

STATISTIK I – Übung 01

Das arithmetische Mittel

1 Kurze Wiederholung

Grundlagen

Das arithmetische Mittel gehört zu den Lagemaßen und ist somit als eine Maßzahl zu verstehen, die das Zentrum einer Verteilung beschreibt. Als das bekannteste Lagemaß wird das arithmetische Mittel häufig auch als „das Standardmittel“ oder einfach nur als „der Mittelwert“ bezeichnet. Seine Berechnung setzt voraus, dass die Daten der Verteilung mindestens metrisch skaliert sind – was in der Praxis (etwa bei Schulnoten) bedauerlicherweise häufig übersehen wird. Ist diese Vorbedingung erfüllt, berechnet sich das arithmetische Mittel wie folgt:

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

mit:

$\bar{x}$ = *Arithmetisches Mittel*

x_i = *Einzelner Wert der Verteilung*

n = *Anzahl der Werte der Verteilung*

Arithmetisches Mittel bei klassierten Daten

Neben der „einfachen“ Berechnung des arithmetischen Mittels wird für unsere Klausur noch die Berechnung des arithmetischen Mittels bei klassierten Daten von Relevanz sein. Hierfür wird auf folgende Formel zurückgegriffen:

$$\overline{x_g} = \sum_{i=1}^k m_i * f_i$$

mit:

$\overline{x_g}$ = *Arithmetisches Mittel (bei klassierten/gruppierten Daten)*

f_i = *relative Häufigkeit (der Klasse i)*

m_i = *Klassenmitte (der Klasse i)*

k = *Anzahl der Klassen*

Robustheit und getrimmtes arithmetisches Mittel

Das arithmetische Mittel wird von (univariaten) Ausreißern (besonders großen oder kleinen Werten im Datensatz) ganz erheblich beeinflusst und wird deshalb auch als „nicht robust“ bezeichnet. Diesen Effekt kann man sich leicht vor Augen führen, indem man sich klarmacht, dass das arithmetische Mittel der Verteilung [1; 2; 3; 4] bei 2,5 liegt, das arithmetische Mittel aus der Verteilung [1; 2; 3; 50] aber bei 14. Bereits ein einzelner besonders großer oder kleiner Wert (wie etwa ein exorbitantes Einkommen bei einer Vermögenserhebung), kann daher den Mittelwert deutlich nach oben oder unten verzerren.

Um diesen Effekt zu begrenzen, kann man entweder auf ein anderes Lagemaß (wie etwa den Median) ausweichen, oder das sogenannte getrimmte arithmetische Mittel berechnen. Hierbei wird der Datensatz vor der Berechnung des arithmetischen Mittels um eine gewisse Anzahl an Werten an den Rändern der Verteilung (symmetrisch) gekürzt, um Ausreißer aus dem Datensatz zu eliminieren. Bei einem Datensatz mit 100 Werten würden bei einer Trimmung um 5% zum Beispiel die 5 größten sowie die 5 kleinsten Werte aus dem Datensatz entfernt und anschließend das arithmetische Mittel auf Basis der bereits bekannten Formel neu berechnet. Dabei ist zu beachten, dass in vielen Fällen auch Nicht-Ausreißer aus den Daten gestrichen werden, die man im Grunde behalten möchte.

Hinweis zu softwaregestützten Analysen

Wird für die Berechnung des arithmetischen Mittels eine Software wie etwa SPSS, PSPP oder PAST eingesetzt, so ist – wie bei vielen anderen Berechnungen auch – zu berücksichtigen, dass die Erfüllung von Vorbedingungen für die Analyse in der Regel nicht von der Software geprüft wird. In diesem Fall betrifft dies das Vorliegen eines metrischen Skalenniveaus. SPSS berechnet das arithmetische Mittel fälschlicherweise nicht nur für Schulnoten, sondern auch für Telefonnummern oder Geschlechter (falls diese mit Zahlen codiert sein sollten) – auch wenn die Ergebnisse vollkommen sinnbefreit sind. Als ganz besonders gefährlich dürfen übrigens solche Fehler betrachtet werden, die – zumindest oberflächlich betrachtet – sinnvolle Ergebnisse darzustellen scheinen (wie eben beim Mittel aus Schulnoten). Beim Einsatz von Software ist daher entscheidend, dass der Anwender / die Anwenderin über die Methodenkenntnisse verfügt, um selbst beurteilen zu können, wann welche Methode zulässig ist.

