+++ 24 Eier je Legehenne im Juni 2014 +++
(Statistisches Bundesamt, Zahl der Woche vom 7. Oktober 2014)

+++ Männer essen mit 1092 g pro Woche doppelt so viel Fleisch, Fleischerzeugnisse und Wurstwaren wie Frauen +++
(DGE aktuell 01/2014)

+++ Mineralbrunnenbranche: Mineralwasserabsatz auf neuem Rekordniveau +++
(07.01.2015, www.presseportal.de)

+++ Rauchen, trinken, faulenzen raubt Lebenszeit +++
Forscher zeigen: So verlängern Sie Ihr Leben um 17 Jahre
(Focus online, 05.01.2015)

+++ Fast 60 % aller Deutschen essen zu wenig Obst +++
(Max Rubner-Institut, Ergebnisse NVS II – Basisauswertung)

Statistische Kennzahlen für die praktische Anwendung

Deskriptive Statistik

Angela Bechthold, Ute Brehme, Bonn

Dieser Beitrag erläutert anhand von Beispielen grundlegende Begriffe und Kennzahlen der deskriptiven Statistik. Er soll einen ersten Überblick bieten und dabei helfen, Statistik zu verstehen und selbst anzuwenden zu können. Die Darstellung ist daher teilweise vereinfacht und verkürzt. Somit richtet sich dieser Beitrag an alle, die bisher wenig Erfahrung mit Statistik haben und einen Zugang zu diesem nützlichen Werkzeug bekommen möchten.

Statistik – wofür braucht man das?

Daten sammeln, darstellen, analysieren und interpretieren – das ist Statistik und sie ist allgegenwärtig. Sie begegnet uns und wir nutzen sie sowohl im Privatleben als auch im Beruf – einfach überall, wo es etwas zu zählen oder zu messen gibt. Wer schon einmal einen Durchschnitt berechnet hat, hat deskriptive Statistik betrieben.

Mögliche Ausgangsfragen, die durch die Anwendung deskriptiver statistischer Methoden beantwortet werden können, können wie folgt lauten (Beispiel Gewichtsreduktionskurs): Wie viele Personen haben am letzten Kurs teilgenommen? Wie hoch war der Anteil von Frauen und Männern? Entspricht das Durchschnittsalter der Teilnehmer dem der bisherigen Kurse? Wie haben die Teilnehmer den Kurs bewertet? Um wieviel Kilogramm hat sich das Körpergewicht der Teilnehmer am Ende des Kurses im Mittel verändert?

Datenmengen zusammenfassen und beschreiben

Die Methoden der deskriptiven Statistik dienen dazu, durch Experimente, Beobachtung oder Befragung gewonnene Datenmengen zusammenzufassen und auf einen überschaubaren Satz an Charakterisierungen zu reduzieren (◆ Kasten 1).

Ziel der deskriptiven Statistik ist es, die Verteilung und Ausprägung von Merkmalen (Eigenschaften) in einer bestimmten Stichprobe zum Erhebungszeitpunkt der Daten zu beschreiben. Das heißt: Die deskriptive Statistik macht ausschließlich Aussagen zum Datensatz selbst, also über genau jene Untersuchungsobjekte (statistische Einheiten, Merkmalsträger), welche tatsächlich untersucht wurden – das können Menschen in einer klinischen Studie, Äpfel im Laden oder Kliniken in einer Stadt sein. Sie macht keine Aussagen über die Grundgesamtheit (alle Menschen, alle Äpfel), sondern beschreibt die Verhältnisse so wie sie sind (messbare Wirklichkeit). Ausnahme sind repräsentative Stichproben oder Vollerhebungen. Mit der deskriptiven Statistik können so bspw. Gesetzmäßigkeiten erkannt und die theoretische Überprüfung durch die schließende Statistik vorbereitet werden.

Dabei ist ein Merkmal eine Eigenschaft, die zu einem Objekt gehört und eine bestimmte Anzahl an Merkmalsausprägungen (Werte, Messwerte) hat (◆ Abbildung 1). Eine Variable ist ein Merkmal, das – im Unterschied zu einer Konstanten – in mindestens zwei Abstufungen (Messwerten) vorkommen kann. Eine zweistufige Variable wäre bspw. das Geschlecht (männlich/weiblich), eine Variable mit beliebig vielen Abstufungen das Alter.

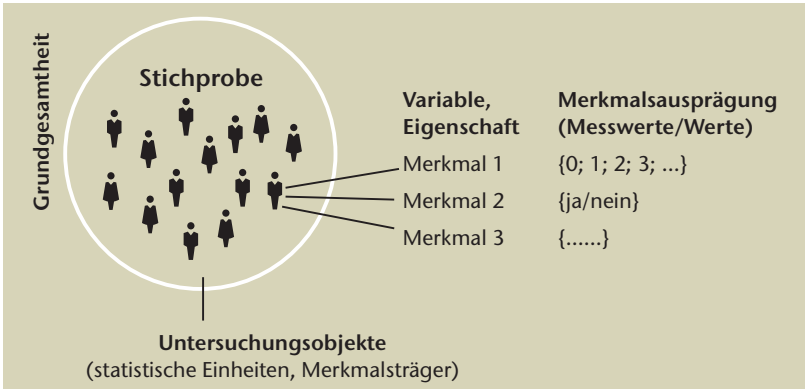


Abb. 1: **Grundgesamtheit, Stichprobe und Variablen/Merkmale**

Art der Daten und Skalenniveaus

An Objekten einer Stichprobe gemessene, beobachtete oder erfragte Daten bilden die Grundlage jeder deskriptiven statistischen Untersuchung. Um diese Daten auswerten zu können, muss klar sein, welcher Art sie sind. Nicht jedes statistische Verfahren passt zu allen Daten. Die verwendeten Verfahren hängen vom Skalenniveau der Daten ab (◆ Tabelle 1).

◆ Tabelle 2 enthält für jeden Klienten eines ambulanten Gewichtsreduktionsprogramms Daten zu neun Variablen. Der Name des behandelnden Arztes und das Geschlecht des Klienten sind Beispiele für **nominale** Daten. Jeder Klient gehört zu einer Kategorie aus einer Menge sich gegenseitig ausschließender Kategorien (Dr. Blum/Dr. Berg/Dr. Holz, weiblich/männlich). Die Kategorien haben keine innere Reihenfolge. Liegen nur zwei Kategorien vor (weiblich/männlich, gesund/krank, ja/nein), so spricht man von dichotomen Variablen. Messen auf Nominalskalenniveau ist gleichzusetzen mit dem Kategorisieren von Objekten; es ist die niedrigste Stufe des Messens, denn die Skalenwerte sind nicht mit quantitativen Ausprägungen der Objekteigenschaften verbunden.

Die Daten zum Rauchverhalten gehören auch zu einer Menge sich gegenseitig ausschließender Kategorien. Jedoch haben die Kategorien eine natürliche, sinnvolle Reihenfolge (Nichtraucher/Gelegenheitsraucher/täglich 1–19 Zigaretten/täglich ≥ 20 Zigaretten). Es handelt sich um **ordinale Daten**. Nur die Reihenfolge der Zahlen ist aussagekräftig, nicht die Zahlen an sich (ein Rauchverhalten mit dem Wert 2 ist nicht „das Doppelte" des Rauchverhaltens mit dem Wert 1), die häufig nur Codierungen sind. Messen auf dem Niveau einer Ordinalskala stellt im Vergleich zur Nominalskala eine höhere Stufe des Messens dar. Unterschiede zwischen den Objekten hinsichtlich der zu messenden Eigenschaft können nicht mehr nur als „gleich" oder „verschieden", sondern auch als „größer" oder „kleiner" charakterisiert werden (◆ Tabelle 1). Daten zum Alter und zum Körpergewicht sind Beispiele für **metrische (quantitative)** Daten. Die möglichen Ausprägungen sind reelle

Statistik und statistische Methoden können grob in zwei Teilgebiete gegliedert werden:

Deskriptive Statistik = Beschreibung von Daten durch Tabellen, statistische Kennzahlen (z. B. Mittelwert, Streuung oder Korrelationskoeffizient) und grafische Darstellung. Die Aussagen treffen nur auf die Objekte der Untersuchung selbst zu, aus der die Daten stammen. Auf die Grundgesamtheit übertragbar sind sie nur im Falle von repräsentativen Stichproben oder Vollerhebungen. Auch beschreibende Statistik genannt.

Schließende Statistik = Ziehen von Schlussfolgerungen. Aus Daten einer Stichprobe wird auf die Grundgesamtheit geschlossen, durch Testen von Hypothesen oder durch Abschätzen von Modellen. Auch prüfende/konfirmatorische/induktive/inferenzielle Statistik genannt.

Kasten 1: **Was versteht man unter deskriptiver Statistik und schließender Statistik?**

Zahlen, und es ist sinnvoll, einen Abstand zwischen zwei Werten zu definieren. Metrische Daten können je nach Messverfahren **diskret bzw. stetig** sein. Variablen wie Alter, Gewicht und Blutdruck von Klienten oder die Temperatur sind reelle Werte und stetig – sie können (zumindest theoretisch) auf einem beliebig genauen Kontinuum beschrieben werden. Zwischen je zwei Werten sind endlich viele Zwischenwerte denkbar. Diskrete Daten können nur eine begrenzte Anzahl möglicher Werte annehmen und es gibt keine Zwischenstufen; ein Beispiel ist die Anzahl der Pulsschläge pro Minute. Ob Werte diskret oder stetig sind, hängt i. d. R. von der Genauigkeit des Messverfahrens ab (bspw. ist die Variable Alkoholkonsum an sich stetig, sie wird aber häufig nur diskret erfasst, z. B. als „nie/1-mal pro Woche/2- bis 3-mal pro Woche usw.).

