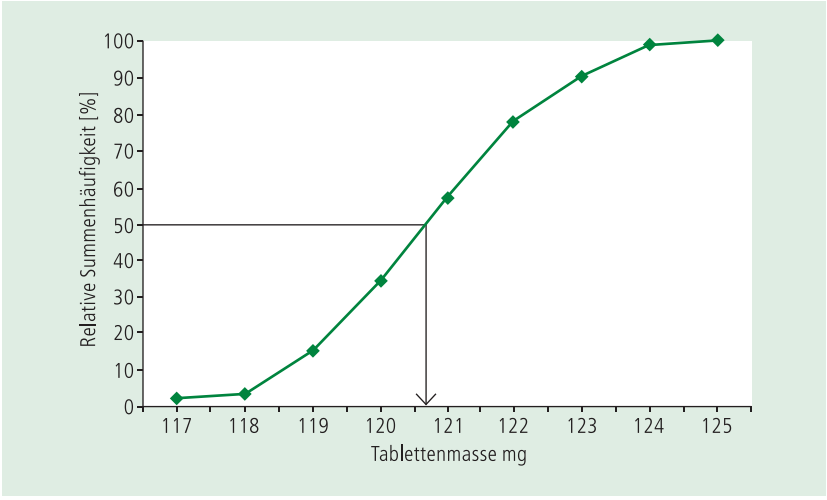

Abb. 4.3.2–3: Polygonzug der Summenhäufigkeitsverteilung von 100 Tablettenmassen (Tab. 4.3.2–2)

Das flächenproportionale Histogramm

Die in einer Leistungskontrolle von den Teilnehmern erzielten Punkte (→ Beispiel 4.3.2–2) sollen in einem Histogramm veranschaulicht werden. Hierzu werden die absoluten Häufigkeiten, wie in der Regel üblich, als Rechtshöhe über den einzelnen Klassen abgetragen.

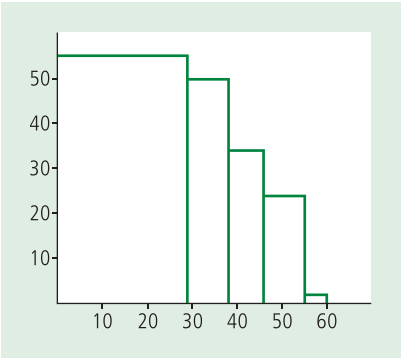

Abb. 4.3.2–4 Histogramm der erreichten Punktzahlen (Werte aus Beispiel 4.3.2–2)

Beispiel 4.3.2–2

In einer Leistungskontrolle wurden von den Teilnehmern folgende Punktzahlen erreicht (Tab. 4.3.2–4):

Erreichte Punktzahl	Häufigkeit
0–29	55
29–38	50
38–46	34
46–55	24
55–60	2

Fehlinterpretation eines Histogramms durch nichtäquidistante Klassenbreiten

Die hierbei erhaltene Darstellung erweckt den Eindruck, als ob die Häufigkeit, mit der 0 bis 29 Punkte erreicht worden sind, wesentlich größer ist als diejenige, mit der 29–38 Punkte erzielt worden sind. Dieser falsche Eindruck entsteht deshalb, weil sich das Auge des Betrachters an der Flächengröße der Rechtecke und nicht an ihrer Höhe orientiert. Die Flächeninhalte der Rechtecke werden mit den absoluten Häufigkeiten der entsprechenden Klassen in Bezug gebracht. Diese nicht realitätsgetreue Darstellung ist auf die unterschiedlichen Klassenbreiten (nichtäquidistant) zurückzuführen, was in der Praxis häufiger auftritt, z. B. bei der Auswertung einer Siebanalyse, bei der die Klassenbreiten durch die Maschenweite der eingesetzten Siebe vorgegeben sind. In solchen Fällen muss die Klassenhäufigkeit auf die Klassenbreite normiert werden. Es gilt dann:

$$Rechteckshöhe = \frac{Klassenhäufigkeit\,[n_i]}{Klassenbreite\,[b_i]}$$ Gleichung 4.3.2–1

Das nun erhaltene Histogramm ist ein flächenproportionales Histogramm. Nur wenn alle Klassen gleich breit (äquidistant) sind, dürfen als Höhen der Rechtecke unmittelbar die absoluten Häufigkeiten der Klassen benutzt werden.