2 Beispielrechnungen

Arithmetisches Mittel

Für eine Gruppe von Studierenden liegt die folgende Altersverteilung vor:

Alter	Absolute Häufigkeit	Relative Häufigkeit	Kumulierte abs. Häufigkeit	Kumulierte rel. Häufigkeit
21 Jahre	5	0,25	5	0,25
22 Jahre	4	0,20	9	0,45
23 Jahre	3	0,15	12	0,60
24 Jahre	4	0,20	16	0,80
25 Jahre	4	0,20	20	1,00
Summe	20	1,00	20	1,00

Das arithmetische Mittel berechnet sich in diesem Fall wie folgt:

$$21+21+21+21+21+22+22+22+22+23+23+23+24+24+24+24+25+25+25+25 = 458$$

$$458 / 20 = \underline{22,9}$$

Alternative Vereinfachung:

$$(21 \cdot 5) + (22 \cdot 4) + (23 \cdot 3) + (24 \cdot 4) + (25 \cdot 4) = 458$$

$$458 / 20 = \underline{22,9}$$

Das arithmetische Mittel liegt somit bei 22,9 Jahren.

Arithmetisches Mittel bei klassierten Daten

Für eine Gruppe von Studierenden liegt die folgende Größenverteilung vor:

Größe	Absolute Klassenhäufigkeit	Relative Klassenhäufigkeit	Absolute kum. Klassenhäufigkeit	Relative kum. Klassenhäufigkeit
[1,55 m – 1,65 m)	12	0,24	12	0,24
[1,65 m – 1,75 m)	16	0,32	28	0,56
[1,75 m – 1,85 m)	22	0,44	50	1,00
Summe	50	1,00	50	1,00

Das arithmetische Mittel berechnet sich in diesem Fall wie folgt:

$$(0,24 * 1,60) + (0,32 * 1,70) + (0,44 * 1,80) = \underline{1,72}$$

Das arithmetische Mittel liegt somit bei 1,72 Metern.

Getrimmtes arithmetisches Mittel

Eine Umfrage unter 10 Personen zum monatlichen Bruttoeinkommen erbrachte folgende Ergebnisse:

2.250 EUR	2.130 EUR
2.320 EUR	2.640 EUR
2.400 EUR	2.550 EUR
2.140 EUR	2.250 EUR
17.380 EUR	2.710 EUR

Das arithmetische Mittel berechnet sich in diesem Fall wie folgt:

$$2250 + 2320 + 2400 + 2140 + 17380 + 2130 + 2640 + 2550 + 2250 + 2710 = 38770$$

$$38770 / 10 = \underline{3877}$$

Das arithmetische Mittel liegt bei 3.877 EUR. Da es offenkundig vom Ausreißer stark beeinflusst wird (alle befragten Personen außer einer verdienen zwischen 2.100 und 2.800 EUR – trotzdem liegt der „Mittelwert“ bei fast 4.000 EUR), soll nachfolgend das um 10% getrimmte Mittel berechnet werden.

Bei einer 10%igen Trimmung sind der größte (17.380 EUR) und der kleinste (2.130 EUR) Wert aus dem Datensatz zu entfernen. Es ergibt sich die folgende neue Grundtabelle:

2.250 EUR	2.640 EUR
2.320 EUR	2.550 EUR
2.400 EUR	2.250 EUR
2.140 EUR	2.710 EUR

Das getrimmte arithmetische Mittel berechnet sich in diesem Fall wie folgt:

$$2250 + 2320 + 2400 + 2140 + 2640 + 2550 + 2250 + 2710 = 19260$$

$$19260 / 8 = \underline{2407,5}$$

Das getrimmte arithmetische Mittel liegt somit (deutlich realitätsnäher) bei 2.407,50 EUR.