Metrische Daten können auf Intervall- oder Verhältnisskalenniveau gemessen werden. Auf Intervallskalenniveau gibt es keinen fixen Nullpunkt. Daher ist eine Aussage über das Verhältnis zweier Werte, z. B. Temperaturen („Heute ist es doppelt so warm wie gestern"), nicht sinnvoll, da ihr Wahrheitsgehalt von der Temperaturskala (°Celsius oder °Fahrenheit mit jeweils unterschiedlichen Nullpunkten) abhängt. Im Gegensatz zur Intervallskala besitzt eine Verhältnisskala einen fixen Nullpunkt, der die völlige Abwesenheit der Eigenschaft bezeichnet. Deshalb sind negative Skalenwerte auf einer Verhältnisskala nicht zulässig. Die Existenz eines fixen Nullpunkts erlaubt Aussagen über Verhältnisse. Ist z. B. ein Objekt doppelt so lang wie ein zweites Objekt, dann bleibt diese Aussage korrekt, egal in welcher Einheit die Länge ausgedrückt wird.

Je nach Untersuchungsgegenstand sind Daten also nominal, ordinal oder metrisch und haben damit jeweils ein definiertes Skalenniveau, von welchem wiederum die passenden statistischen Methoden abhängen (◆ Tabelle 1).

Statistische Kennzahlen und Darstellungsformen

Nach Durchführung eines Kurses liegen – wie nach einer Befragung, einem Experiment oder einer epidemiologischen Studie – von einer bestimmten Anzahl an Objekten Merkmalswerte vor. Sie werden in einer Datenmatrix (vgl. Auszug in ◆ Tabelle 2) zusammengestellt. Bei sehr umfangreichen Originaldaten ist es nicht sinnvoll und auch nicht möglich, alle Daten zu betrachten. Hier sind Kennzahlen hilfreich. Sie liefern komprimierte Informationen über die charakteristischen Eigenschaften der untersuchten Stichprobe.

Möglichkeiten zur numerischen Zusammenfassung durch Kennzahlen sind

- **Häufigkeiten.** Sie informieren über die gesamte Verteilung der Daten einer Variablen („Wie häufig ist welcher Wert?").
- **Lagemaße.** Sie charakterisieren das Zentrum bzw. den Schwerpunkt der Daten („Wo liegt die ‚Mitte' der Verteilung der Werte?").
- **Streuungsmaße.** Sie machen eine Aussage über die Breite der Verteilung („Wie stark streuen die Werte?").
- **Korrelations- bzw. Assoziationsmaße.** Sie sagen etwas über den Zusammenhang verschiedener Variablen aus („Wie stark hängen die Werte zweier Variablen zusammen?").

Abhängig von der Anzahl der involvierten Variablen gliedert sich die deskriptive Statistik in drei Bereiche, in denen die genannten Kennzahlen ihre Anwendung finden:

Univariate Statistik

Die Verteilung einer Variablen wird analysiert (man würde sich in ◆ Tabelle 2 bspw. nur die Körperge-

Skalenniveau	Datenniveau	Mögliche Aussagen	Zulässige Kennzahlen (Auswahl)	Geeignete grafische Darstellung (Auswahl)	Beispiele
Nominalskala (verschiedene Kategorien, Zwischenwerte nicht möglich)	nominal (qualitativ)	**gleich oder verschieden**	Häufigkeiten, Modus, Kontingenzkoeffizient	Balkendiagramm, Kreisdiagramm	Geschlecht, Blutgruppe, Nationalität
Ordinalskala (Rangfolge, Zwischenwerte möglich, aber nicht vorhanden)	ordinal (qualitativ)	zusätzlich: **größer oder kleiner**	Häufigkeiten, Median, Interquartilsabstand, Rangkorrelationskoeffizient	Balkendiagramm, Kreisdiagramm, Box-Plot	Noten, soziale Schicht
Intervallskala (Messskala mit konstanten Abständen, kein fixer Nullpunkt)	metrisch (quantitativ)	zusätzlich: **Vergleichbarkeit von Differenzen** (nicht der Werte selbst!)	arithmetisches Mittel, Standardabweichung, Maßkorrelationskoeffizient	Histogramm, Box-Plot	Temperatur, Datum
Verhältnisskala (Messskala mit fixem Nullpunkt)	metrisch (quantitativ)	zusätzlich: **Gleichheit von Verhältnissen**	arithmetisches Mittel, Variationskoeffizient	Histogramm, Box-Plot	Gewicht, Länge, Umsatz

Tab. 1: **Skalenniveaus und ihre Unterscheidungsmerkmale**

Klienten-Nr.	Behandelnde Arztpraxis	Alter des Klienten (Jahre)	Geschlecht 1 = weiblich 2 = männlich	Körpergröße (m)	Körpergewicht (kg) am 04.01.2014	BMI (kg/m²)	Rauchverhalten 0 = Nichtraucher 1 = raucht gelegentlich 2 = tägl. 1–19 Zigaretten 3 = tägl. ≥ 20 Zigaretten	Bewertung des Programms (Note)	Blutdruck systolisch (mm Hg)
K01	Dr. Blum	69	1	1,65	79	29,0	0	1	130
K02	Dr. Blum	51	1	1,76	102	32,9	1	2	178
K03	Dr. Berg	37	1	1,73	78	26,1	0	2	96
K04	Dr. Holz	39	2	1,86	123	35,6	0	2	142
K05	Dr. Holz	41	1	1,59	66	26,1	1	2	96
K06	Dr. Berg	44	2	1,76	91	29,4	2	3	151
K07	Dr. Blum	28	1	1,74	75	24,8	2	4	122
K08	Dr. Blum	22	1	1,67	82	29,4	0	4	142
K09	Dr. Berg	25	2	1,75	119	38,9	3	5	170
K10	Dr. Holz	55	2	1,81	103	31,4	0	2	151
K11	Dr. Berg	47	1	1,58	68	27,2	0	2	87
K12	Dr. Holz	32	1	1,82	110	33,2	0	3	145
K13	Dr. Holz	19	1	1,78	98	30,9	3	2	147
K14	Dr. Blum	46	2	1,89	145	40,6	2	4	181
K15	Dr. Holz	45	1	1,71	103	35,2	0	2	159

Tab. 2: **Auszug aus der Datenmatrix von 31 Teilnehmern eines ambulanten Gewichtsreduktionsprogramms**

wichte anschauen und deren Verteilung über alle Klienten untersuchen). Hierbei kommen Häufigkeiten, Lage- und Streuungsmaße zum Einsatz (s. u.).

Bivariate Statistik

Die Beziehung zweier Variablen zueinander wird analysiert (man würde in ◆ Tabelle 2 bspw. das Geschlecht und die Körpergewichte betrachten – für jeden Klienten hat man so zwei verbundene Informationen und kann erkennen, wie beide Variablen zusammenhängen). Hier kommen u. a. die Korrelations- bzw. Assoziationsmaße ins Spiel (s. u.).

Multivariate Statistik

Die Beziehung von drei und mehr Variablen zueinander wird analysiert (◆ Tabelle 2 als Ganzes ist ein Beispiel für multivariate Daten, sie beinhaltet für jeden Klienten neun miteinander verbundene Informationen, die komplexere Analysen zur Untersuchung des Zusammenhangs erlauben). Hier wird's komplizierter. Der Zusammenhang zwischen multiplen Variablen kann mit Methoden der statistischen Modellierung dargestellt werden. Dies dient weniger der einfachen Beschreibung, sondern ist vielmehr im Kontext der schließenden Statistik erforderlich und daher nicht Teil dieses Beitrags.

Einige der gängigsten Kennzahlen und Darstellungsformen uni- und bivariater Statistik werden im Folgenden vorgestellt.

Häufigkeiten

Am einfachsten lassen sich **nominale und ordinale Daten** numerisch zusammenfassen, indem man Häufigkeiten auszählt. Das heißt, für jede Merkmalsausprägung (z. B. Geschlecht weiblich/männlich, Noten 1–6) wird die Anzahl angegeben (absolute Häufigkeit). Alternativ oder zusätzlich kann man das Verhältnis oder den Anteil der Patienten in jeder Kategorie berechnen (relative und prozentuale Häufigkeit; ◆ Kasten 2).

Für Daten auf Nominal- und Ordinalskalenniveau ist die Berechnung von Häufigkeiten oft die einzige Möglichkeit zur Zusammenfassung; z. B. kann für das Merkmal Geschlecht (Nominalskala) lediglich die Häufigkeit von Männern und Frauen ermittelt werden. Die Berechnung eines „mittleren" Geschlechts ist nicht möglich.

