Springer-Lehrbuch

Statistik für Human- und Sozialwissenschaftler

Limitierte Sonderausgabe

Bearbeitet von
Jürgen Bortz, Christof Schuster

7., vollst. überarb. u. aktualisierte Aufl. 2010. Buch. xix, 655 S.
ISBN 978 3 642 12769 4
Format (B x L): 19,3 x 26 cm
Gewicht: 1960 g

Kapitel 2 Statistische Kennwerte

ÜBERSICHT

Arithmetisches Mittel – Modalwert – Medianwert – Varianz – Standardabweichung – Interquartilbereich – Perzentil – *z*-Wert

Die Anwendung statistischer Verfahren setzt voraus, dass quantitative Informationen über den jeweiligen Untersuchungsgegenstand bekannt sind. Die Aussage: „Herr *X* ist neurotisch" mag zwar als qualitative Beschreibung der genannten Person informativ sein; präziser wäre diese Information jedoch, wenn sich die Ausprägung des Neurotizismus durch eine bestimmte Zahl kennzeichnen ließe, die beispielsweise Vergleiche hinsichtlich der Ausprägungsgrade des Neurotizismus bei verschiedenen Personen ermöglicht.

Liegen quantitative Informationen über mehrere Personen bzw. eine Stichprobe vor, erleichtern summarische Darstellungen der Daten die Interpretation der in der Stichprobe angetroffenen Merkmalsverteilung. Die Altersangaben der Klienten einer therapeutischen Ambulanz beispielsweise könnten folgendermaßen statistisch „verdichtet" werden:

- *Maße der zentralen Tendenz* geben an, welches Alter alle Klienten am besten charakterisiert.
- *Maße der Variabilität* kennzeichnen die Unterschiedlichkeit der behandelten Klienten in Bezug auf das Alter.

Kennwerte, die entweder die zentrale Tendenz oder die Variabilität eines Merkmals charakterisieren, sollen nun dargestellt werden.

2.1 Maße der zentralen Tendenz

Eine Stichprobe von n Untersuchungseinheiten soll hinsichtlich eines Merkmals beschrieben werden. Beispielsweise soll die Fähigkeit, aus einzelnen Teilstücken eine vorgegebene Figur zusammenzusetzen (Puzzle), untersucht werden. An der Untersuchung nehmen 90 Patienten mit hirnorganischen Schäden teil. Das uns interessierende Merkmal ist die Bearbeitungszeit, die die Versuchspersonen zum Zusammenlegen der Figur benötigen. Tabelle 2.1 enthält die Bearbeitungszeiten der Patienten.

Tabelle 2.1. Bearbeitungszeiten eines Puzzles in Sekunden

131,8	106,7	116,4	84,3	118,5	93,4	65,3	113,8	140,3
119,2	129,9	75,7	105,4	123,4	64,9	80,7	124,2	110,9
86,7	112,7	96,7	110,2	135,2	134,7	146,5	144,8	113,4
128,6	142,0	106,0	98,0	148,2	106,2	122,7	70,0	73,9
78,8	103,4	112,9	126,6	119,9	62,6	116,6	84,6	101,0
68,1	95,9	119,7	122,0	127,3	109,3	95,1	103,1	92,4
103,0	90,2	136,1	109,6	99,2	76,1	93,9	81,5	100,4
114,3	125,5	121,0	137,0	107,7	69,0	79,0	111,7	98,8
124,3	84,9	108,1	128,5	87,9	102,4	103,7	131,7	139,4
108,0	109,4	97,8	112,2	75,6	143,1	72,4	120,6	95,2

Nun überlegen wir, durch welchen Wert alle Bearbeitungszeiten am besten beschrieben werden können. In der Tat gibt es zu diesem Zweck mehrere *Kennwerte*. Die gebräuchlichsten Maße sind der Mittelwert, der Median und der Modalwert (s. Exkurs 2.2 für weitere Maße).

2.1.1 Mittelwert

Der Mittelwert ist das gebräuchlichste Maß zur Kennzeichnung der zentralen Tendenz der Verteilung eines metrischen Merkmals. Er wird berechnet, indem die Summe aller Werte durch die Anzahl der Werte dividiert wird. Die Formel für den Mittelwert lautet

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (2.1)$$

Der Mittelwert wird häufig auch „arithmetisches Mittel" genannt. (Exkurs 2.1 enthält Hinweise für das Rechnen mit dem Summenzeichen Σ.)

Zunächst illustrieren wir die Berechnung des Mittelwerts mit Hilfe eines Beispiels.

BEISPIEL 2.1

Fünf Schüler schreiben ein Diktat im Fach Englisch. Die Anzahl der Orthographiefehler lauten: 3, 5, 6, 8, und 14. Berechnen wir nun den Mittelwert der Fehler, so ergibt sich

$$\bar{x} = (3+5+6+8+14)/5 = 36/5 = 7{,}2.$$

Im Durchschnitt machte ein Schüler etwa sieben Fehler.

Das arithmetische Mittel ist hinsichtlich jedes einzelnen Wertes, welcher in seine Berechnung eingeht, sensitiv. Wird ein einzelner Wert der vorliegenden Stichprobe verändert, so ändert sich ebenfalls der Mittelwert. Diese Eigenschaft ist der Grund, wieso der Mittelwert ein sehr guter Schätzer des *Zentrums* einer Verteilung ist.

Auf der anderen Seite birgt diese Sensitivität des Mittels hinsichtlich jeder einzelnen Beobachtung auch die Gefahr, dass das Mittel durch ungewöhnliche Beobachtungen stark beeinflusst wird. Solche ungewöhnlichen Beobachtungen, die oft als *Ausreißer* oder *Extremwerte* bezeichnet werden, können aus vielen verschiedenen Gründen in den Daten enthalten sein.

Eine mögliche Ursache sind Fehler, die sich in die Berechnung einschleichen. So könnte es im Beispiel 2.1 bei der Berechnung der durchschnittlichen Fehleranzahl mit Hilfe eines Taschenrechners bei der Eingabe zu einem Tippfehler kommen. Wir nehmen folgendes Szenario an: Anstatt der Zahl 14 wird versehentlich die Zahl 24 eingegeben. Aufgrund dieses Fehlers lautet die durchschnittliche Fehlerzahl nun 9,2. Dieser Wert ist kein guter Repräsentant der beobachteten Fehler in Beispiel 2.1 mehr.

Eine weitere Eigenschaft des Mittels ist, dass die Summe der Abweichungen vom arithmetischen Mittel immer *null* ergeben muss, d. h. es gilt für beliebige Werte $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Um diese Aussage herzuleiten, wird die Summe der Abweichungen vom Mittel folgendermaßen umgeformt:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - n \cdot \bar{x} \\ &= \sum_{i=1}^{n} x_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.\end{aligned}$$

Um diese Eigenschaft zu illustrieren, berechnen wir die Summe der Abweichungen für die Werte des Beispiels 2.1. Man erhält

$$(3-7{,}2) + (5-7{,}2) + \cdots + (14-7{,}2) = 0.$$

Eine weitere interessante Eigenschaft des Mittels ist, dass die Summe der quadrierten Abweichungen aller Werte vom Mittel ein *Minimum* ergibt. Mit anderen Worten, sucht man einen Wert, den wir mit $\tilde{x}$ bezeichnen wollen, für den die Summe

$$\sum_{i=1}^{n}(x_i - \tilde{x})^2$$

so klein wie möglich wird, so ist nicht offensichtlich, wie $\tilde{x}$ gewählt werden muss, um diese Summe zu minimieren. Es lässt sich zeigen, dass der Mittelwert diese Bedingung erfüllt. Zum Beweis siehe den Abschnitt zur *Methode der kleinsten Quadrate* auf S. 90.

BEISPIEL 2.2

Auch diese „Minimums"-Eigenschaft des Mittels soll mit den Werten des Beispiels 2.1 illustriert werden. Wir berechnen dazu die Quadratsumme $\sum(x_i - \bar{x})^2$. Man erhält

$$(3-7{,}2)^2 + (5-7{,}2)^2 + \cdots + (14-7{,}2)^2 = 70{,}8.$$

Ersetzt man 7,2 durch einen beliebigen anderen Wert, so wird die resultierende Quadratsumme auf keinen Fall kleiner als 70,8 (s. Aufgabe 2.9).