3 Übungsaufgaben

Arithmetisches Mittel

Aus einem Produktionslos von 1.000 Karosserieteilen wird eine Stichprobe von 20 Teilen gezogen und gewogen. Es ergeben sich die folgenden Werte:

Teil 1	1,72 kg	Teil 6	1,74 kg	Teil 11	1,76 kg	Teil 16	1,76 kg
Teil 2	1,74 kg	Teil 7	1,73 kg	Teil 12	1,77 kg	Teil 17	1,77 kg
Teil 3	1,72 kg	Teil 8	1,73 kg	Teil 13	1,71 kg	Teil 18	1,71 kg
Teil 4	1,72 kg	Teil 9	1,76 kg	Teil 14	1,72 kg	Teil 19	1,75 kg
Teil 5	1,73 kg	Teil 10	1,72 kg	Teil 15	1,77 kg	Teil 20	1,76 kg

a) Fassen Sie diese Werte in einer kumulierten Häufigkeitstabelle (ohne Klassierung) zusammen.

b) Berechnen Sie das arithmetische Mittel.

Arithmetisches Mittel bei klassierten Daten

Eine Gruppe von 50 Studierenden wird nach ihrem ungefähren Lernaufwand für eine Statistikklausur (in Tagen) befragt. Es ergeben sich die folgenden (klassierten) Werte:

Größe	Absolute Klassenhäufigkeit	Relative Klassenhäufigkeit	Absolute kum. Klassenhäufigkeit	Relative kum. Klassenhäufigkeit
[1 Tag – 3 Tage)	17			
[3 Tage – 5 Tage)	23			
[5 Tage – 7 Tage)	10			
Summe				

a) Füllen Sie die restlichen Felder der kumulierten Häufigkeitstabelle aus.

b) Berechnen Sie das arithmetische Mittel.

Getrimmtes arithmetisches Mittel

Eine Gruppe von Studierenden befragt Passantinnen und Passanten auf dem Campus. Erhoben wird dabei unter anderem das Alter (in Jahren). Hierfür ergeben sich für zwanzig Personen folgende Werte:

Person 1	23 Jahre	Person 11	22 Jahre
Person 2	24 Jahre	Person 12	23 Jahre
Person 3	19 Jahre	Person 13	37 Jahre
Person 4	21 Jahre	Person 14	30 Jahre
Person 5	23 Jahre	Person 15	85 Jahre
Person 6	26 Jahre	Person 16	21 Jahre
Person 7	31 Jahre	Person 17	12 Jahre
Person 8	27 Jahre	Person 18	11 Jahre
Person 9	34 Jahre	Person 19	27 Jahre
Person 10	23 Jahre	Person 20	28 Jahre

a) Berechnen Sie das um 5% getrimmte arithmetische Mittel.

b) Berechnen Sie das um 10% getrimmte arithmetische Mittel.

STATISTIK I – Lösung 01**Das arithmetische Mittel**Arithmetisches Mittel

Aus einem Produktionslos von 1.000 Karosserieteilen wird eine Stichprobe von 20 Teilen gezogen und gewogen. Es ergeben sich die folgenden Werte:

Teil 1	1,72 kg	Teil 6	1,74 kg	Teil 11	1,76 kg	Teil 16	1,76 kg
Teil 2	1,74 kg	Teil 7	1,73 kg	Teil 12	1,77 kg	Teil 17	1,77 kg
Teil 3	1,72 kg	Teil 8	1,73 kg	Teil 13	1,71 kg	Teil 18	1,71 kg
Teil 4	1,72 kg	Teil 9	1,76 kg	Teil 14	1,72 kg	Teil 19	1,75 kg
Teil 5	1,73 kg	Teil 10	1,72 kg	Teil 15	1,77 kg	Teil 20	1,76 kg

a) Fassen Sie diese Werte in einer kumulierten Häufigkeitstabelle (ohne Klassierung) zusammen.

Gewicht	Absolute Klassenhäufigkeit	Relative Klassenhäufigkeit	Absolute kum. Klassenhäufigkeit	Relative kum. Klassenhäufigkeit
1,71 kg	2	0,10	2	0,10
1,72 kg	5	0,25	7	0,35
1,73 kg	3	0,15	10	0,50
1,74 kg	2	0,10	12	0,60
1,75 kg	1	0,05	13	0,65
1,76 kg	4	0,20	17	0,85
1,77 kg	3	0,15	20	1,00
Summe	20	1,00	20	1,00

b) Berechnen Sie das arithmetische Mittel.

$$(1,71 \cdot 2) + (1,72 \cdot 5) + (1,73 \cdot 3) + (1,74 \cdot 2) + (1,75 \cdot 1) + (1,76 \cdot 4) + (1,77 \cdot 3) = 34,79$$

$$34,79 / 20 = \underline{1,7395}$$

Das arithmetische Mittel liegt bei 1,7395 kg.