Daten auf **Intervallskalenniveau** lassen sich zu Messwertklassen (Ordinalskala) zusammenfassen, in denen die Häufigkeiten der Messwerte gezählt werden. Bspw. kann der BMI (Intervallskala) zu vier Messwertklassen (Untergewicht/Normalgewicht/Präadipositas/Adipositas) zusammengefasst werden, sodass die Berechnung von Häufigkeiten in einzelnen Klassen möglich wird.

Die Verteilung der Häufigkeiten auf die einzelnen Merkmalsausprägungen (im Beispiel ◆ Kasten 2: Noten) lässt sich zur besseren Übersichtlichkeit auch grafisch darstellen.

Nominale Daten können mithilfe von Kreisdiagrammen oder Balkendiagrammen grafisch zusammengefasst werden. *Ordinale Daten* werden

Die 31 Klienten eines Kurses zur Gewichtsreduktion haben die Inhalte des Programms mit den Noten 1 (sehr gut) bis 6 (schlecht) bewertet. Jede einzelne Bewertung stellt eine Messung dar. Man erhält damit 31 Messwerte der Variable „Bewertung des Programms“. Der Größe nach aufgelistet bilden diese Daten eine geordnete **Urliste (Primärtabelle):**

1,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,6

Bereits erkennbar ist: Die meisten Klienten vergaben die Noten 2 und 3.

Absolute Häufigkeit: Die Primärtabelle lässt sich übersichtlicher darstellen, indem die einzelnen (der Größe nach geordneten) Merkmalsausprägungen mit der Häufigkeit ihres Auftretens versehen werden, so wie es in der zweiten Spalte von ◆ Tabelle 3 gemacht wurde.

Relative Häufigkeit: Die relative Häufigkeit ergibt sich, wenn man die absolute Häufigkeit einer Merkmalsausprägung durch die Gesamtzahl der Objekte (n = 31 Klienten) teilt. Relative Häufigkeiten können nur Werte zwischen 0 und 1 annehmen. Die Summe aller relativen Häufigkeiten ergibt 1 (◆ Tabelle 3, dritte Spalte).

Prozentuale Häufigkeit: Die prozentualen Häufigkeiten (Prozentwerte) ergeben sich, wenn man die relative Häufigkeit einer Merkmalsausprägung mit 100 multipliziert. Prozentuale Häufigkeitsangaben sollten nicht unzulässig genau sein: Wenn die Gesamtzahl n < 100 ist, sollten Anteile ganzzahlig berechnet werden (durch Rundungen können sich vereinzelt Anteilsangaben zu etwas über oder unter 100 % addieren, ◆ Tabelle 3). Am Beispiel in ◆ Tabelle 3, vierte Spalte, ist somit folgende Aussage ablesbar: „39 % der insgesamt 31 Klienten bewerteten den Kurs mit der Note 2.“

Anteilsangaben haben den Vorteil, dass sie Vergleiche mit anderen Datensätzen erleichtern, bei denen die Gesamtzahl der Objekte anders ist. Werden allerdings Anteile (relative und prozentuale Häufigkeit) anders als in ◆ Tabelle 3 ohne Angaben zur absoluten Häufigkeit aufgeführt, so sollte zumindest die Gesamtzahl n der Beobachtungen genannt werden. Bei der Aussage „50 % aller Klienten haben ihr Zielgewicht erreicht“ sollte bekannt sein, ob es sich um 50 % von 100 Klienten oder um 50 % von zwei Klienten handelt.

Merkmalsausprägungen (Note 1 bis 6)	absolute Häufigkeit (Anzahl der Klienten)	relative Häufigkeit = absolute Häufigkeit/31 (Anteil der Klienten)	prozentuale Häufigkeit = relative Häufigkeit×100 (%-Anteil der Klienten)
1	1	0,0323	3
2	12	0,3871	39
3	10	0,3226	32
4	5	0,1613	16
5	2	0,0644	6
6	1	0,0323	3
gesamt	**n = 31**	**1**	**~100**

Tab. 2: **Häufigkeitsangaben am Beispiel der Verteilung der Variable „Bewertung des Programms“ bei 31 Klienten eines Kurses zur Gewichtsreduktion**

Kasten 2: **Beispiel für die Berechnung von Häufigkeiten**

mit den gleichen Methoden zusammengefasst wie nominale Daten, nur werden in den Tabellen und Diagrammen die Kategorien ordinaler Variablen ihrer natürlichen Reihenfolge entsprechend aufgeführt.

Kreisdiagramm

Ein Kreisdiagramm ist eine effektive Methode, um die Anteile der Beobachtungen verschiedener Merkmalsausprägungen (Kategorien) zu veranschaulichen (◆ Abbildung 2). Die Beträge der relativen Häufigkeiten werden als Segmente eines Kreises (Tortenstück) dargestellt. Jede Kategorie wird durch einen Kreisabschnitt repräsentiert, wobei die Fläche des Abschnitts dem Anteil der Beobachtungen in dieser Kategorie entspricht.

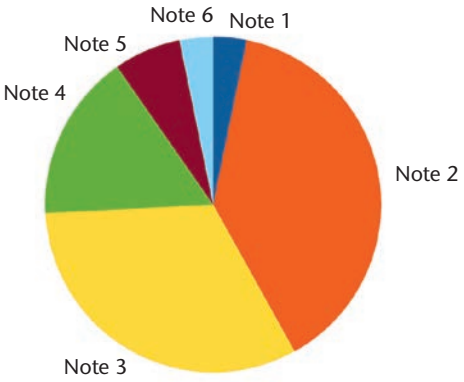


Abb. 2: **Kreisdiagramm am Beispiel der Verteilung der Variable „Bewertung des Programms“ bei 31 Klienten eines Kurses zur Gewichtsreduktion**

Balkendiagramm

In einem Balkendiagramm (Säulendiagramm) gehört zu jeder Merkmalsausprägung (Kategorie) ein Balken; die Höhe der einzelnen Balken stellt die Häufigkeit der Merkmalsausprägung dar. Die Zahlenwerte der Häufigkeiten sollten angegeben werden – z. B. wie in ◆ Abbildung 3 mit einer Skala. Manchmal wird die Skala unterbrochen, wenn die Differenz zwischen zwei Häufigkeiten verglichen mit den absoluten

Häufigkeitswerten sehr klein ist. Das führt dazu, dass die Differenz zwischen den Kategorien größer erscheint als sie ist. Solche Unterbrechungen sollten deshalb möglichst vermieden werden. Werden sie dennoch benutzt, muss die Unterbrechung der Skala und der Balken deutlich gekennzeichnet werden.
Alle Balken sollten die gleiche Breite haben, üblicherweise sind die Balken voneinander getrennt. Werden Daten auf *Nominalskalenniveau* als Häufigkeiten in einem Balkendiagramm dargestellt, so ist ein Abstand zwischen den einzelnen Balken beizubehalten. Ein solches Balkendiagramm wird auch als **Stabdiagramm** bezeichnet; die einzelnen Klassen haben keine Berührungspunkte. Balkendiagramme sind sinnvoll, wenn es nur wenige unterschiedliche Merkmalsausprägungen bzw. Kategorien gibt.

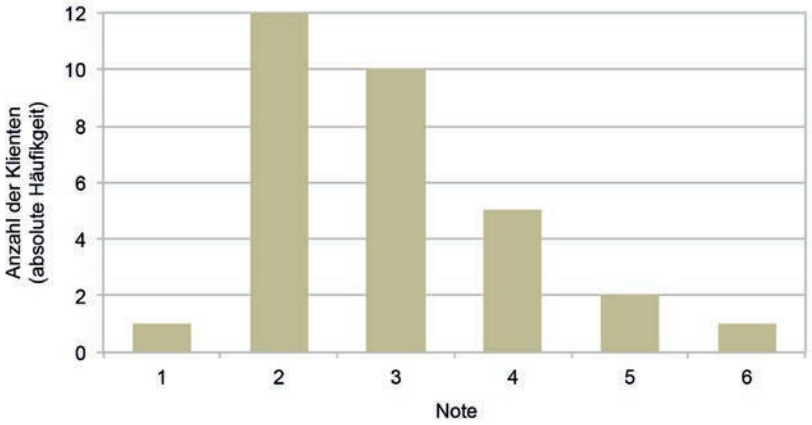


Abb. 3: **Balkendiagramm am Beispiel der Verteilung der Variable „Bewertung des Programms" bei 31 Klienten eines Kurses zur Gewichtsreduktion**

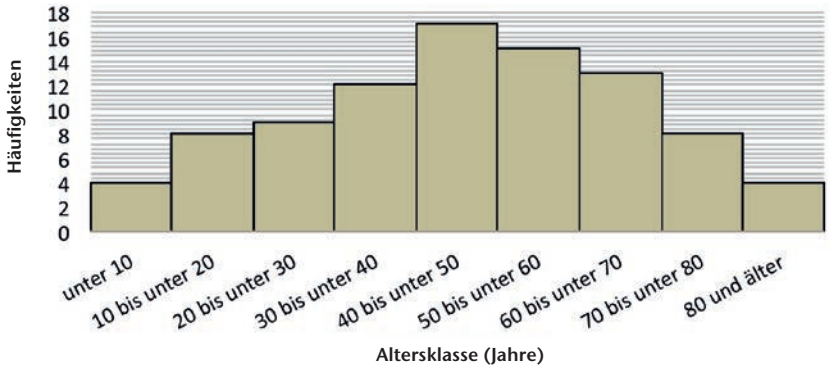


Abb. 4: **Histogramm mit zugehöriger Häufigkeitstabelle für in Klassen eingeteilte Altersangaben von 90 Personen**