BEISPIEL 2.3

Als zweites numerisches Beispiel für die Berechnung des Mittelwertes bestimmen wir das Mittel für die Bearbeitungszeiten der Tab. 2.1. Die Berechnung von Hand ist zwar nicht weiter schwierig, aber aufgrund der großen Anzahl von Beobachtungen aufwändig. Man erhält

$$\bar{x} = (131{,}8 + \cdots + 95{,}2)/90 = 9619{,}9/90 = 106{,}89.$$

Damit beträgt die durchschnittliche Bearbeitungszeit des Puzzles etwa 107 Sekunden.

2.1.2 Median

Der Median einer Stichprobe von Werten ist definiert als der Wert, der größer als 50% der Werte der Stichprobe ist. Der Median kennzeichnet auf einfache Weise die *Mitte* der Stichprobenwerte, da die Hälfte der Werte kleiner und die andere Hälfte der Werte größer ist als der Median.

EXKURS 2.1 Das Rechnen mit dem Summenzeichen

Ein in der Statistik häufig benötigtes Operationszeichen ist das Summenzeichen, das durch $\sum$ gekennzeichnet wird. Unter Verwendung des Summenzeichens schreiben wir z. B.:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5.$$

Dabei bedeutet die linke Seite der Gleichung: „Summe aller x_i-Werte für $i = 1$ bis 5“. Der Laufindex i kann durch beliebige andere Buchstaben ersetzt werden. Unterhalb des Summenzeichens wird die untere Grenze des Laufindex angegeben, oberhalb des Summenzeichens steht die obere Grenze. Die folgenden Beispiele verdeutlichen einige Operationen mit dem Summenzeichen:

$$\sum_{i=3}^{6} x_i = x_3 + x_4 + x_5 + x_6,$$

$$\sum_{i=2}^{4} x_i \cdot y_i = x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4,$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2,$$

$$\sum_{i=1}^{n} (x_i + a) = (x_1 + a) + \cdots + (x_n + a) = \sum_{i=1}^{n} x_i + n \cdot a,$$

$$\sum_{i=1}^{n} c \cdot x_i = c \cdot x_1 + \cdots + c \cdot x_n = c \sum_{i=1}^{n} x_i,$$

Wenn aus dem Kontext die Grenzen der zu summierenden Werte klar hervorgehen, kann die ausführliche Schreibweise für eine Summation durch folgende einfachere Schreibweise ersetzt werden:

$$\sum_{i=1}^{n} x_i = \sum_{i} x_i.$$

Häufig sind Daten nicht nur nach einem, sondern nach mehreren Kriterien gruppiert, sodass eine eindeutige Kennzeichnung nur über mehrere Indizes möglich ist. Wenn beispielsweise p Variablen bei n Personen gemessen werden, kennzeichnen wir die 3. Messung der 2. Personen durch x_{23} oder allgemein die i-te Messung der m-ten Person durch x_{mi}. Will man die Summe aller Messwerte der 2. Person bestimmen, verwenden wir folgende Rechenvorschrift:

$$\sum_{i=1}^{p} x_{2i} = x_{21} + x_{22} + x_{23} + \cdots + x_{2p}.$$

Die Summe aller Messwerte für die Variable 5 hingegen lautet:

$$\sum_{m=1}^{n} x_{m5} = x_{15} + x_{25} + x_{35} + \cdots + x_{n5}.$$

Die Summe der Werte der m-ten Person ermitteln wir nach der Beziehung:

$$\sum_{i=1}^{p} x_{mi} = x_{m1} + x_{m2} + \cdots + x_{mp}$$

bzw. die Summe aller Werte auf der i-ten Variablen:

$$\sum_{m=1}^{n} x_{mi} = x_{1i} + x_{2i} + \cdots + x_{ni}.$$

Sollen die Messwerte über alle Personen und alle Variablen summiert werden, kennzeichnen wir dies durch ein doppeltes Summenzeichen:

$$\sum_{i=1}^{p} \sum_{m=1}^{n} x_{mi}.$$

Entsprechendes gilt für Messwerte, die mehr als zweifach indiziert sind.

Die Berechnung des Medians wird folgendermaßen bewerkstelligt: Zunächst werden die Rohwerte nach ihrer Größe sortiert. Die sortierten Rohwerte bezeichnet man auch als „Ordnungsstatistik“ und schreibt sie als

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

Mit $x_{(1)}$ ist also der kleinste Wert einer Stichprobe gemeint. Ganz analog bezeichnet $x_{(n)}$ den größten beobachteten Wert. Dagegen bezeichnet x_1 bzw. x_n die erste und die letzte Beobachtung, die in der Stichprobe gemacht wurde.

Mit Hilfe der Schreibweise für die sortierten Rohdaten kann man den Median, den wir mit Md abkürzen, folgendermaßen ausdrücken; dabei muss zwischen geradem und ungeradem Stichprobenumfang unterschieden werden. Es gilt:

$$\mathrm{Md} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{falls } n \text{ ungerade,} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)/2 & \text{falls } n \text{ gerade.} \end{cases}$$

BEISPIEL 2.4

Haben fünf Versuchspersonen die Messwerte 5, 2, 3, 7, und 8 erhalten, so müssen die Daten zunächst sortiert werden. Man erhält somit

$$2, 3, 5, 7, 8.$$

Man erkennt, dass die dritte Beobachtung die Rohwerte in zwei gleich große Hälften teilt. Damit lautet der Median $\mathrm{Md} = 5$.

Natürlich kann auch die obige Formel angewendet werden, um den Median zu bestimmen. Da die Anzahl der Beobachtungen n ungerade ist, bestimmt man zunächst $(n + 1)/2 = 3$.

Somit ergibt sich $\text{Md} = x_{(3)}$. Der Median ist also die drittgrößte Beobachtung, die den Wert 5 besitzt.

Ist der Stichprobenumfang geradzahlig, so liegen zwei Werte in der Mitte der sortierten Messwerte. Wir gehen diesmal von den sechs Messwerten 2, 8, 6, 4, 12 und 10 aus. Wiederum werden diese Werte zunächst nach ihrer Größe sortiert. Man erhält

$$2, 4, 6, 8, 10, 12.$$

Die beiden mittleren Beobachtungen sind 6 und 8. Der Median wird nun als arithmetisches Mittel dieser beiden Beobachtungen bestimmt. Also $\text{Md} = (6 + 8)/2 = 7$.

Wiederum überzeugen wir uns, dass die Verwendung der Fomel für den Median zum gleichen Ergebnis führt. Da $n = 6$, erhält man

$$\text{Md} = (x_{(6/2)} + x_{(6/2+1)})/2 = (x_{(3)} + x_{(4)})/2 = 7.$$

■

Zunächst zwei Bemerkungen zur Definition des Medians für eine Stichprobe. Erstens macht das Beispiel deutlich, dass die Definition des Medians als der Wert, unter dem die Hälfte der Beobachtungen liegt, nur für gerades n sinnvoll ist. Trotzdem liegt es nahe, den Median für ungerades n so festzulegen, wie wir dies getan haben, da $x_{((n+1)/2)}$ die mittlere Beobachtung ist, welche die Messwerte in zwei gleich große Hälften teilt.

Zweitens zeigt das Beispiel, dass der Median nicht notwendigerweise eindeutig festgelegt ist, denn für das Beispiel mit geradem Stichprobenumfang erfüllt jede Zahl zwischen 6 und 8, also zwischen $x_{(3)}$ und $x_{(4)}$, die Bedingung, dass die Hälfte aller Beobachtungen kleiner ist. Insofern wären beispielsweise auch die Werte 6,1, 6,5, 7,3 usw. legitime Stichprobenmediane. Trotzdem ist es eine weit verbreitete Konvention, für gerade n den Median als Mittel der beiden mittleren Beobachtungen festzulegen.