Arithmetisches Mittel bei klassierten Daten

Eine Gruppe von 50 Studierenden wird nach ihrem ungefähren Lernaufwand für eine Statistiklausur (in Tagen) befragt. Es ergeben sich die folgenden (klassierten) Werte:

Größe	Absolute Klassenhäufigkeit	Relative Klassenhäufigkeit	Absolute kum. Klassenhäufigkeit	Relative kum. Klassenhäufigkeit
[1 Tag – 3 Tage)	17			
[3 Tage – 5 Tage)	23			
[5 Tage – 7 Tage)	10			
Summe				

a) Füllen Sie die restlichen Felder der kumulierten Häufigkeitstabelle aus.

Größe	Absolute Klassenhäufigkeit	Relative Klassenhäufigkeit	Absolute kum. Klassenhäufigkeit	Relative kum. Klassenhäufigkeit
[1 Tag – 3 Tage)	17	0,34	17	0,34
[3 Tage – 5 Tage)	23	0,46	40	0,80
[5 Tage – 7 Tage)	10	0,20	50	1,00
Summe	50	1,00	50	1,00

b) Berechnen Sie das arithmetische Mittel.

$$(2 * 0,34) + (4 * 0,46) + (6 * 0,20) = 0,68 + 1,84 + 1,20 = \underline{3,72}$$

Das arithmetische Mittel beträgt 3,72 Tage.

Getrimmtes arithmetisches Mittel

Eine Gruppe von Studierenden befragt Passantinnen und Passanten auf dem Campus. Erhoben wird dabei unter anderem das Alter (in Jahren). Hierfür ergeben sich für zwanzig Personen folgende Werte:

Person 1	23 Jahre	Person 11	22 Jahre
Person 2	24 Jahre	Person 12	23 Jahre
Person 3	19 Jahre	Person 13	37 Jahre
Person 4	21 Jahre	Person 14	30 Jahre
Person 5	23 Jahre	Person 15	85 Jahre
Person 6	26 Jahre	Person 16	21 Jahre
Person 7	31 Jahre	Person 17	12 Jahre
Person 8	27 Jahre	Person 18	11 Jahre
Person 9	34 Jahre	Person 19	27 Jahre
Person 10	23 Jahre	Person 20	28 Jahre

a) Berechnen Sie das um 5% getrimmte arithmetische Mittel.

$20 * 0,05 = 1$ -> Der Datensatz wird um je einen Wert an beiden Enden der Verteilung gekürzt.

$$12 + 19 + 21 + 21 + 22 + 23 + 23 + 23 + 23 + 24 + 26 + 27 + 27 + 28 + 30 + 31 + 34 + 37 = 451$$

$$451 / 18 = \underline{25,06}$$

Das um 5% getrimmte arithmetische Mittel liegt bei 25,06 Jahren.

b) Berechnen Sie das um 10% getrimmte arithmetische Mittel.

$$19 + 21 + 21 + 22 + 23 + 23 + 23 + 23 + 24 + 26 + 27 + 27 + 28 + 30 + 31 + 34 = 402$$

$$402 / 16 = \underline{25,125}$$

Das um 10% getrimmte arithmetische Mittel liegt bei 25,125 Jahren.

STATISTIK I – Übung 02

Median und Modus

1 Kurze Wiederholung

Median

Beim sogenannten Median handelt es sich ebenfalls um ein statistisches Lagemaß, welches – im Gegensatz zum arithmetischen Mittel – auch für ordinalskalierte Daten berechnet werden kann. Der Median ist als derjenige Wert definiert, der genau in der Mitte der geordneten Werte (die Ordnung von Werten setzt bekanntlich mindestens ordinalskalierte Daten voraus) einer Verteilung liegt. Da es bei einer ungeraden Anzahl von Werten tatsächlich genau einen „mittigen“ Wert gibt, während bei einer geraden Anzahl von Werten zwei Werte in der Mitte der Verteilung liegen, existieren für die Berechnung des Medians zwei unterschiedliche Formeln.