Werden *metrische Daten* in Klassen eingeteilt, so kann bei der grafischen Umsetzung auf den Abstand zwischen den Balken verzichtet werden. Ein solches Balkendiagramm wird auch als **Histogramm** bezeichnet (◆ Abbildung 4). Am Ende der einen Klasse beginnt direkt die nächste Klasse. Jede Klasse wird durch ein Rechteck repräsentiert, dessen Breite dem Wertebereich des Intervalls (z. B. „10 bis unter 20 Jahre") entspricht. Die Fläche, die von den Rechtecken eingeschlossen wird, repräsentiert die Gesamtzahl der Häufigkeiten. Von der gewählten Klasseneinteilung (hierzu gibt es keine zwingenden Vorgaben zur Anzahl und Breite) hängt ab, wieviel Information transportiert und wie aussagekräftig das Histogramm ist. Die Klassen werden in Histogrammen häufig durch die Klassenmitte gekennzeichnet (statt „10 bis unter 20 Jahre" also z. B. „15 Jahre"). Die Darstellung in Form eines Histogramms ist weit verbreitet und gibt einen guten Überblick über die Verteilung stetiger Daten.
Wenn ein Histogramm auf einer sehr großen Anzahl von Messungen beruht, kann deren Wertebereich (theoretisch) in relativ schmale Intervalle unterteilt werden, wobei der Umriss der Rechtecke näherungsweise die Form einer glatten Kurve annimmt. Diese Kurve zeigt die Häufigkeitsverteilung (symmetrisch oder asymmetrisch; s. u.).

Lagemaße

Lagemaße geben Auskunft über die Zentralität, also den typischen (Modus), den zentralen (Median) oder den durchschnittlichen (arithmetisches Mittel) Wert einer Verteilung.

Der Modus (Modalwert) ist der am häufigsten gemessene Wert

Der Modus einer Verteilung ist derjenige Messwert, der am häufigsten vorkommt. Um ihn zu bestimmen, muss nur für jeden beobachteten Rohwert ausgezählt werden, wie oft er in der Stichprobe vertreten ist. Am Beispiel der Verteilung der Variable „Bewertung des Programms" (◆ Tabelle 3) wäre dies die Note 2.
Der Modus wird kaum verwendet, da sich schnell Probleme bei seiner Bestimmung ergeben können. Beispielsweise, wenn jeder Wert nur einmal beobachtet worden ist oder ein Datensatz zwei verschiedene Werte gleich häufig enthält (bimodale Verteilung). Darüber hinaus hat der Modus den Nachteil, dass er über vergleichbare Stichproben hinweg sehr variabel ist, d. h. sehr unterschiedliche Werte annehmen kann.
Der Modus ist anwendbar ab nominalem Messniveau. Seine Angabe ist eigentlich nur bei Merkmalen sinnvoll,

die kategorisiert erfasst oder nachträglich in Kategorien eingeteilt wurden.

Der Median ist der Messwert in der Mitte

Der Median ist der Wert, der eine der Größe nach geordnete Messwertreihe halbiert. Er kennzeichnet auf einfache Weise die Mitte der Stichprobenwerte, da die Hälfte (50 %) der Werte kleiner und die andere Hälfte der Werte größer ist als der Median. Der Median entspricht dem 50. Perzentil (◆ Kasten 3).
Der Median wird ermittelt, indem alle Daten in aufsteigender Reihenfolge geordnet werden und dann der mittlere Wert ausgewählt wird (◆ Kasten 4). Wenn die Gesamtzahl der Daten in einem Datensatz eine gerade Zahl ist (also nicht ein Wert in der Mitte liegt, sondern zwei), dann ist der Median der Durchschnitt der beiden mittleren Werte der Rangfolge.
Der Median ist stabil gegenüber Ausreißern (Extremwerten). Anwendbar ist er ab ordinalem Skalenniveau.

Das arithmetische Mittel ist der Durchschnitt

Das arithmetische Mittel (arithmetischer Mittelwert, durchschnittlicher Wert) ist der Durchschnittswert einer Verteilung. Es ist das wohl bekannteste Lagemaß. Es wird berechnet, indem alle Messungen addiert werden und anschließend die Summe durch die Anzahl der Messungen geteilt wird (◆ Kasten 4).
Das arithmetische Mittel wird u. a. – im Gegensatz zum Median – stark von Ausreißern beeinflusst. Es kann nur für Daten mit metrischem Skalenniveau berechnet werden.

Symmetrieeigenschaften einer Verteilung

Die drei Lagemaße geben auch Auskunft über die Symmetrieeigenschaften (Schiefe) einer Verteilung.

a) Arithmetisches Mittel ≈ Median ≈ Modus: Die Verteilung ist symmetrisch.

Bei der symmetrischen Häufigkeitsverteilung sind extrem hohe oder extrem niedrige Werte selten und mittlere Werte häufig. Besonders häufig Bezug genommen wird auf die Normalverteilung (Gauß-Verteilung; ◆ Abbildung 5) mit ihrer symmetrischen Glockenform, die durch eine mathematische Formel beschrieben werden kann. Eine Normalverteilung ist u. a. dadurch gekennzeichnet, dass arithmetisches Mittel, Median und Modus nahe beieinander liegen.
Egal, wie groß die Streuung der Daten ist (wie hoch und schmal oder wie flach und breit die Glocke ist) – bei einer Normalverteilung befinden sich im Bereich der einfachen Standardabweichung (s. u.) über- und unterhalb des arithmetischen Mittels immer 68 % aller Messwerte (◆ Abbildung 5). Innerhalb der zwei- oder dreifachen Standardabweichung beidseits des arithmetischen Mittels befinden sich 95 % bzw. 99 % aller Messwerte).

b) Modus < Median < arithmetisches Mittel: Die Verteilung ist rechtsschief (linkssteil, linksgipflig (◆ Abbildung 6).

c) Arithmetisches Mittel < Median < Modus: Die Verteilung ist linksschief (rechtssteil, rechtsgipflig (◆ Abbildung 7).

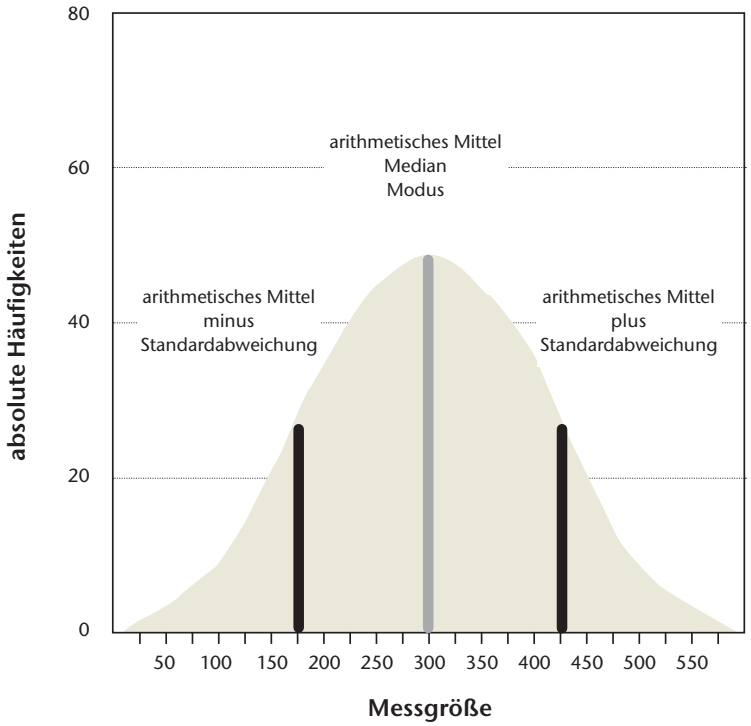


Abb. 5: **Normalverteilung** [mod. nach SCHNEIDER (1997)]

Ein Perzentil (auch Quantil genannt) bringt die relative Position eines Messwerts innerhalb der Stichprobe zum Ausdruck. Ein Perzentil bezieht sich immer auf einen vorgegebenen Prozentsatz. Bspw. bezeichnet das 20. Perzentil (P20) den Wert, unterhalb dem 20 % der Werte der Stichprobe liegen, und das 50. Perzentil (P50) bezeichnet den Wert, unterhalb dem 50 % der Stichprobenwerte liegen. Das heißt: **Das 50. Perzentil entspricht dem Median.**

Perzentile können verwendet werden, um Bereiche zu kennzeichnen, in denen ein bestimmter Prozentsatz der Stichprobenwerte liegt. Oftmals wird dabei die Stichprobe in gleich große Anteile zerlegt. Bspw. sind die 25 %-, 50 %- und 75 %-Perzentile diejenigen Werte, welche die Stichprobe in vier gleich große Anteile zerlegen. Man spricht deshalb auch von „Quartilen".