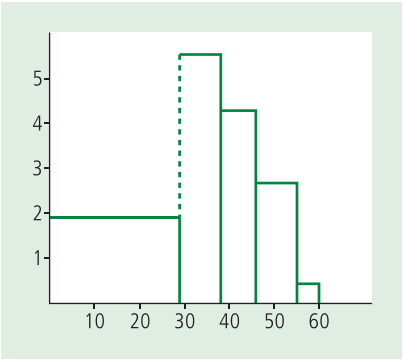

Abb. 4.3.2–5 Flächenproportionales Histogramm für Werte aus Beispiel 4.3.2–2

Tab. 4.3.2–4 Ergebnisse einer Leistungskontrolle (Beispiel 4.3.2–2), normiert auf die-Klassenbreite zur Darstellung als flächenproportionales Histogramm

Erreichte Punktzahl	Häufigkeit $[n_i]$	Klassenbreite $[b_i]$	$\frac{n_i}{b_i}$
0 – 29	55	29	1,897
29 – 38	50	9	5,556
38 – 46	34	8	4,25
46 – 55	24	9	2,667
55 – 60	2	5	0,4

Der Boxplot

4.3.3

Veranschaulichung von Lage, Streuung und Schiefe einer Verteilung

In einem Boxplot (auch Box-and-Whisker-Plot; engl. *plot*: Graph, Diagramm) werden Lage und Streuung einer Messreihe graphisch dargestellt. Er eignet sich insbesondere zum visuellen Vergleich mehrerer Datensätze. Für die Darstellung eines Boxplots können unterschiedliche Lage- und Streuungsmaße ausgewählt werden, so dass es eine Vielzahl von Varianten gibt. Häufig werden Median, Quartilsabstand und Spannweite dargestellt. Zusätzlich kann der arithmetische Mittelwert eingetragen werden. Zur Erstellung des Boxplots werden das 1. Quartil (auch als „hinge“ (Angelpunkt) bezeichnet), das 2. Quartil (Median) und das 3. Quartil bestimmt. Als Box wird ein Rechteck zwischen dem 1. und dem 3. Quartil (Quartilsabstand Q, auch h-spread) gezeichnet. Innerhalb der Box liegen somit 50 % aller Werte (Hälftespielraum). Der Quartilsabstand ist damit ein Maß für die Streuung der Werte. In die Box wird der Median eingetragen. Aus der Lage des Median innerhalb der Box ist erkennbar, ob eine symmetrische oder eine schiefe Verteilung vorliegt; im zweiten Fall besitzen erstes und drittes Quartil (bzw. kleinster und größter Messwert) verschieden große Abstände vom Median.

Als Whisker (engl. *whisker*: Schnurrhaar) werden die horizontalen Linien bezeichnet, deren Länge in unterschiedlicher Weise festgelegt werden kann. Häufig erstrecken sich die whiskers bis zum kleinsten bzw. größten Wert. Der Abstand zwischen den beiden äußeren Enden der Linien beschreibt somit die Spannweite.