Der Median wird im Gegensatz zum Mittelwert nicht oder zumindest nur wenig von Ausreißern beeinflusst. Man kann sich dies klar machen, wenn man im Beispiel 2.4 den größten Wert noch weiter erhöht (oder auch den kleinsten Wert noch weiter verringert). Weder für einen geraden noch für einen ungeraden Stichprobenumfang wird dadurch der Median verändert. Der Betrag der Erhöhung (Verminderung) spielt dabei keine Rolle.

Die „Robustheit“ des Medians gegenüber Ausreißern bzw. Extremwerten ist ein nicht zu unterschätzender Vorteil des Medians gegenüber dem Mittelwert. Vermutet man untypische Beobachtungen in den Daten, sollte der Median als Kennwert der zentralen Tendenz verwendet werden.

Eine weitere Eigenschaft des Medians ist, dass die Summe der Abweichungsbeträge vom Median

Tabelle 2.2. Sortierte Bearbeitungszeiten in Sekunden

62,6	64,9	65,3	68,1	69,0	70,0	72,4	73,9	75,6
75,7	76,1	78,8	79,0	80,7	81,5	84,3	84,6	84,9
86,7	87,9	90,2	92,4	93,4	93,9	95,1	95,2	95,9
96,7	97,8	98,0	98,8	99,2	100,4	101,0	102,4	103,0
103,1	103,4	103,7	105,4	106,0	106,2	106,7	107,7	108,0
108,1	109,3	109,4	109,6	110,2	110,9	111,7	112,2	112,8
112,9	113,4	113,8	114,3	116,4	116,6	118,5	119,2	119,7
119,9	120,6	121,0	122,0	122,7	123,4	124,2	124,3	125,5
126,6	127,3	128,5	128,6	129,9	131,7	131,8	134,7	135,2
136,1	137,0	139,4	140,3	142,0	143,1	144,8	146,5	148,2

ein Minimum ergibt. Mit anderen Worten, gesucht ist ein Wert $\tilde{x}$, für den die Summe

$$\sum_{i=1}^{n} |x_i - \tilde{x}|$$

so klein wie möglich wird. Es lässt sich zeigen, dass das Ersetzen von $\tilde{x}$ durch den Median diese Summe minimiert.

BEISPIEL 2.5

Für die 90 Bearbeitungszeiten der Tab. 2.1 soll ebenfalls der Median berechnet werden. Die Berechnung von Hand wäre allerdings aufwändig, da zunächst die Ordnungsstatistik – also die nach ihrer Größe sortierten Bearbeitungszeiten – bestimmt werden müsste. Um die Berechung des Medians zu erleichtern, sind die sortierten Bearbeitungszeiten in Tab. 2.2 enthalten. Da der Stichprobenumfang $n = 90$ gerade ist, müssen die beiden mittleren Bearbeitungszeiten ermittelt werden. Dies sind die 45. und 46. Beobachtung. Der Median ist also

$$\text{Md} = \frac{x_{(45)} + x_{(46)}}{2} = \frac{108{,}0 + 108{,}1}{2} = 108{,}05.$$

Der Mittelwert für diese Daten betrug 106,89. Wie man erkennt, ergibt sich für den Median der Bearbeitungszeiten ein sehr ähnlicher Wert. Für dieses Beispiel beträgt der Unterschied zwischen den beiden Kennwerten nur etwas mehr als eine Sekunde.

■

2.1.3 Modalwert

Der Modalwert einer Verteilung ist derjenige Messwert, der am häufigsten vorkommt. Den Modalwert zu bestimmen ist also sehr einfach und bedarf keiner Formel. Es muss nur für jeden beobachteten Rohwert ausgezählt werden, wie oft er in der Stichprobe vertreten ist. Der Wert mit der größten Häufigkeit ist der Modalwert.

Der Modalwert wird in den Sozialwissenschaften kaum verwendet, da sich schnell Probleme bei seiner Berechnung ergeben können. Beispielsweise sind die Bearbeitungszeiten in Tab. 2.1 so genau gemessen worden, dass jeder Wert nur einmal be-

EXKURS 2.2 Weitere Maße der zentralen Tendenz

Geometrisches Mittel. Werden subjektive Empfindungsstärken gemittelt, kann man aufgrund psychophysischer Gesetzmäßigkeiten zeigen, dass die durchschnittliche Empfindungsstärke verschiedener Reize nicht durch das arithmetische Mittel, sondern besser durch das geometrische Mittel (GM) abgebildet wird. Soll beispielsweise in einem psychophysischen Experiment eine Versuchsperson die durchschnittliche Helligkeit von drei verschiedenen Lampen mit den Helligkeiten 100 Lux, 400 Lux und 1000 Lux einstellen, erwarten wir, dass die eingestellte durchschnittliche Helligkeit nicht dem arithmetischen Mittel (= 500 Lux), sondern dem geometrischen Mittel entspricht. Das geometrische Mittel setzt voraus, dass alle Werte positiv sind. Es wird nach folgender Beziehung berechnet:

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}.$$

Das geometrische Mittel in unserem Zahlenbeispiel lautet: $GM = \sqrt[3]{100 \cdot 400 \cdot 1000} = 342$.
Ein wichtiges Anwendungsfeld für das geometrische Mittel sind durchschnittliche Wachstumsraten wie beispielsweise durchschnittliche Umsatzsteigerungen pro Jahr, durchschnittliche Veränderungen der Bevölkerungszahlen pro Jahr oder durchschnittliche Preissteigerungen pro Jahr, wobei die Wachstumsrate als prozentuale Veränderung gegenüber dem Vorjahr definiert ist (ausführlicher hierzu vgl. z. B. Sixtl, 1993, S. 61 ff.).

Harmonisches Mittel. Ein Autofahrer fährt staubedingt 50 km mit einer Geschwindigkeit von 20 km/h und danach 50 km mit 125 km/h. Wie lautet die Durchschnittsgeschwindigkeit für die Gesamtstrecke von 100 km?
Die vielleicht spontan einfallende Antwort $(20\,\text{km/h} + 125\,\text{km/h})/2 = 72{,}5\,\text{km/h}$ ist falsch, denn die Durchschnittsgeschwindigkeit ergibt sich als Gesamtstrecke/Gesamtzeit. Für die 100 km benötigt der Fahrer $50/20 + 50/125 = 2{,}5 + 0{,}4 = 2{,}9$ Stunden, sodass sich eine Durchschnittsgeschwindigkeit von $100\,\text{km}/2{,}9\,\text{h} = 34{,}48\,\text{km/h}$ ergibt. Dieser Wert entspricht dem harmonischen Mittel der beiden Geschwindigkeiten. Die allgemeine Berechnungsvorschrift für das harmonische Mittel lautet:

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Berechnen wir das harmonische Mittel für das Beispiel, resultiert

$$\frac{2}{\frac{1}{20\,\text{km/h}} + \frac{1}{125\,\text{km/h}}} = 34{,}48\,\text{km/h}.$$

Das harmonische Mittel kommt zur Anwendung, wenn Indexzahlen (Kilometer pro Stunde, Preis pro Liter, Einwohner pro Quadratkilometer etc.) zu mitteln sind und die Zählervariable (Kilometer, Preis, Einwohnerzahl) konstant ist. Ist die Nennervariable (Fahrzeit, Litermenge, Flächengröße) konstant, ergibt sich der durchschnittliche Index über das arithmetische Mittel.

obachtet worden ist. In einem solchen Fall ist der Modalwert nicht definiert. Des Weiteren ist nicht ausgeschlossen, dass in einem Datensatz zwei verschiedene Werte die gleiche maximale Häufigkeit besitzen. Sind diese beiden Werte nicht unmittelbar nebeneinander, so liegt eine sog. *bimodale Verteilung* vor. Auch in diesem Fall ist unklar, wie der Modalwert bestimmt werden sollte. Darüber hinaus hat der Modalwert den Nachteil, dass er über vergleichbare Stichproben hinweg sehr variabel ist, d. h. sehr unterschiedliche Werte annehmen kann.

Trotzdem kann es gelegentlich von Interesse sein, den häufigsten Rohwert einer Stichprobe zu berichten. In diesem Fall kann dies einfach über die Verwendung des Begriffs „Modalwert“ kommuniziert werden. Wir werden den Modalwert mit Mo abkürzen.