Kasten 3: **Perzentile (Quantile) einer Stichprobe**

Bei asymmetrischen Häufigkeitsverteilungen treten entweder kleinere (rechtsschief) oder größere Werte (linksschief) gehäuft auf (◆ Abbildungen 6 und 7). Es gibt auch U-förmige Häufigkeitsverteilungen, bei denen extrem hohe und extrem niedrige Werte häufig, mittlere Werte selten sind.
Generell gilt: Je näher arithmetisches Mittel, Median und Modus beieinander liegen, desto symmetrischer ist die Verteilung. Je stärker die drei Werte differieren, desto „schiefer" ist die Verteilung.

Arithmetisches Mittel oder Median?

Ob und wann das arithmetische Mittel oder der Median als statistische Kennzahl anzuwenden ist, hängt vom Skalenniveau der Messwerte, dem Häufigkeitsverteilungstyp und der Zahl der Untersuchungsobjekte ab.
Arithmetisches Mittel und entsprechende Streuungsmaße sind anwendbar für metrische Daten (ab Intervallskalenniveau). Sie fassen symmetrisch bzw. normalverteilte Daten gut zusammen; asymmetrische Verteilungen können sie aufgrund der Empfindlichkeit gegenüber Extremwerten nur unzureichend charakterisieren. Die Zahl der Messwerte sollte ausreichend groß sein.
Median und Perzentile sind anwendbar für Daten ab Ordinalskalenniveau. Der Median ist dem arithmetischen Mittel bei asymmetrischer Verteilung der Daten und bei sehr kleiner Zahl an Messwerten vorzuziehen.

Streuungsmaße

Die Lagemaße allein beschreiben eine Datenverteilung nicht ausreichend, da sie die Streuung der Daten nicht berücksichtigen. Bei der Beschreibung statistischer Ergebnisse müssen daher zusätzlich Streuungsmaße angegeben werden. Leider werden diese oft „vergessen".
Streuungsmaße geben an, inwieweit die Daten einer Verteilung vom Lagemaß abweichen. Sie zeigen also, wie gut oder wie schlecht die Lagemaße eine Verteilung repräsentieren.

Die Spannweite ist empfindlich gegenüber Ausreißern

Das einfachste Maß für die Streuung ist die Spannweite (Variationsbreite, Range) einer Variable – also die Differenz zwischen dem größten und dem kleinsten Wert. Sie wird angegeben, indem diese beiden Extremwerte notiert werden – am Beispiel in ◆ Kasten 5: 66 bis 145 kg. Die Nachteile der Spannweite sind, dass sie keine Aussage über die Streuung der übrigen Werte erlaubt und empfindlich gegenüber Ausreißern ist – schon ein einzelner extremer Wert kann die Spannweite stark vergrößern.

Der Interquartilsabstand deckt die mittleren 50 % ab

Letztgenannter Nachteil kann umgangen werden, indem statt der Spannweite der Interquartilsabstand verwendet wird. Dieser drückt die Länge des Bereichs aus, über den die mittleren 50 % einer Rohwerteverteilung streuen. Er wird ermittelt, indem alle Daten in aufsteigender Reihenfolge sortiert werden und dann das untere und obere Quartil bestimmt wird: der Wert bei einem Viertel und bei drei Vierteln der Liste. Der Interquartilsabstand ist die Differenz zwischen dem unteren und oberen Quartil (◆ Kasten 5).

Die Varianz ist die „durchschnittliche Streuung"

Die Varianz ist – wie das arithmetische Mittel – auch eine Art Durchschnittswert, nämlich die „durchschnittliche Streuung". Die Varianz ist definiert als die Summe der quadrierten Abweichungen der einzelnen Messwerte von ihrem arithmetischen Mittel, dividiert durch die Anzahl der Messwerte minus 1 (◆ Kasten 6). Je größer die Streuung, desto größer ist die Varianz. Die Interpretation der Varianz wird dadurch erschwert, dass sie durch das Quadrieren nicht mehr die Einheit der Messwerte besitzt. Ein besser verständliches Maß, welches sich direkt aus der Varianz ableitet, ist die Standardabweichung.

Die Standardabweichung ist ein Gütemaß für das arithmetische Mittel

Das am häufigsten benutzte Streuungsmaß ist die Standardabweichung (SD). Sie ist die Wurzel der Varianz (◆ Kasten 6). Durch Zie-

In ◆ Tabelle 2 wurden von 15 Klienten folgende Körpergewichte (kg) dokumentiert:
79 • 102 • 78 • 123 • 66 • 91 • 75 • 82 • 119 • 103 • 77 • 110 • 98 • 145 • 103

Der **Median** wird ermittelt, indem alle Messungen zuerst in aufsteigender Reigenfolge sortiert werden:
66 • 68 • 75 • 78 • 79 • 82 • 91 • **98** • 102 • 103 • 103 • 110 • 119 • 123 • 145

Der Median ist der **mittlere Wert** in der Reihenfolge, hier also: **98**

Wäre die Anzahl der Messungen geradzahlig (z. B. Körpergewicht von 16 Klienten), entspräche der Median dem Durchschnitt (arithmetischem Mittel) der beiden mittleren Werte:
66 • 68 • 75 • 78 • 79 • 82 • 91 • **98** • **102** • 103 • 103 • 110 • 119 • 123 • 145 • 147

Der Median wäre in diesem Fall (98 + 102)/2 = 100

Das **arithmetische Mittel** wird berechnet, indem alle Messungen addiert werden und dann durch die Anzahl der Messungen dividiert wird:
(79 + 102 + 78 + 123 + 66 + 91 + 75 + 82 + 119 + 103 + 68 + 110 + 98 + 145 + 103)/15 = 1457/15 = 96,1

Kasten 4: **Beispiel für die Berechnung von Median und arithmetischem Mittel**

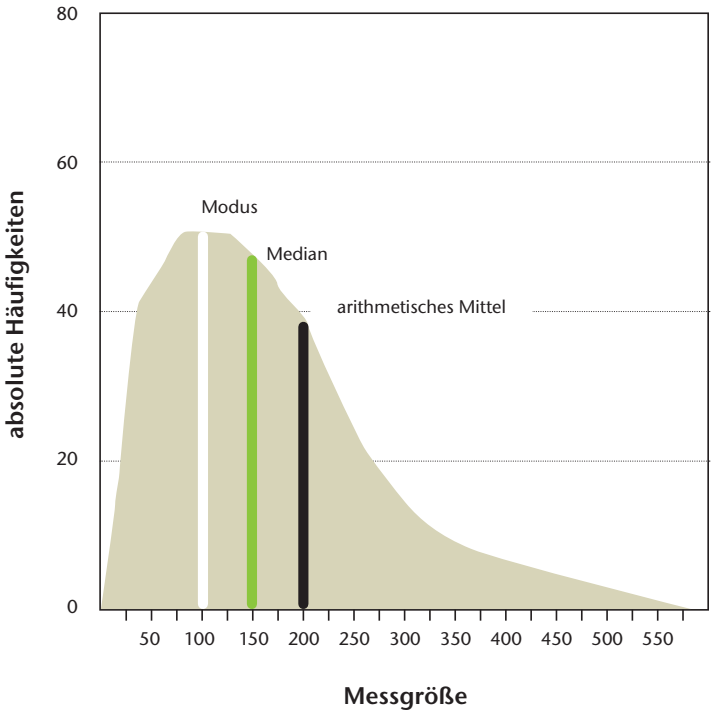


Abb. 6: **Asymmetrische Verteilung, rechtsschief** [mod. nach SCHNEIDER (1997)]

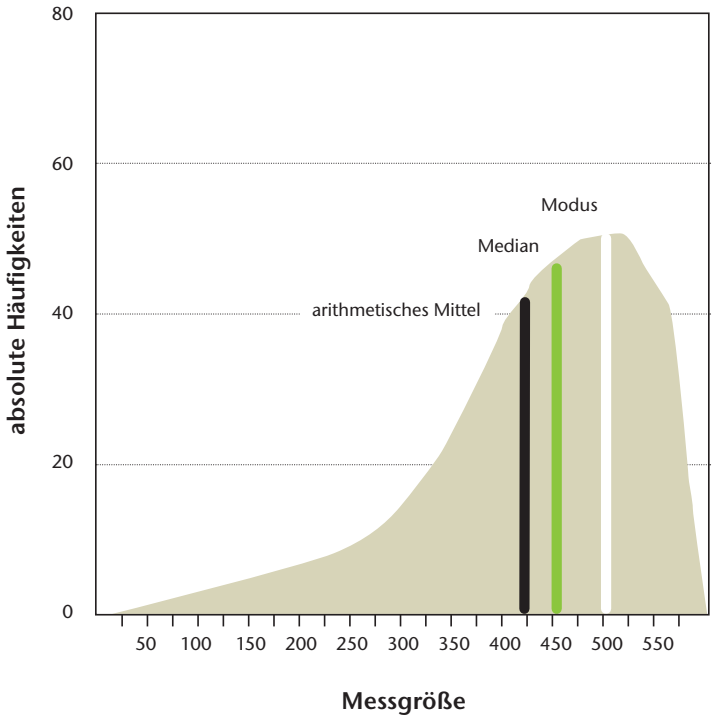


Abb. 7: **Asymmetrische Verteilung, linksschief** [mod. nach SCHNEIDER (1997)]

hung der Wurzel wird die Quadrierung wieder rückgängig gemacht. Dadurch ist die SD in der ursprünglichen Maßeinheit der Variablen zu lesen und besser zu interpretieren als die Varianz. Die SD ist ein Gütemaß für das arithmetische Mittel, denn: Je geringer die Streuung, desto weniger weichen die einzelnen Messwerte von ihrem arithmetischen Mittel ab. Das bedeutet: Je geringer die SD, desto besser repräsentiert das arithmetische Mittel die gesamte Verteilung.