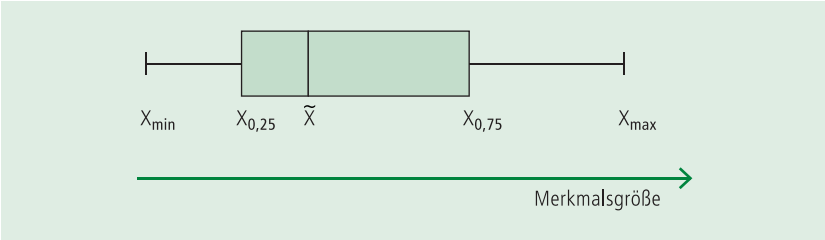

Abb. 4.3.3–1 Boxplot

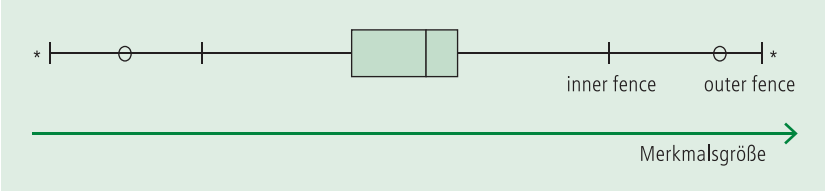

Abb. 4.3.3–2 Modifizierter Boxplot

Eine Modifikation des Boxplots ermöglicht eine Visualisierung von Werten, die als potenzielle Ausreißer in Frage kommen. Die Länge der whiskers wird hierbei durch die sog. inneren und äußeren Zäune (inner fences und outer fences) bestimmt (siehe Abb. 4.3.3–2).
Es gilt:

- Grenzen des inneren Zaunes: $x_{0,25} - 1{,}5\ Q$ und $x_{0,75} + 1{,}5\ Q$.
- Grenzen des äußeren Zaunes: $x_{0,25} - 3\ Q$ und $x_{0,75} + 3\ Q$.

Werte zwischen inner- und outer fence werden häufig einzeln als Kreise (o) markiert. Bei diesen Werten besteht der Verdacht, dass es sich um Ausreißer handelt. Werte, die außerhalb des outer fence liegen, werden als extreme Ausreißer angesehen und häufig durch Kreuze (*) dargestellt.

Beispiel 4.3.3–1
Bei einer Messung der Körpergröße von Frauen und Männern wurden folgende Werte ermittelt, die bereits der Größe nach geordnet sind:
Körpergröße (in cm)
Frauen
152; 154; 154; 155; 156; 157; 158; 158; 160; 160;
163; 164; 166; 166; 168; 168; 169; 170; 172; 174
Männer
161; 165; 170; 171; 173; 173; 174; 174; 174; 175;
175; 176; 176; 177; 177; 180; 183; 184; 184; 192
Beide Messreihen sollen in Form von Boxplots graphisch dargestellt werden.

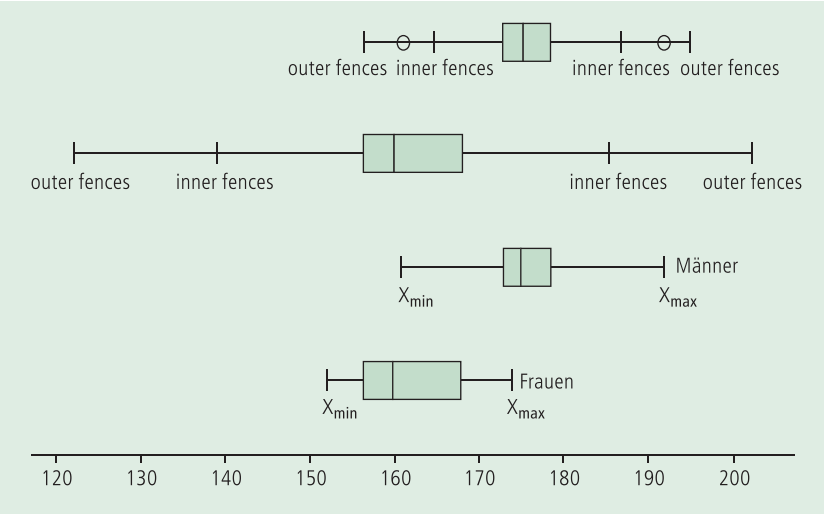

Abb. 4.3.3–3 Boxplot für Körpergröße von Frauen und Männern (Beispiel 4.3.3–1)

Zunächst müssen die hierzu erforderlichen Werte berechnet werden:

Frauen:

$x_{0,25} = 156{,}6$	$\tilde{x} = 160{,}0$	$x_{0,75} = 168{,}0$	$Q = 11{,}5$
inner fences:	$156{,}5 - (1{,}5 \cdot 11{,}5) = 139{,}25$		$168{,}0 + (1{,}5 \cdot 11{,}5) = 185{,}25$
outer fences:	$156{,}5 - (3 \cdot 11{,}5) = 122{,}0$		$168{,}0 + (3 \cdot 11{,}5) = 202{,}5$