Bei einer ungeraden Anzahl von Werten wird der mittlere Wert der geordneten Verteilung gewählt:

$$x_{med} = x_{(\frac{n+1}{2})}$$

Bei einer geraden Anzahl von Werten wird das arithmetische Mittel der „mittigen“ Werte gebildet:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

mit:

$$x_{med} = \text{Median}$$

$$n = \text{Anzahl der Werte der Verteilung}$$

$$x_n = \text{Wert an } n\text{-ter Stelle der geordneten Verteilung}$$

Robustheit des Medians

Im Gegensatz zum arithmetischen Mittel ist der Median Ausreißern gegenüber äußerst robust. Dies zeigt sich am bereits bekannten Beispiel der Verteilungen [1; 2; 3; 4] und [1; 2; 3; 50]. Während das arithmetische Mittel der ersten Verteilung bei 2,5 liegt, liegt das arithmetische Mittel der zweiten Verteilung bei 14 – das arithmetische Mittel wird also durch den einzelnen, aus dem Rahmen fallenden Wert deutlich sichtbar nach oben verzerrt. Betrachtet man hingegen den Median, so ist festzustellen, dass dieser sowohl in der ersten als auch in der zweiten Verteilung bei 2,5 liegt – der Median wird durch den einzelnen Ausreißer also überhaupt nicht beeinträchtigt.

Der Grund hierfür liegt auf der Hand: Während in die Berechnung des arithmetischen Mittels sämtliche Werte der Verteilung mit exakt dem gleichen Gewicht eingehen (also auch sämtliche Ausreißer), werden für die Berechnung des Medians in diesem Fall lediglich zwei Werte (bei einer ungeraden Anzahl von Werten sogar lediglich ein Wert) benötigt, die in der Mitte der geordneten Verteilung liegen und daher unmöglich Ausreißer sein können. (Der Sonderfall einer Verteilung mit zwei Werten ist per se unsinnig und wird an dieser Stelle nicht weiter betrachtet.) Im Datensatz eventuell vorhandene Ausreißer gehen daher nicht in die Berechnung des Medians ein – und können diesen somit auch nicht beeinflussen. Das Gedankenexperiment zeigt: Auch wenn man in der zweiten Verteilung [1; 2; 3; 50] den Ausreißer auf 500, 5.000 oder 50.000 setzen würde, bliebe der Median stabil bei 2,5.

Perzentile

Im vorangegangenen Abschnitt haben wir die Definition des Medians als den Wert kennengelernt, der exakt in der Mitte der geordneten Verteilung liegt. Diese Definition kann man abgewandelt auch wie folgt formulieren: 50% der Werte einer Verteilung sind entweder kleiner oder gleich dem Median, während die anderen 50% der Werte einer Verteilung entweder größer oder gleich dem Median sind. Diese Betrachtungsweise macht deutlich, dass es sich beim Median lediglich um den Sonderfall eines anderen Lagemaßes – des Perzentils – handelt, der den Datensatz genau an der Marke 50/50 teilt. Ebenso sind aber natürlich auch andere Perzentile vorstellbar, die den Datensatz etwa an der Marke 20/80, 45/55 oder 95/5 teilen. Der Median ist insofern lediglich das bekannteste – und meistberechnete – Perzentil. Gemeinsam mit den 25%-Perzentil sowie dem 75%-Perzentil, bildet der Median die sogenannten Quartile, die einen Datensatz exakt in vier gleich große Wertebereiche unterteilen:

- 25%-Perzentil (25% aller Werte liegen unterhalb dieses Wertes, 75% liegen oberhalb)
- 50%-Perzentil – Median (50% aller Werte liegen unter- bzw. oberhalb dieses Wertes)
- 75%-Perzentil (75% aller Werte liegen unterhalb dieses Wertes, 25% liegen oberhalb)

Anders formuliert handelt es sich bei dem 25%-Perzentil um den Median der Werte unterhalb des 50%-Perzentils, während das 75%-Perzentil den Median der Werte oberhalb des 50%-Perzentils darstellt. Die Differenz zwischen 75%-Perzentil und 25%-Perzentil spielt als sogenannter Interquartilsabstand (IQR = Inter Quartile Range) eine bedeutende Rolle bei der Konstruktion von Box-Plots und stellt zudem das einzige Streuungsmaß dar, das Ausreißern gegenüber robust ist.