Der Variationskoeffizient erlaubt Vergleiche zwischen Streuungen mehrerer Verteilungen

Der Variationskoeffizient ist ein Maß für die Kompaktheit von Daten. Er ist definiert als der Anteil der SD am arithmetischen Mittel und wird errechnet, indem die SD durch das arithmetische Mittel geteilt wird.
Im Gegensatz zu anderen Streuungsmaßen quantifiziert der Variationskoeffizient die Variabilität einer Verteilung maßstabsunabhängig. Er eignet sich damit sehr gut als Streuungs-Vergleichsmaß bei mehreren Verteilungen und kann als Prozentzahl gelesen werden: Im ◆ Kasten 6 ist das mittlere Körpergewicht der Stichprobe A 96,1 ± 22,3 kg. In einer anderen Stichprobe B sei es 87,2 ± 15,9 kg. Die Frage, in welcher Stichprobe die Streuung der Daten größer ist, beantwortet der Variationskoeffizient. Für Stichprobe A beträgt er 22,3 kg/96,1 kg = 0,23 und für Stichprobe B beträgt er 15,9 kg/87,2 kg = 0,18. Die Streuung ist in Stichprobe A mit 23 % somit größer als in Stichprobe B mit 18 %.

Die in ◆ Tabelle 2 dokumentierten Körpergewichte (kg) von 15 Klienten werden zuerst in aufsteigender Reigenfolge sortiert:

66 • 68 • 75 • **78** • 79 • 82 • 91 • 98 • 102 • 103 • 103 • **110** • 119 • 123 • 145

Das untere Quartil ist der Wert, der bei einem Viertel der Liste liegt, hier also der vierte: **78**. Das obere Quartil ist der Wert, der bei drei Vierteln der Liste liegt, hier also der zwölfte: **110**. Der Interquartilsabstand ist die Differenz zwischen dem oberen und dem unteren Quartil: 110 - 78 = **32**

Das heißt: Die mittleren 50 % der Körpergewichte streuen über einen Bereich von 32 kg.

In diesem Beispiel mit 15 Messungen ist die Bestimmung von Quartilen einfach. Bei anderen Anzahlen von Beobachtungen müssen ggf. gewichtete Mittel (ähnlich wie bei der Berechnung des Medians für eine gerade Anzahl an Messungen) berechnet werden. In der Praxis werden Quartile – sowie alle anderen Kennzahlen – normalerweise mit einem Statistikprogramm berechnet.

Kasten 5: **Beispiel für die Bestimmung des Interquartilsabstands**

Ein Box-Plot stellt Lage- und Streuungsmaße grafisch dar

Ein Box-Plot (◆ Abbildung 8) veranschaulicht grafisch gleichzeitig Lage- und Streuungsmaße. Er ermöglicht Aussagen über Symmetrie

sowie Zahl und Lage extremer Beobachtungen. Die „Box" umfasst dabei den Median als Zentrum der Daten sowie das untere und das obere Quartil (25 %- und 75 %-Perzentile). Damit werden durch die „Box" die mittleren 50 % der Daten repräsentiert. Die Werte an den Rändern der Verteilung werden jeweils vom Ende der Box ausgehend durch Striche – auch „Whisker" genannt – markiert. Der eine Whisker kennzeichnet die Verteilung der Messwerte, die kleiner als das erste Quartil sind, und der andere Whisker kennzeichnet die Verteilung der Messwerte, die größer als das dritte Quartil sind. Sofern keine Ausreißer vorliegen, zeigt das jeweilige Ende der Whisker den niedrigsten bzw. höchsten Messwert (Minimum bzw. Maximum) an.

Box-Plots sind gut geeignet, Ausreißer in den Daten zu identifizieren, die individuell eingezeichnet werden (meist mit einem * oder °). Die untere Grenze für Ausreißer (Länge des Whiskers) wird bestimmt, indem vom unteren Quartil das 1,5-Fache des Interquartilsabstands (◆ Kasten 5) subtrahiert wird. Für das in ◆ Abbildung 8 gezeigte Beispiel bedeutet dies: $78-1{,}5\times32 = 30$.

Messwerte, die diesen Grenzwert unterschreiten, werden als Ausreißer betrachtet. Wenn der kleinste Messwert diese Grenze nicht unterschreitet, endet der Whisker beim kleinsten Messwert (hier: 66). Die entsprechende obere Grenze ist die Summe aus oberem Quartil und dem 1,5-Fachen des Interquartilsabstands, hier: $110+1{,}5\times32 = 158$.

Messwerte, die diesen Grenzwert überschreiten, werden als Ausreißer betrachtet. Der größte Messwert im Beispiel ist 145; er überschreitet diese Grenze nicht, deshalb endet der obere Whisker bei diesem Maximum.

In ◆ Tabelle 2 wurden von 15 Klienten folgende Körpergewichte (kg) dokumentiert:

79 • 102 • 78 • 123 • 66 • 91 • 75 • 82 • 119 • 103 • 77 • 110 • 98 • 145 • 103

Der Mittelwert der Körpergewichte der 15 Klienten wurde in ◆ Kasten 4 berechnet mit 96,1 kg.

Die **Varianz** wird ermittelt, indem zuerst für jeden einzelnen Messwert dessen Differenz zum Mittelwert berechnet wird (seine Abweichung vom Mittelwert). Dann wird jede Abweichung quadriert (sodass lauter positive Werte entstehen). Danach werden alle quadrierten Abweichungen des Datensatzes addiert und anschließend durch die Anzahl der Messungen minus 1 geteilt:

$(79\text{-}96{,}1)^2 + (102\text{-}96{,}1)^2 + \ldots\ldots + (145\text{-}96{,}1)^2 + (103\text{-}96{,}1)^2 \,/\, (15\text{-}1) = 496{,}6$

Die **Standardabweichung** ist die Wurzel der Varianz: $\sqrt{496{,}6} = 22{,}3$ kg

Das heißt: Das mittlere Körpergewicht der 15 Klienten beträgt 96,1 ± 22,3 kg.

Kasten 6: **Beispiel für die Bestimmung von Varianz und Standardabweichung**

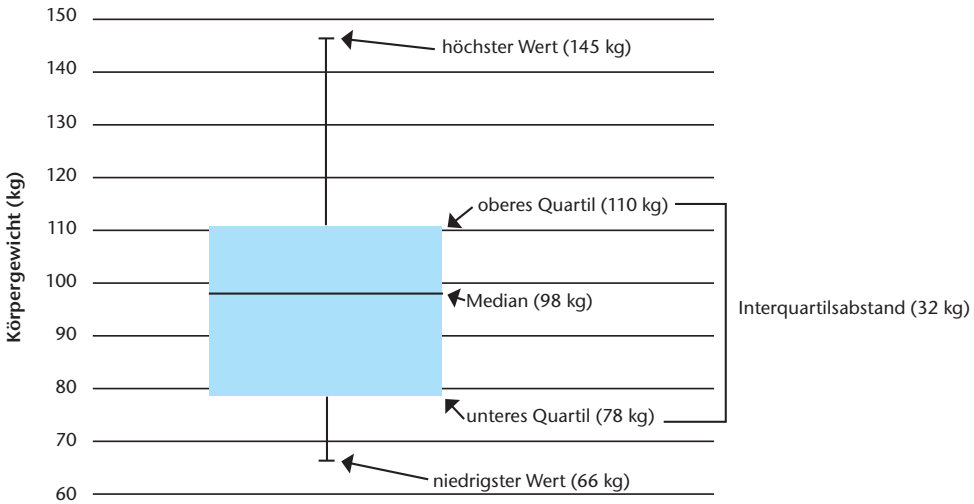


Abb. 8: **Box-Plot anhand der Körpergewichte aus ◆ Tabelle 2**
Wie erkennbar ist, liegen keine Ausreißer vor, sodass die Whisker bis zum niedrigsten und höchsten Wert reichen. Der Median befindet sich nicht mittig, sondern eher im oberen Bereich der Box und die Whisker haben unterschiedliche Längen. An diesen Aspekten ist erkennbar, dass die Verteilung nicht ganz symmetrisch ist.

Korrelations- und Assoziationsmaße

Bisher wurde nur die Datenreihe einer Variable betrachtet (univariate Statistik). Wenn eine an den gleichen Untersuchungsobjekten erhobene zweite Datenreihe einer anderen Variable einbezogen wird (bivariate Statistik), so kann die Frage beantwortet werden, ob ein Zusammenhang zwischen zwei Variablen vorliegt und wie stark dieser ist. Kennzahlen, die diese Frage beantworten können, werden als Korrelationsmaße oder Assoziationsmaße bezeichnet. Für nominale, ordinale und metrische Variablen gibt es unterschiedliche Korrelationsmaße. Solange es um metrische Variablen (hohes Skalenniveau) geht, können Zusammenhänge mit relativ einfachen statistischen Mitteln beschrieben werden – bei niedrigerem Niveau der Daten wird es schwieriger. Darauf kann hier nicht im Einzelnen eingegangen werden, es wird auf die Literatur verwiesen (z. B. Vierfeldertafel, Kontingenztabelle, 3-D-Balkendiagramme für nominale und/oder ordinale Variablen).