2.2 Maße der Variabilität

Ähneln sich die Werte zweier Stichproben hinsichtlich ihrer zentralen Tendenz, können sie dennoch hinsichtlich der Variabilität ihrer Werte stark voneinander abweichen. Während Maße der zentralen Tendenz angeben, welcher Wert die Mitte bzw. das Zentrum aller Werte am besten repräsentiert, informieren Maße der Variabilität über die *Unterschiedlichkeit* der Werte.

Für die empirische Forschung sind Maße der Variabilität denen der zentralen Tendenz ebenbürtig. Ein wichtiges allgemeines Forschungsanliegen ist die Beantwortung der Frage, wie die bezüglich eines Merkmals angetroffene Unterschiedlichkeit von Personen oder anderen Untersuchungseinheiten erklärt werden kann. Wir stellen z. B. fest, dass Schüler unterschiedlich leistungsfähig sind, dass Patienten auf eine bestimmte Behandlung unterschiedlich gut ansprechen, dass Wähler unterschiedliche Parteien präferieren etc. und suchen nach Gründen, die für die Verschiedenartigkeit verantwortlich sein könnten. Viele statistische Verfahren zur Überprüfung von Hypothesen tragen dazu bei, auf diese Frage eine Antwort zu finden.

Das Bemühen, Unterschiedlichkeit erklären zu wollen, setzt jedoch zunächst voraus, dass sich die in einer Untersuchung festgestellten Unterschiede angemessen beschreiben oder quantifizieren lassen. Hierfür wurden verschiedene *Variabilitätsmaße* entwickelt.

2.2.1 Varianz

Ein wichtiges Maß zur Kennzeichnung der Variabilität von Messwerten ist die Varianz, deren Berechnung ein metrisches Merkmal voraussetzt.

Definition 2.1

Die Varianz einer Stichprobe des Umfangs n ist definiert als die Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel, dividiert durch $n - 1$. Wir bezeichnen die Stichprobenvarianz mit s^2.

Die Berechnungsformel der Varianz lautet

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{2.2}$$

Folgende Überlegung führt zu dieser Definition der Varianz: Da die Unterschiedlichkeit der Messwerte zum Ausdruck gebracht werden soll, ist es naheliegend, die Abweichungen der Beobachtungen vom Zentrum der Stichprobenwerte zu betrachten. Verwendet man den Mittelwert, um das Zentrum der Verteilung zu kennzeichnen, dann sind also die Abweichungen $(x_i - \bar{x})$ von Interesse. Um eine repräsentative Abweichung zu erhalten, mag man auf die Idee kommen, diese Abweichungen zu mitteln. Allerdings führt dies nicht zu einem geeigneten Variabilitätsmaß, da die Abweichungen sowohl positiv als auch negativ sind und sich somit wechselseitig eliminieren. Um die Abweichungen von ihrem Vorzeichen zu befreien, kann man sie quadrieren oder einfach deren Betrag verwenden. An dieser Stelle verfolgen wir den ersten Vorschlag und betrachten die *quadrierten Abweichungen*.

Da ein einzelner Kennwert zur Kennzeichnung der Variabilität aus den Stichprobenwerten errechnet werden soll, werden die quadrierten Abweichungen summiert, d. h. wir berechnen die Quadratsumme

$$\mathrm{QS} = \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.3}$$

Die Quadratsumme ist immer positiv und steigt mit zunehmenden Abweichungen vom Mittel an. Dies sind Eigenschaften, die ein geeignetes Maß der Variabilität erfüllen sollte. Die Quadratsumme hat aber den Nachteil, dass sie auch vom Stichprobenumfang abhängt, weswegen sie am Stichprobenumfang relativiert werden sollte. Am naheliegendsten ist es, die Quadratsumme durch den Stichprobenumfang n zu dividieren. Aus Gründen, deren Erläuterung an dieser Stelle zu weit gehen würden, dividiert man statt dessen durch $n - 1$. Den Ausdruck $n-1$ werden wir später (Exkurs 8.1) als „Freiheitsgrade" der Varianz kennenlernen.

Wie man durch den Vergleich mit der Formel (2.2) erkennt, kann die Varianz auch als $s^2 = \mathrm{QS}/(n-1)$ geschrieben werden.

BEISPIEL 2.6

Zur Illustration wollen wir die Varianz für zwölf Noten x_i, $i = 1, \ldots, 12$ ermitteln. Wir fertigen dazu folgendes Rechenschema an, wobei die Spaltensummen in der letzten Zeile stehen.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3,3	0,8	0,64
1,7	−0,8	0,64
2,0	−0,5	0,25
4,0	1,5	2,25
1,3	−1,2	1,44
2,0	−0,5	0,25
3,0	0,5	0,25
2,7	0,2	0,04
3,7	1,2	1,44
2,3	−0,2	0,04
1,7	−0,8	0,64
2,3	−0,2	0,04
30,0		7,92

Da die Summe der zwölf Noten 30,0 beträgt, ergibt sich für den Mittelwert $\bar{x} = 2{,}5$. Die erste Abweichung lautet somit $3{,}3 - 2{,}5 = 0{,}8$. Quadrieren wir diese Abweichung, resultiert der Wert 0,64. Wie man der letzten Zeile des Rechenschemas entnimmt, beträgt die Summe aller quadrierten Abweichungen 7,92. Damit ergibt sich für die Varianz der Wert

$$s^2 = \frac{\mathrm{QS}}{n-1} = \frac{7{,}92}{11} = 0{,}72.$$

■

Für die Berechnung der Quadratsumme ist folgende Formel oft hilfreich, da sie nicht die vorherige Berechnung des Mittels voraussetzt:

$$\mathrm{QS} = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\,. \tag{2.4}$$

Wie wird die Varianz interpretiert? Die Interpretation wird dadurch erschwert, dass die Varianz durch das Quadrieren nicht mehr die *Einheiten* der Messwerte besitzt. Die Stichprobenvarianz von 0,72 aus Beispiel 2.6 lässt sich deshalb nicht auf die Notenskala beziehen. Ein Maß der Variabilität, welches sich direkt aus der Varianz ableitet, das aber keine Schwierigkeiten bei der Interpretation bereitet, ist die Standardabweichung.

2.2.2 Standardabweichung

Mit der Varianz haben wir ein Maß, dem durch die Quadrierung der individuellen Abweichungen das Quadrat der ursprünglichen Einheit der Messwerte zugrunde liegt. Da ein solches Maß nur schwer interpretierbar ist, wird die Quadrierung wieder rückgängig gemacht, indem man die Wurzel der Varianz berechnet. Der so ermittelte Wert wird als Standardabweichung s bezeichnet. Die Standardabweichung berechnen wir also als

$$s = \sqrt{s^2}. \tag{2.5}$$

Gelegentlich wird die Standardabweichung auch als *Streuung* bezeichnet. Allerdings wird der Begriff „Streuung“ auch synonym mit „Variabilität“ verwendet, sodass seine Verwendung missverständlich sein könnte.

BEISPIEL 2.7

Um auch die Berechnung der Standardabweichung numerisch zu zeigen, greifen wir das Beispiel 2.6 erneut auf. Dort hatten wir die Varianz von zwölf Notenwerten berechnet. Wir erhielten $s^2 = 0{,}72$. Die Standardabweichung der Notenwerte lautet somit

$$s = \sqrt{0{,}72} = 0{,}85.$$

Die Standardabweichung drückt die Variabilität der beobachteten Werte auf der Notenskala aus.

■

Da die Standardabweichung die gleiche Einheit wie die Messwerte besitzt, kann ihre Größe direkt mit den Messwerten in Beziehung gesetzt werden. Dabei kann die Standardabweichung – wie der Name schon suggeriert – als eine „repräsentative“ Abweichung vom Zentrum der Verteilung interpretiert werden.

Für die Daten des Beispiels 2.6 errechneten wir $\bar{x} = 2{,}5$ und $s = 0{,}85$. Damit können wir sagen, dass die Abweichungen vom Notendurchschnitt etwa 0,85 betragen. Dies darf natürlich nicht so interpretiert werden, dass alle Abweichungen diesen Wert besitzen. Die Beträge der Abweichungen werden sowohl über als auch unter dem Wert 0,85 liegen. Insofern ist die Standardabweichung ein Maß, mit dem die Größe der Abweichungen gut repräsentiert wird. Genauere Aussagen sind allerdings ohne zusätzliche Annahmen nicht möglich. Wir werden aber auf die Interpretation der Standardabweichung später im Zusammenhang mit normalverteilten Merkmalen zurückkommen.