Männer:

$x_{0,25} = 173{,}6$	$\tilde{x} = 175{,}0$	$x_{0,75} = 178{,}5$	$Q = 5{,}5$
inner fences:	$173{,}0 - (1{,}5 \cdot 5{,}5) = 164{,}7$		$178{,}5 + (1{,}5 \cdot 5{,}5) = 168{,}75$
outer fences:	$173{,}0 - (3 \cdot 5{,}5) = 156{,}5$		$178{,}5 + (3 \cdot 5{,}5) = 195{,}0$

An den Boxplots ist deutlich erkennbar, dass bei den Frauen im Vergleich zu den Männern

- die Körpergröße geringer ist (Lage des Boxplots),
- die Streuung der Werte geringer ist ($R = 22$, im Vergleich dazu bei den Männern $R = 31$),
- eine linkssteile Verteilung vorliegt.

Der Quartilsabstand ist allerdings mit $Q = 11{,}5$ größer als der entsprechende Wert ($Q = 5{,}5$) bei den Männern. Bei den Frauen liegt kein Wert zwischen dem inneren und dem äußeren Zaun, während bei den Männern mit 161 kg und 192 kg zwei ausreißerverdächtige Werte vorliegen, da sie in diesem Bereich liegen.

Literatur

Burkschat M, Cramer E, Kramps U. Beschreibende Statistik. Springer Verlag, Berlin 2004
Harms V. Biomathematik, Statistik und Dokumentation. 7. Aufl., Harms Verlag, Kiel 1998
Lorenz RJ. Grundbegriffe der Biometrie. 4. Aufl., Gustav Fischer Verlag, Stuttgart 1996
Sachs L, Hedderich J. Angewandte Statistik. 12. Aufl., Springer, Berlin 2006
Timischl W. Qualitätssicherung, Statistische Methoden. 3. Aufl., Hanser Verlag, München 2003
Toutenburg H, Heumann C. Deskriptive Statistik. 5. Aufl., Springer, Berlin 2006

4.4 Übungsaufgaben

Aufgabe 1

Fünfzehn befragte Personen geben ihre monatlichen Ausgaben in € wie folgt an:

1200	300	250	300	3000	1400	700	750
1450	1500	800	900	950	1300	300	

a) Berechnen Sie den arithmetischen Mittelwert, den Median und den Modus!
b) Erklären Sie, warum sich die Lagemaße unterscheiden!
c) Welche Maßzahl charakterisiert Ihrer Meinung nach die Stichprobe am besten?

Aufgabe 2

Berechnen Sie arithmetischen Mittelwert, Median, Spannweite und Quartilsabstand der folgenden Datenreihe:

3	12	6	10	12	14	7	7	5	9

Aufgabe 3

Die Häufigkeitstabelle zeigt die Anzahl fehlerhafter Ampullen, die bei der Qualitätskontrolle an 30 aufeinander folgenden Arbeitstagen beobachtet wurde

Anzahl fehlerhafter Ampullen	0	1	2	3	4	6	7	9
Absolute Häufigkeit	1	3	4	5	8	3	2	4

Stellen Sie die Verteilung in einem Säulendiagramm dar!
Berechnen Sie den Median und den arithmetischen Mittelwert!
Berechnen Sie die mittlere Abweichung der Werte vom Median und Mittelwert!

Aufgabe 4

Die Körpergewichte der Schüler einer Klasse sind nach Geschlechtern aufgeteilt
Körpergewichte der Schüler in kg

w	55	57	63	52	60	62	51	62	62	51	54	58	59	
m	70	73	79	85	68	67	72	70	66	64	61	60	63	71

w: weiblich, m: männlich

Berechnen Sie nach den Geschlechtern getrennt die Spannweite und den Median!
Stellen Sie beide Datenreihen der Körpergewichte in einem Boxplot dar und vergleichen Sie die beiden Darstellungen!