Streudiagramm und Korrelation

Als anschauliche grafische Darstellungsweise für die Betrachtung von zwei metrischen Merkmalen bietet sich das Streudiagramm (Punktwolke) an. Die ein Objekt betreffenden Wertepaare werden durch einen Punkt in einem Koordinatensystem

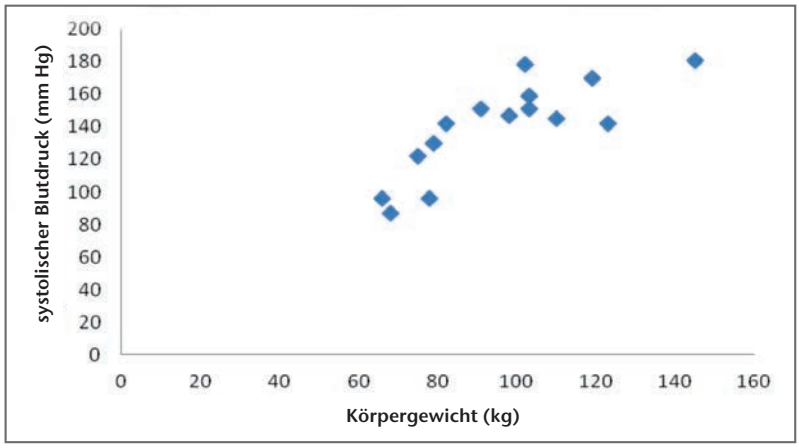


Abb. 9: **Streudiagramm (Punktwolke) für den Zusammenhang zwischen Körpergewicht und Blutdruck der in ◆ Tabelle 2 beschriebenen 15 Klienten**
Der Korrelationskoeffizient beträgt 0,8, d. h., es besteht ein positiver Zusammenhang zwischen Körpergewicht und systolischem Blutdruck.

abgebildet (◆ Abbildung 9). So ist zu sehen, ob ein Zusammenhang – eine Korrelation – zwischen beiden Merkmalen besteht und wenn ja, wie stark er ist. In dem in ◆ Abbildung 9 dargestellten Beispiel haben verschiedene Objekte gleichzeitig höhere Werte der einen Variable (Körpergewicht) als auch der anderen Variable (systolischer Blutdruck). Niedrigere Werte der ersten Variable sind mit niedrigeren Werten der zweiten Variable verbunden. Ein solcher Zusammenhang wird als positive Korrelation bezeichnet (s. u.). Wenn zu hohen Werten der einen Variable niedrige Werte der anderen Variable gehören, liegt eine negative Korrelation vor. Die Stärke des Zusammenhangs kann grafisch mittels einer Ellipse veranschaulicht werden, die die Messwerte umhüllt. Je schmaler die Ellipse, desto stärker ist der Zusammenhang.

Der Korrelationskoeffizient zeigt den linearen Zusammenhang

Eine Angabe über die Stärke eines Zusammenhangs zwischen zwei Variablen ist mit dem Zusammenhangsmaß des Korrelationskoeffizienten möglich. Er ist ein Maß dafür, inwiefern die Punkte des Streudiagramms auf einer Geraden liegen. In die Berechnung des Korrelationskoeffizienten geht die Differenz jedes einzelnen Messwerts zum Mittelwert ein (das wird normalerweise mit einem Statistikprogramm berechnet). Der Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen:

- Korrelationskoeffizient -1: perfekt/stark negativer Zusammenhang
- Korrelationskoeffizient < 0: beide Variablen sind negativ linear abhängig
- Korrelationskoeffizient = 0: es besteht kein linearer Zusammenhang (lineare Unabhängigkeit)
- Korrelationskoeffizient > 0: beide Variablen sind positiv linear abhängig
- Korrelationskoeffizient +1: perfekt/stark positiver Zusammenhang

Je näher der Koeffizient bei 0 liegt, desto schlechter ist die Anpassung der Geraden an die Punkte. Es ist zu beachten, dass ein Korrelationskoeffizient von 0 nicht bedeutet, dass die zwei Variablen nicht zusammenhängen – sondern nur, dass der Zusammenhang nicht einer Gerade entspricht (er kann z. B. nichtlinear U-förmig sein).
Wichtig für die Interpretation von Korrelationen sind folgende Aspekte:

- **Korrelation ist nicht Kausalität!** Eine Korrelation sagt nicht zwangsläufig etwas über die Verursachung aus. Eine vorhandene Korrelation zeigt zunächst nur einen Zusammenhang, der viele Ursachen haben kann (bekanntermaßen war die Abnahme der Storchenbestände trotz vorhandener Korrelation kein ursächlicher Faktor für den Geburtenrückgang im letzten Jahrhundert – die zunehmende Industrialisierung hatte Einfluss auf beide Faktoren).
- Der Korrelationskoeffizient sagt bei nichtlinearen Zusammenhängen möglicherweise nichts über die Stärke des Zusammenhangs aus.
- Der Korrelationskoeffizient kann stark durch Ausreißerwerte beeinflusst werden.
- Abhängig vom Skalenniveau sollten unterschiedliche Verfahren zur Berechnung der Zusammenhangsmaße verwendet werden (Kontingenz-, Rangkorrelations-, Maßkorrelationskoeffizient, ◆ Tabelle 1 und weiterführende Literatur).
- Eine Korrelation erlaubt keine Vorhersage! Der Korrelationskoeffizient sagt nichts über die Art des Zusammenhangs aus. Und er lässt keine Aussage darüber zu, welcher Wert in der einen Messgröße zu erwarten ist, wenn die andere Messgröße um einen bestimmten Betrag verändert wird. Für eine solche Vorhersage von Werten ist die Berechnung einer Regressionsgeraden erforderlich.

Die lineare Regression schätzt Werte

Die Vorhersage von (nicht gemessenen) Werten einer Variablen aus den Werten einer zweiten Variable ist mit der **Regressionsgeraden** möglich, die die Art des Zusammenhangs von zwei metrischen Variablen beschreibt. Dies ist allerdings nur möglich, wenn eine Variable eindeutig von der anderen abhängt und der umgekehrte Fall nicht denkbar ist. So kann die Höhe des Blutdrucks vom Gewicht abhängen, das Gewicht jedoch nicht vom Blutdruck. Der Blutdruck ist deshalb die abhängige Variable (Outcome; im Streudiagramm i. d. R. auf der y-Achse),

das Gewicht die unabhängige Variable (Prädiktor; im Streudiagramm i. d. R. auf der x-Achse).

◆ Abbildung 10 stellt eine Regressionsgerade dar. Das ist die Gerade mit der besten Anpassung an die Punkte im Streudiagramm. Sie wird üblicherweise per Regressionsanalyse mit einem Statistikprogramm berechnet und – nach definierten mathematischen Regeln (Methode der kleinsten Quadrate) – durch die Punkte hindurch gezeichnet sowie auch als Gleichung angegeben. Die Steigung der Geraden wird als Regressionskoeffizient bezeichnet.

Die Regressionsgleichung liefert den Wert der abhängigen Variable, wenn die unabhängige Variable bekannt ist. Im hier gewählten Beispiel (◆ Abbildung 10) zeigt sie an, dass der Wert des systolischen Blutdrucks im Mittel um ca. 1,1 mm Hg ansteigt, wenn das Körpergewicht um 1 kg zunimmt. Bei einer 90 kg schweren Person ist mit einem Blutdruck von $1{,}1 \times 90 + 38{,}4 = 137$ mm Hg zu rechnen.

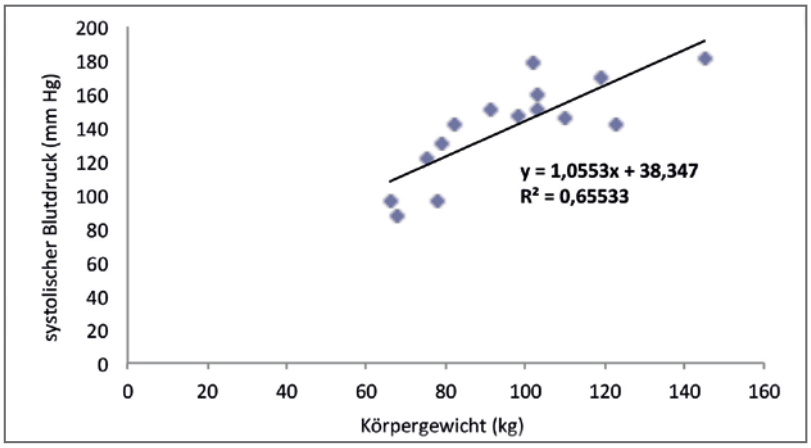


Abb. 10: Streudiagramm (Punktwolke) mit Regressionsgerade und Regressionsgleichung für den Zusammenhang zwischen Körpergewicht (unabhängige Variable) und Blutdruck (abhängige Variable) der in ◆ Tabelle 2 beschriebenen 15 Klienten

Ein Maß für die Güte einer Regressionsanalyse ist das **Bestimmtheitsmaß B.** Diese statistische Kennzahl berechnet sich durch Quadrieren des Korrelationskoeffizienten. Dadurch kann sie nur positive Werte annehmen. Ein Korrelationskoeffizient von 0,8 zwischen den Körpergewichten und Blutdruckwerten entspricht einem Bestimmtheitsmaß von $0{,}8 \times 0{,}8 \approx 0{,}6$. Das heißt, 60 % der Variation in den Blutdruckwerten können durch das Körpergewicht „erklärt" werden (im linearen Modell). Oder: 40 % der Variation in der abhängigen Variable (Blutdruck) können nicht durch die berücksichtigte unabhängige Variable (Körpergewicht) erklärt werden. Diese nicht erklärte Variation wird durch nicht berücksichtigte bzw. unbekannte Variablen verursacht. Ein hohes Bestimmtheitsmaß belegt also einen starken Zusammenhang zwischen den Variablen und ist ein Gütekriterium für eine Regressionsanalyse.