Die Standardabweichung sowie die Varianz sind die mit Abstand populärsten Maße der Variabilität. Obwohl die Varianz nicht einfach in Bezug zu den Messwerten zu interpretieren ist, ist sie doch für die Statistik von großer Bedeutung, da sie im Rahmen vieler Analyseverfahren in Anteile zerlegt wird, aus deren relativer Größe wichtige Schlussfolgerungen gezogen werden können. Trotz der großen Popularität von Varianz und Standardabweichung gibt es zahlreiche andere Maße der Variabilität, von denen einige kurz besprochen werden sollen.

2.2.3 AD-Streuung

Im Zusammenhang mit der Varianz bzw. der Standardabweichung hatten wir erläutert, dass Abweichungen vom Mittel zwar indikativ für die Variabilität der Daten sind, aber deren Summierung nicht zielführend ist, da die unterschiedlichen Vorzeichen der Abweichungen dazu führen, dass sie sich gegenseitig ausgleichen. Wir hatten deshalb die Vorzeichen durch Quadrieren der Abweichungen eliminiert. Das Vorzeichen lässt sich jedoch auch einfach dadurch eliminieren, dass man die Beträge der Abweichungen betrachtet. Mittelt man die Abweichungsbeträge, so erhält man die sog. AD-Streuung („average deviation“). Sie bringt den Durchschnitt der in Absolutbeträgen gemessenen Abweichungen aller Messwerte vom arithmetischen Mittel zum Ausdruck. Als Formel geschrieben:

$$\mathrm{AD} = \frac{\sum_{i=1}^{n} (|x_i - \bar{x}|)}{n}. \tag{2.6}$$

BEISPIEL 2.8

Um die Berechnung der AD-Streuung zu illustrieren, greifen wir auf die Daten in Beispiel 2.6 zurück. Dort wurden die Examensnoten von zölf Prüflingen betrachtet und deren Abweichungen vom Mittelwert bereits berechnet. Wir geben die Daten hier noch einmal wieder.

x_i	$\lvert x_i - \bar{x}\rvert$
3,3	0,8
1,7	0,8
2,0	0,5
4,0	1,5
1,3	1,2
2,0	0,5
3,0	0,5
2,7	0,2
3,7	1,2
2,3	0,2
1,7	0,8
2,3	0,2
30,0	8,4

Die Summe der Abweichungsbeträge beträgt also 8,4. Damit ergibt sich

$$\text{AD} = \frac{8{,}4}{12} = 0{,}70.$$

Die durchschnittliche Abweichung vom Mittel beträgt somit 0,70 Notenpunkte.

■

2.2.4 Variationsbreite

Das einfachste Variabilitätsmaß ist die Variationsbreite, der entnommen werden kann, wie groß der *Bereich* ist, in dem sich die Messwerte befinden. Die Variationsbreite ermittelt man, indem man die Differenz aus dem größten und kleinsten Wert bildet, d. h.

$$x_{(n)} - x_{(1)}.$$

Die Variationsbreite wird häufig auch mit dem englischen Wort „Range" bezeichnet.

Für die Beispieldaten aus Tab. 2.1 lauten die kürzeste bzw. längste Bearbeitungszeit 62,6 s und 148,2 s. Somit ist die Variationsbreite

$$148{,}2\,\text{s} - 62{,}6\,\text{s} = 85{,}6\,\text{s}.$$

Die Variationsbreite als Maß der Variabilität ist sehr sensitiv gegenüber Ausreißern, da schon ein einzelner extremer Wert die Variationsbreite erheblich vergrößern kann.

2.2.5 Interquartilbereich

Die Variationsbreite ist kein sehr nützliches Maß zur Charakterisierung der Variabilität, da sie stark durch Ausreißer beeinflusst wird. Stabiler sind eingeschränkte *Streubereiche*, bei denen ein gewisser Prozentsatz der größten und kleinsten Beobachtungen nicht berücksichtigt wird. Beispielsweise könnte man den Bereich kennzeichnen, in dem sich die mittleren 50% einer Verteilung befinden. Dies lässt sich bewerkstelligen, indem eine Messwertreihe mit Hilfe des Medians in zwei gleich große Hälften geteilt wird, wobei in der einen Hälfte alle Werte kleiner als der Median enthalten sind und die andere Hälfte die Werte größer als der Median enthält.

Wenn wir nun den Median der Werte berechnen, welche unterhalb des Medians liegen, so dürfen wir erwarten, dass wir einen Wert erhalten, unter dem etwa 25% der Beobachtungen liegen. Den so ermittelten Wert nennt man *unteren Angelpunkt* Q_1. Ganz analog kennzeichnet der Median aller Messwerte über dem Median den Wert, über dem etwa 25% der Beobachtungen liegen. Diese ist der *obere Angelpunkt* Q_3. Der Median entspricht Q_2. Gelegentlich werden die Angelpunkte auch nach ihrem Erfinder als „Tukey-Angelpunkte" bezeichnet (Tukey, 1977). Die englische Bezeichnung für Angelpunkte ist „Hinge".

Ein Variabilitätsmaß, welches sich robust gegenüber Ausreißern verhält, ist der Abstand der Angelpunkte, welcher auch als „Interquartilbereich" bezeichnet wird. Der Interquartilbereich ist also

$$\text{IQR} = Q_3 - Q_1.$$

Der IQR drückt die Länge des Bereichs aus, über den die mittleren 50% einer Rohwerteverteilung streuen. Das Akronym IQR steht für die englische Bezeichnung „Inter-Quartile-Range".

Wie werden die Angelpunke nun konkret berechnet? Zunächst werden die Rohdaten nach ihrer Größe sortiert und der Median bestimmt. Um den unteren Angelpunkt zu bestimmen, wird ein neuer Datensatz gebildet, der die Hälfte der sortierten Daten umfasst. Und zwar sind dies die Werte: $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$. Bei geradem Stichprobenumfang n ist $m = n/2$ und bei ungeradem n ist $m = (n+1)/2$. Der untere Angelpunkt Q_1 ist der Median dieses Datensatzes.

Der obere Angelpunkt wird ganz analog berechnet d. h., es wird der Datensatz $x_{(m)}, x_{(m+1)}, \ldots, x_{(n)}$ gebildet. Der obere Angelpunkt Q_3 ist der Median dieses Datensatzes.

BEISPIEL 2.9

Für die Bearbeitungszeiten, die in Tab. 2.2 bereits sortiert vorliegen, lassen sich die Angelpunkte direkt aus der Tabelle ablesen. Da der Stichprobenumfang mit $n = 90$ gerade ist, muss man, um Q_1 zu erhalten, nur den Median der 45 kleinsten Beobachtungen bestimmen. Mit anderen Worten $Q_1 = x_{(23)}$. Ganz analog ergibt sich $Q_3 = x_{(68)}$. Liest man die Werte aus Tab. 2.2 ab, so erhält man $Q_1 = 93{,}4$ s und $Q_3 = 122{,}7$ s. Somit beträgt der Interquartilbereich

$$\text{IQR} = 122{,}7\,\text{s} - 93{,}4\,\text{s} = 29{,}3\,\text{s}.$$

Die mittleren 50% der Bearbeitungszeiten streuen also über einen Bereich von fast 30 s.

■

Eine weiterführende Diskussion der Angelpunkte sowie verwandter Kennwerte findet man bei Hoaglin (1983).

2.2.6 MAD

Da der Mittelwert nicht robust gegenüber Ausreißern ist, liegt es nahe, Abweichungen vom Median zu betrachten. Wiederum eliminieren wir das Vorzeichen der Abweichungen, indem wir die Beträge betrachten, also $|x_i - \text{Md}|$, für $i = 1, \ldots, n$. Der Median dieser Abweichungsbeträge wird MAD genannt. Das MAD-Maß ist also der Median der absoluten Abweichungen vom Median, wobei MAD die Abkürzung für „median absolut deviation from the median" ist.

Als Formel wird es folgendermaßen geschrieben:

$$\text{MAD} = \text{Md}(|x - \text{Md}|).$$

BEISPIEL 2.10

Auch die MAD-Streuung soll an einem Beispiel verdeutlicht werden. Dazu greifen wir erneut die Daten auf, welche bereits in den Beispielen 2.6 und 2.8 verwendet wurden. Die Notenwerte sind in folgender Übersicht enthalten.

x_i	$\lvert x_i - \text{Md}\rvert$
3,3	1,0
1,7	0,6
2,0	0,3
4,0	1,7
1,3	1,0
2,0	0,3
3,0	0,7
2,7	0,4
3,7	1,4
2,3	0,0
1,7	0,6
2,3	0,0

Der Median der zwölf Notenpunkte beträgt 2,3. Für die absolute Abweichung vom Median ergibt sich beispielsweise für den ersten Wert $|3{,}3 - 2{,}3| = 1{,}0$. Alle absoluten Abweichungen sind ebenfalls in der oben dargestellten Tabelle enthalten. Berechnet man den Median, indem man die absoluten Abweichungen zuerst sortiert und dann aufgrund des geraden Stichprobenumfangs das Mittel der beiden mittleren Abweichungen berechnet, so erhält man MAD = 0,6. Dieser Wert ist der AD-Streuung, für die wir 0,7 berechneten, sehr ähnlich. ■

Weitere Informationen zum MAD-Maß findet man bei Maronna et al. (2006).

2.3 Stichprobenperzentile

Perzentile sind Kennwerte, die in vielen Kontexten der Statistik eine Rolle spielen. Ein Perzentil bringt die *relative Position* eines Messwertes innerhalb der Stichprobe zum Ausdruck. Wie die folgende Definition erläutert, bezieht sich ein Perzentil immer auf einen vorgegebenen Prozentsatz.

Definition 2.2

Stichprobenperzentil. Das Perzentil einer Stichprobe x_p ist der Messwert, unter dem p-Prozent der Werte in der Stichprobe liegen.

Beispielsweise bezeichnet $x_{30\%}$ den Wert, unterhalb dem 30% der Stichprobe liegen, und $x_{50\%}$ bezeichnet den Wert, unterhalb dem 50% der Stichproben liegen. Insofern entspricht $x_{50\%}$ dem Median.

Perzentile können verwendet werden, um Bereiche zu kennzeichnen, in denen ein bestimmter Prozentsatz der Stichprobe liegt. Oftmals wird dabei die Stichprobe in gleich große Anteile zerlegt.

Beispielsweise sind die drei Perzentile $x_{25\%}$, $x_{50\%}$ und $x_{75\%}$ diejenigen Werte, welche die Stichprobe in vier gleich große Anteile zerlegen. Man spricht deshalb auch von „Quartilen" bzw. dem ersten, zweiten und dritten Quartil. (Den oben besprochenen unteren bzw. oberen Angelpunkt bezeichnet man auch als erstes bzw. drittes „Pseudo-Quartil".) Ganz analog werden die neun Perzentile $x_{10\%}, x_{20\%}, \ldots, x_{90\%}$ auch *Dezile* genannt. Viele Autoren verwenden anstelle des Begriffs „Perzentil" den Begriff „Quantil".

Die praktische Berechnung von Perzentilen wird dadurch erschwert, dass es nicht nur eine einzige Berechnungsvorschrift gibt. Dies liegt daran, dass es einer Konvention bedarf, um Perzentile für eine beliebige Prozentangabe bestimmen zu können, es aber keine allgemein anerkannte „beste" Konvention dafür gibt.

Dass die Bestimmung von Perzentilen für Stichprobendaten oft nicht eindeutig möglich ist, hatten wir schon im Zusammenhang mit dem Median gesehen, der für eine gerade Anzahl von Beobachtungen als arithmetisches Mittel der beiden Werte $x_{(n/2)}$ und $x_{(n/2+1)}$ definiert wurde. Natürlich liegt diese Definition nahe, trotzdem könnte aber jeder Wert im Intervall zwischen $x_{(n/2)}$ und $x_{(n/2+1)}$ genauso gut als Median verwendet werden. Der Median ist also für gerades n nicht eindeutig bestimmt und wird erst durch die Festlegung einer Berechnungsvorschrift eindeutig bestimmbar.

Als zweites Beispiel stelle man sich eine Stichprobe vor, welche zehn Werte umfasst, wobei wir zur Vereinfachung annehmen, dass jeder Wert nur einmal aufgetreten ist. Fragen wir beispielsweise nach $x_{10\%}$, so könnte man einen beliebigen Wert wählen, der zwischen den beiden kleinsten be-

obachteten Werten liegt, also zwischen $x_{(1)}$ und $x_{(2)}$. Dieser Wert hätte die Eigenschaft, dass genau 10% der Werte kleiner und 90% der Werte größer als dieser Wert wären. Aber wie lassen sich $x_{11\%}, x_{12\%}, \ldots$ sinnvoll festlegen? Hierzu bedarf es wiederum einer Konvention.

Eine nahe liegende Vorgehensweise, um Perzentile für beliebige Prozentanteile bestimmen zu können, ist die „lineare Interpolation". Wir illustrieren den Grundgedanken mit einem Beispiel.

BEISPIEL 2.11

Nehmen wir an, es liegt eine Stichprobe von neun Testwerten vor, die mit Hilfe eines Fragebogens, der 20 „Ja-Nein" Fragen enthält, ermittelt wurden. Jede Ja-Antwort wird dabei als ein Punkt gewertet. Die folgenden neun Werte wurden beobachtet, wobei die Testwerte bereits nach ihrer Größe sortiert wurden:

$$2, 3, 5, 9, 10, 12, 14, 15, 19.$$

Wir betrachten nun die Abb. 2.1, in der die nach Größe sortierten Messwerte gegen die Stützstellen $p_k = k/(n+1)$ abgetragen sind. Da unsere Stichprobe neun Werte umfasst, sind die neun Stützstellen die Werte 0,1, 0,2, ... 0,9. Die so festgelegten Punkte der Abbildung werden nun durch Geradensegmente (lineare Interpolation) miteinander verbunden. Mit Hilfe dieser Abbildung kann man für einen beliebigen auf der Abszisse wählbaren Anteil das entsprechende Perzentil ablesen. Beispielsweise lassen sich die drei Quartile anhand der Abbildung bestimmen. Man erhält: $x_{25\%} = 4$, $x_{50\%} = 10$ und $x_{75\%} = 14{,}5$.

$x_{25\%}$ und $x_{75\%}$ sind nicht notwendigerweise mit den oben dargestellten Angelpunkten Q_1 und Q_3 identisch. Trotzdem gilt ganz allgemein: $x_{25\%} \approx Q_1$ und $x_{75\%} \approx Q_3$. Berechnet man die Angelpunkte, so erhält man $Q_1 = 5$ und $Q_3 = 14$.

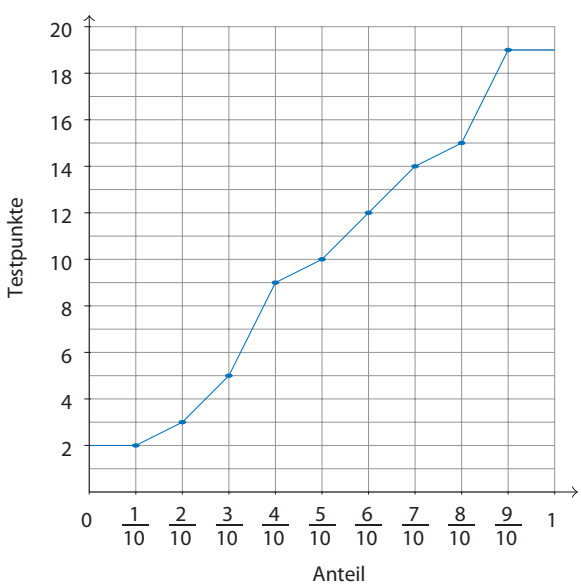

Abbildung 2.1. Lineare Interpolation zwischen den sortierten Testpunkten

Die unterschiedlichen Konventionen, Perzentile linear zu interpolieren, unterscheiden sich hauptsächlich in der Wahl der Stützstellen p_k, die zur Interpolation benötigt werden. Detaillierte Information zur Wahl der Stützstellen p_k findet man bei Hyndman und Fan (1996).