Die Berechnung von Regressionsgeraden (Regressionsanalyse) sollte nur durchgeführt werden, wenn

- zwischen den beiden Variablen ein linearer Zusammenhang besteht, der nicht zu schwach ist (Korrelationskoeffizient) – sonst hat die Regressionsgerade wenig Aussagekraft;
- zwischen den beiden Variablen eine gerichtete Abhängigkeit besteht (eine Variable hängt von der anderen ab, die Abhängigkeit ist nicht umkehrbar);
- die Daten auf mindestens Intervallskalenniveau vorliegen.

Vorhersagen mit Regressionsanalysen sind Schätzungen! Geschätzt werden sollte nur für den Bereich, für den auch Werte gemessen wurden (im Beispiel Körpergewichte zwischen ca. 60 und 145 kg). Nur für diesen Bereich gibt es Anhaltspunkte für den Verlauf der Geraden, in anderen Bereichen könnte er anders sein.

Resümee

Ist ein arithmetisches Mittel ohne Standardabweichung angegeben? Eine prozentuale Häufigkeit ohne Gesamtfallzahl? Eine Grafik ohne eindeutige Beschriftung? Zeigt eine Studie, dass die Körpergröße Einfluss auf das Einkommen hat, weil größere Menschen mehr verdienen als kleinere? Als Leser dieses Beitrags werden Sie in Zukunft solche Fehlinformationen durch beschreibende Statistik erkennen und bei der Erstellung von eigenen Berechnungen vermeiden können.

Folgende Geschichte von Dubben und Beck-Bornholdt (2006) verdeutlicht, wie Fehlschlüsse allein durch die Verwechslung von Anzahl und Anteil entstehen können: „Auf dem Küchentisch liegen sieben Würstchen und drei Eier, also zusammen zehn Dinge. Die Eier machen somit 30 % (3/10 = 0,3 oder 30 %) der Gegenstände auf dem Tisch aus. Dann betritt ein Hund unbeobachtet den Raum. Er frisst fünf Würstchen auf ... Jetzt sind noch zwei Würstchen und drei Eier übrig, also insgesamt fünf Dinge. Der Eieranteil beträgt jetzt 60 % (3/5 = 0,6 oder 60 %). Durch die wunderbare Tat des Hundes hat er sich verdoppelt! So weit ist noch alles richtig, aber zu folgern, dass sich die Anzahl der Eier verdoppelt hat und der Hund somit Eier legen kann, ist schlicht falsch. Diese Geschichte … hat zahlreiche Parallelen in der wissenschaftlichen Literatur."

Angela Bechthold, Diplom-Oecotrophologin, Fachjournalistin (FJS)
Dr. Ute Brehme
Deutsche Gesellschaft für Ernährung e. V.
Godesberger Allee 18
53175 Bonn

Quellen für diesen Beitrag und weiterführende Literatur sowie Hinweise auf Info- und Lernmaterialien zur Statistik finden Sie unter www.ernaehrungs-umschau.de/service/literaturverzeichnisse

Anerkannte Fortbildung für die Zertifikate der DGE, des VDD, des VDOE und der Landesärztekammer (Hessen)

Online-Fortbildung

10 Fragen

Weitere Informationen zur Online-Fortbildung finden Sie unter **www.ernaehrungs-umschau.de**

Bei allen Fragen ist jeweils nur eine Antwort richtig:

1. Die deskriptive Statistik erlaubt Aussagen über Eigenschaften von Objekten, ...

- ○ A deren Merkmale ausschließlich metrisch skalierte Messwerte haben.
- ○ B die immer repräsentativ für die Grundgesamtheit sind.
- ○ C die in einer Stichprobe tatsächlich untersucht wurden.
- ○ D deren Ausprägungen durch schließende Statistik nicht vorhersagbar sind.

2. Die Häufigkeiten von Blutdruckwerten von 278 Patienten sollen grafisch dargestellt werden. Welche der folgenden Darstellungsformen ist geeignet und warum?

- ○ A Ein Histogramm, da der Blutdruck eine metrische Variable ist.
- ○ B Ein Kreisdiagramm, da der Blutdruck ein nominales Datenniveau hat.
- ○ C Ein Streudiagramm, da es die Verteilung der Werte einer Variablen am besten abbildet.
- ○ D Ein Balkendiagramm, da es jeden einzelnen Blutdruckwert zeigen kann.

3. 16 Teilnehmer eines Kurses machten folgende Angaben zur Anzahl der von ihnen täglich eingenommenen Medikamente:

0 • 1 • 2 • 5 • 4 • 3 • 1 • 2 • 2 • 0 • 0 • 1 • 3 • 3 • 3 • 2

Die relative Häufigkeit der Angabe „2 Medikamente täglich“ beträgt ...

- ○ A 0,17
- ○ B 0,20
- ○ C 0,33
- ○ D 0,25

4. Welche der folgenden Aussagen zum Median trifft NICHT zu? Der Median ist ...

- ○ A wie das arithmetische Mittel ein Lagemaß.
- ○ B anfällig für Ausreißerwerte.
- ○ C der Wert, der eine der Größe nach geordnete Messwertreihe halbiert.
- ○ D das 50. Perzentil.

5. In einer Messwertreihe der Altersangaben 10, 10, 20, 25 und 35 Jahre betragen sowohl das arithmetische Mittel als auch der Median 20 Jahre. Statt der 35-jährigen Person wird eine 70-jährige Person in die Messwertreihe aufgenommen. Hierdurch ...

- ○ A erhöhen sich beide Lagemaße auf 27 Jahre.
- ○ B bleibt das arithmetische Mittel unverändert, der Median erhöht sich auf 27 Jahre.
- ○ C erhöht sich das arithmetische Mittel auf 25 Jahre, der Median bleibt unverändert.
- ○ D erhöht sich das arithmetische Mittel auf 27 Jahre, der Median bleibt unverändert.

6. Welche Aussage zur Symmetrie einer Verteilung trifft zu?

- ○ A Als Normalverteilung wird eine Verteilung bezeichnet, in der es keine Ausreißerwerte gibt.
- ○ B Je näher Modus, Median und arithmetisches Mittel beieinander liegen, umso symmetrischer ist eine Verteilung.
- ○ C Stimmen Median und arithmetisches Mittel nicht überein, liegt eine Gauß-Verteilung vor.
- ○ D Bei einer „linksschiefen“ Verteilung treten kleinere Werte gehäuft auf.

7. Für welche der folgenden Variablen ist es möglich, arithmetisches Mittel und Standardabweichung zu berechnen?

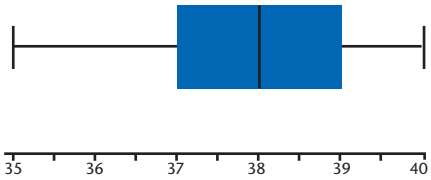
1. Adipositas in Klassen mit dem Schweregrad 1, 2 und 3
2. Geschlecht der Patienten einer Stichprobe
3. Serumcholesterinkonzentration (in mmol/L) bei Teilnehmern einer Verzehrstudie
4. Körpergröße (in m) einer Stichprobe von Patienten

- ○ A Nur Antwort 3 ist richtig.
- ○ B Nur die Antworten 1 und 4 sind richtig.
- ○ C Nur die Antworten 3 und 4 sind richtig.
- ○ D Alle Antworten sind richtig.

8. Bei welcher der folgenden Angaben zum arithmetischen Mittel und seiner Standardabweichung repräsentiert der Mittelwert die Verteilung der Messwerte am besten?

- ○ A 120 ± 10
- ○ B 120 ± 120
- ○ C 120 ± 40
- ○ D 120 ± 80

9. Welche Informationen sind anhand des folgenden Box-Plots abzulesen?



- ○ A Der Median der Verteilung beträgt 37.
- ○ B Das arithmetische Mittel beträgt 38.
- ○ C Der seltenste Wert liegt bei 35, der häufigste Wert bei 40.
- ○ D Die mittleren 50 % der Messwerte liegen zwischen 37 und 39.

10. Es soll beschrieben werden, ob ein linearer Zusammenhang zwischen dem Ballaststoffgehalt und der Energiedichte von ausgewählten Lebensmitteln besteht. Welche statistische Kennzahl ist geeignet?

- ○ A Korrelationskoeffizient
- ○ B Variationskoeffizient
- ○ C arithmetisches Mittel
- ○ D Interquartilsabstand