SOFTWAREHINWEIS 2.1

Die Berechnung der Perzentile mit Hilfe von linearer Interpolation ist relativ mühsam, sodass ein Statistikprogramm dafür verwendet werden sollte. In diesem Zusammenhang ist von Interesse, welche Wahl der Stützstellen p_k durch das Programm getroffen wird bzw. welche Möglichkeiten der Auswahl zur Verfügung stehen.

Die im obigen Beispiel gewählten Stützstellen $p_k = k/(n+1)$ entsprechen der Default-Einstellung (HAVERAGE) der SPSS-Prozedur „EXAMINE". In der R-Funktion „quantile()" kann die Wahl der Stützstellen über den „type"-Parameter beeinflusst werden. Man erhält das obige Ergebnis für „type=6".

BEISPIEL 2.12

Berechnet man mit SPSS für die Daten aus Tab. 2.1 die Perzentile $x_{5\%}, x_{10\%}, \ldots$ so erhält man die folgenden Ergebnisse:

Prozent	Perzentil
5	68,595
10	75,610
25	93,150
50	108,050
75	122,875
90	136,010
95	142,495

Wir wollen eines der Perzentile zur Illustration mit Hilfe linearer Interpolation nachrechnen. Wir greifen willkürlich $x_{5\%}$ heraus.

Für einen Stichprobenumfang von $n = 90$ gibt es keinen Messwert, unter dem genau 5% der Beobachtungen liegen. Wiederum gehen wir von den Stützstellen $p_k = k/(n+1)$ aus, wobei $k = 1, \ldots, 90$, da die genannten Werte mit Hilfe der Default-Einstellung von SPSS produziert wurden. Berechnet man die Stützstellen für $k = 1, 2, 3, 4, 5$, so erkennt man, dass 5% zwischen $p_4 = 4/91 = 0{,}044$ und $p_5 = 5/91 = 0{,}055$ liegt.

Aus den sortierten Bearbeitungszeiten ergibt sich, dass die viert- und fünftschnellste Bearbeitungszeit 68,1 s und 69,0 s betragen. Die Steigung der Geraden, anhand der linear interpoliert wird, beträgt somit

$$b = \frac{x_{(5)} - x_{(4)}}{p_5 - p_4} = \frac{69{,}0 - 68{,}1}{(5/91) - (4/91)} = 81{,}9.$$

Nun erhält man das gesuchte 5%-Perzentil, indem man folgenden Ausdruck berechnet:

$$\begin{aligned} x_{5\%} &= b \cdot (0{,}05 - p_4) + x_{(4)} \\ &= 81{,}9 \cdot (0{,}05 - 4/91) + 68{,}1 = 68{,}595. \end{aligned}$$

Das Ergebnis entspricht genau dem am Anfang des Beispiels genannten Wert.

2.4 Transformierte Messwerte

2.4.1 Kennwerte transformierter Messwerte

Häufig werden aus den beobachteten Messwerten neue Werte berechnet. Wir fragen deshalb, in welcher Beziehung die Kennwerte der transformierten Werte, welche wir mit y bezeichnen, zu den Kennwerten der ursprünglichen x-Werte stehen. Zu fordern ist, dass sich die Kennwerte der y-Werte auf sinnvolle Weise aus den Kennwerten der x-Werte ergeben. Dies ist z. B. für den Mittelwert der Fall.

Wenn zu jedem x-Rohwert die gleiche Konstante a addiert wird, so erwarten wir, dass sich auch der Mittelwert um diese Konstante ändert. Dies ist in der Tat der Fall. Mit anderen Worten, das Mittel der Werte

$$x_1 + a, \ldots, x_n + a$$

beträgt $\bar{x}+a$. Des Weiteren ist zu fordern, dass sich durch die Multiplikation der Messwerte mit einer Konstanten b der Mittelwert der transformierten Messwerte um den gleichen Faktor verändert. Mit anderen Worten, das Mittel von

$$b \cdot x_1, \ldots, b \cdot x_n$$

beträgt $b \cdot \bar{x}$.

Beide Ergebnisse kann man in folgender Formel zusammenfassen: Transformiert man x-Werte linear mit Hilfe der Gleichung $y_i = a+b\cdot x_i$, so lässt sich das Mittel der transformierten Werte mit Hilfe der Beziehung

$$\bar{y} = a + b \cdot \bar{x} \tag{2.7}$$

bestimmen.

Nun wollen wir fragen, wie sich die Varianz durch eine additive Konstante verändert. Wird zu jedem x-Wert eine Konstante a addiert, so erhöht dies die Variabilität der Messwerte nicht. Mit anderen Worten, die Varianz der transformierten Werte ist identisch mit der Varianz der x-Werte.

Werden die x-Werte dagegen alle mit der Konstanten b multipliziert, so verändert dies die Variabilität der Daten, sobald $b \neq 1$. Und zwar besitzt die Varianz der Werte

$$b \cdot x_1, \ldots, b \cdot x_n$$

den Wert $b^2 \cdot s_x^2$.

Wiederum kann man beide Ergebnisse in einer Formel zusammenfassen: Transformiert man x-Werte linear mit Hilfe der Gleichung $y_i = a+b\cdot x_i$, so lässt sich die Varianz der transformierten Werte mit Hilfe der Beziehung

$$s_y^2 = b^2 \cdot s_x^2 \tag{2.8}$$

bestimmen. Für die Standardabweichung der y-Werte ergibt sich die Beziehung

$$s_y = |b| \cdot s_x. \tag{2.9}$$

2.4.2 z-Transformation

Gelegentlich steht man vor der Aufgabe, den Testwert einer Person mit den Testwerten anderer Personen in Beziehung zu setzen, um zu beurteilen, ob es sich bei diesem Wert um einen „hohen“ bzw. „niedrigen“ Wert handelt. Im Alltag verwenden wir oft den Mittelwert als Referenzpunkt und bezeichnen einen Wert als über- oder unterdurchschnittlich. Um zu genaueren Aussagen zu gelangen, könnte man die Dezile bestimmen und dann feststellen, zwischen welchen Dezilen sich der Testwert der Person befindet. Allerdings wird häufig ein anderes Vorgehen gewählt, das den Mittelwert sowie die Standardabweichung der Stichprobe verwendet, um den Testwert zu transformieren. Es handelt sich dabei um die z-Transformation.

Zunächst betrachten wir wieder die Abweichung vom Mittel, wobei diesmal das Vorzeichen der Abweichung von Bedeutung ist, da es erkennen lässt, ob der Wert über- bzw. unterdurchschnittlich ist. Beispielsweise möge die Körpergröße einer männlichen Person 190 cm betragen. Wenn die durchschnittliche Größe der männlichen Personen in der Stichprobe 175 cm beträgt, so entspricht die Abweichung vom Mittel +15 cm. Da wir mit den Einheiten des Längenmaßes vertraut sind, ist diese Aussage bereits informativ. Allerdings dürfte es nicht leicht fallen, zu beurteilen, ob diese Abweichung „gewöhnlich“ ist. Dies wäre dann der Fall, wenn viele Personen um 15 cm oder mehr vom Mittel abweichen würden. Es wäre auch möglich, dass eine Abweichung von 15 cm im Vergleich zu den anderen Stichprobenwerten bereits so groß ist, dass wir sie als Ausreißer betrachten sollten. Da wir durch die Maße der Variabilität repräsentative Abweichungen leicht bestimmen können, liegt es nahe, die Abweichung vom Mittel an einem Maß für die Variabilität der Werte zu *relativieren*.

Die sog. *z-Werte* erhält man, indem man die Abweichung vom Mittel an der Standardabweichung relativiert. Man berechnet also

$$z = \frac{x - \bar{x}}{s}. \tag{2.10}$$

Definition 2.3

z-Transformation. Das Umrechnen des Rohwertes x in den z-Wert mit Hilfe von Gl. (2.10) wird auch „z-Transformation“ genannt. Somit gibt der z-Wert an, um wie viele Standardabweichungen ein Rohwert unter bzw. über dem Mittelwert liegt.

Haben wir in der Stichprobe männlicher Personen, deren Körpergrößen ermittelt wurden, eine Standardabweichung von 15 errechnet, so können wir den z-Wert einer Person mit einer Größe von 190 cm berechnen. Wir erhalten

$$z = \frac{190\,\text{cm} - 175\,\text{cm}}{15\,\text{cm}} = 1.$$

Der z-Wert dieser Person beträgt also 1,0. Dies bedeutet, dass ihr Rohwert den Mittelwert um die Länge einer Standardabweichung übersteigt. Da es sich bei der Standardabweichung – wie der Name schon sagt – um eine repräsentative Abweichung handelt, ist eine Größe von 190 cm sicherlich noch kein extrem großer Wert im Vergleich zu den anderen Körpergrößen, die sich in der Stichprobe befinden. Ein z-Wert von 0,0 entspricht einer durchschnittlichen Ausprägung des Rohwertes. Ein negativer z-Wert zeigt einen unterdurchschnittlichen Rohwert an.

Wie man an dem Beispiel gut erkennen kann, besitzen z-Werte nicht mehr die Einheiten der Rohwerte. Sie sind also *dimensionslose* Zahlen.

In Aufgabe 2.7 überlegen wir uns, weshalb für z-Werte folgende Aussage gilt:

z-transformierte Werte haben einen Mittelwert von 0 und eine Standardabweichung von 1.

BEISPIEL 2.13

Im Beispiel 2.6 hatten wir zwölf Notenwerte betrachtet, für die wir bereits Mittel und Standardabweichung bestimmt haben. Wir ermittelten $\bar{x} = 2{,}5$ und $s = 0{,}85$. Mit Hilfe dieser Werte bestimmen wir nun den z-Wert der ersten Person, für die eine Note von 3,3 berichtet wurde. Wir erhalten

$$z = \frac{3{,}3 - 2{,}5}{0{,}85} = 0{,}94.$$

Da der z-Wert dieser Person 0,94 beträgt, ist die Note um fast eine Standardabweichung höher als der Durchschnitt.

■

Folgende Anwendungen für z-Werte sind weit verbreitet:

1. Es sollen die Werte zweier Personen verglichen werden, die zu unterschiedlichen Stichproben bzw. Gruppen gehören. Beispielsweise möchte man die Examensnoten zweier Personen vergleichen, die zu unterschiedlichen Jahrgängen gehören. Selbst wenn beide Personen die gleiche Note erhielten, ist nicht auszuschließen, dass die Examensbedingungen beim älteren Jahrgang einfacher (oder schwerer) waren, sodass die beiden Leistungen nicht ohne Weiteres gleichgesetzt werden können. Mit Hilfe der z-Transformation lassen sich die Werte aber vergleichbar machen.

2. Die Grade der relativen Merkmalsausprägung zweier Merkmale einer Person sollen miteinander verglichen werden. So könnte man für jede Person nicht nur die Körpergröße, sondern auch deren Gewicht in Kilogramm ermitteln. Da beide Variablen unterschiedliche Einheiten besitzen, lassen sie sich nicht direkt vergleichen. Durch die z-Transformation wird die relative Position einer Person aber dimensionslos zum Ausdruck gebracht, sodass der Vergleich durch z-Werte sinnvoll ist. Hätte eine Person z. B. sowohl für das Merkmal Körpergröße als auch für das Merkmal Gewicht einen z-Wert von 1,0, so könnten wir uns zumindest ein ungefähres Bild von der Person machen, denn sie wäre in vergleichbarem Ausmaß überdurchschnittlich groß und schwer. Es handelte sich also um eine große Person, bei der das Verhältnis von Größe zu Gewicht aber „normal“ sein dürfte. Wie sähe eine Person aus, deren Körpergröße einem z-Wert von 1,0 entspricht, aber deren Körpergewicht einem z-Wert von −1,0 entspräche? Diese Person wäre überdurschnittlich groß, hätte aber ein unterdurchschnittliches Gewicht. Es müsste sich also um eine große, schlanke Person handeln.

ÜBUNGSAUFGABEN

Summenzeichen

Aufgabe 2.1 Gegeben sind die fünf Werte $x_1 = 1$, $x_2 = 4$, $x_3 = 5$, $x_4 = 8$, $x_5 = 10$. Berechnen Sie folgende Summen: a) $\sum_{i=1}^{5} x_i$, b) $\sum_{i=1}^{5} x_i^2$, c) $(\sum_{i=1}^{5} x_i)^2$, d) $\sum_{i=2}^{5} x_i$, e) $\sum_{i=1}^{5} x_i + 5$, f) $\sum_{i=1}^{5} (x_i + 5)$, g) $\sum_{i=1}^{5} (2x_i)$, h) $\sum_{i=2}^{4} (x_i + i^2)$ und i) $\sum_{i=1}^{5} (x_3 + i^2)$.

Aufgabe 2.2 Formen Sie folgende Ausdrücke um: a) $\sum_{i=1}^{n} (x_i + a)$, b) $\sum_{i=1}^{n} bx_i$ und c) $1/n \sum_{i=1}^{n} (a + bx_i)$.

Statistische Kennwerte

Aufgabe 2.3 Bei einer Erhebung der Intelligenz von 20 Studenten fallen folgende Werte an:

109	92	93	94	96
96	97	98	100	101
101	102	103	103	103
104	105	105	107	91

Berechnen Sie:
a) Mittelwert, Median und Modalwert
b) QS, Varianz und Standardabweichung
c) AD-Streuung
d) MAD-Streuung
e) beide Tukey-Angelpunkte sowie den IQR

Transformationen

Aufgabe 2.4 Fünf Personen bearbeiten einen psychologischen Test. Es treten folgende Messwerte auf:

$$x_1 = 80, x_2 = 70, x_3 = 60, x_4 = 50 \text{ und } x_5 = 40.$$

a) Berechnen Sie Mittelwert und Standardabweichung.
b) Standardisieren Sie die Testwerte mit Hilfe der z-Transformation.
c) Berechnen Sie Mittelwert und Standardabweichung der z-Werte.

Aufgabe 2.5 Eine Reihe von Messwerten besitzt einen Mittelwert von 10 und eine Standardabweichung von 3. Die Messwerte werden anhand der Gleichung $y = 4 \cdot x + 25$ transformiert. Berechnen Sie a) $\bar{y}$, b) s_y^2 und c) s_y.

Aufgabe 2.6 Zeigen Sie, dass der Mittelwert einer Stichprobe von x-Werten mit dem Mittelwert der durch die Transformation $y = b \cdot x + a$ gewonnenen y-Werte in der Beziehung $\bar{y} = b \cdot \bar{x} + a$ steht.

Aufgabe 2.7 Zeigen Sie, dass für z-transformierte Werte gilt: $\bar{z} = 0$ und $s_z = 1{,}0$, wobei s_z die Standardabweichung der z-Werte bezeichnet.

Verschiedenes

Aufgabe 2.8 Eine Möglichkeit die Quadratsumme von n Werten zu berechnen, ist durch folgende Formel gegeben:

$$\text{QS} = \frac{1}{n} \sum_{i<j} (x_i - x_j)^2.$$

Die Summation erfolgt über alle möglichen Wertepaare, wobei jedes Paar nur einmal berücksichtigt wird. Berechnen Sie zunächst die Quadratsumme der fünf Werte $1, 2, 3, 4, 5$ mit Hilfe der Formel (2.3). Überprüfen Sie dann das Ergebnis mit der oben angegebenen Formel.

Aufgabe 2.9 Um die kleinste-Quadrate Eigenschaft des Mittels zu illustrieren, berechne man für die in Beispiel 2.1 enthaltenen Fehlerzahlen ($3, 5, 6, 8$ und 14) die Quadratsumme

$$\sum_i (x_i - \tilde{x})^2,$$

wobei für $\tilde{x}$ die Werte a) 7,1, b) 7,2, c) 7,3, d) 7,4 und e) 7,5 einzusetzen sind.