Für die Berechnung beliebiger Perzentile existieren – analog zur Berechnung des Median – wiederum zwei Formeln. Ergibt die Multiplikation der gewünschten Perzentilgrenze p mit der Anzahl der Werte der Verteilung n (d.h. $n * p$), keinen ganzzahligen Wert, berechnet sich das Perzentil wie folgt:

$$x_p = x_{(k)}$$

Dabei ist k die nächste auf das Ergebnis der Multiplikation ($n * p$) folgende, ganze Zahl. Ergibt ($n * p$) dagegen einen ganzzahligen Wert (in diesem Falle k), berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

Modus

Der Modus – den wir der Kürze wegen an dieser Stelle noch an die Betrachtung des Medians anhängen wollen – ist das einzige Lagemaß, das auch für nominalskalierte Daten bestimmt werden kann. Er ist als der in den unklassierten Daten am häufigsten auftretende Wert definiert, bei gleichbreit klassierten Daten entspricht der Modus dagegen der Klassenmitte der Klasse, welche die meisten Werte auf sich vereinen kann. Der Modus lässt sich ohne Berechnung unmittelbar aus den Daten herauslesen – allerdings nur dann, wenn ein eindeutiges Maximum (d.h. eine unimodale Verteilung) vorliegt. Bei bi- oder multimodalen Verteilungen kann der Modus in der Regel (es sei denn, zwei Werte treten tatsächlich exakt gleich oft auf) zwar rechnerisch bestimmt, jedoch nicht mehr vernünftig interpretiert werden. Da keine scharfen Kriterien existieren, ist es der Anwenderin bzw. dem Anwender überlassen, wann eine Verteilung „gerade noch“ als unimodal gelten kann bzw. als bimodal gelten muss.

Auf eine Besonderheit bei der Verwendung von Software wie SPSS, PAST oder PSPP sei abschließend noch hingewiesen: Liegt ein bimodaler oder multimodaler Datensatz vor, wird zumeist nur der in der Häufigkeitstabelle zuoberst stehende, am häufigsten auftretende Wert als Modus ausgegeben. Dies ist dann selbstverständlich kein rechnerisch gültiger Modus – auch wenn SPSS das anders sieht.

2 Beispielrechnungen

Median, Perzentile und Modus

Für eine Gruppe von Studierenden liegt die folgende Altersverteilung vor:

Alter	Absolute Häufigkeit	Relative Häufigkeit	Kumulierte abs. Häufigkeit	Kumulierte rel. Häufigkeit
21 Jahre	5	0,25	5	0,25
22 Jahre	4	0,20	9	0,45
23 Jahre	3	0,15	12	0,60
24 Jahre	4	0,20	16	0,80
25 Jahre	4	0,20	20	1,00
Summe	20	1,00	20	1,00

Um den Median berechnen zu können, müssen die Werte der Verteilung zunächst in eine geordnete Reihenfolge gebracht werden:

21; 21; 21; 21; 21; 22; 22; 22; 22; 23; 23; 23; 24; 24; 24; 24; 25; 25; 25; 25

Da es sich um eine gerade Anzahl an Werten handelt, steht kein einzelner Wert direkt in der Mitte der geordneten Verteilung. Für die Bestimmung des Median wird in diesem Fall daher auf die zweite Medianformel zurückgegriffen:

$$x_{med} = \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

Es ist also das arithmetische Mittel des 10 (20/2) und des 11 (20/2+1) Wertes zu berechnen:

$$(23+23) / 2 = \underline{23}$$

Der Median liegt somit bei 23 Jahren.

Zusätzlich zum Median sollen nun noch das 25%-Perzentil sowie das 75%-Perzentil berechnet werden, um die Quartile vollständig angeben und den Interquartilsabstand bestimmen zu können. Hierzu wird wie folgt vorgegangen:

$$(0,25 * 20) = 5 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 5$$

$$(0,75 * 20) = 15 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 15$$

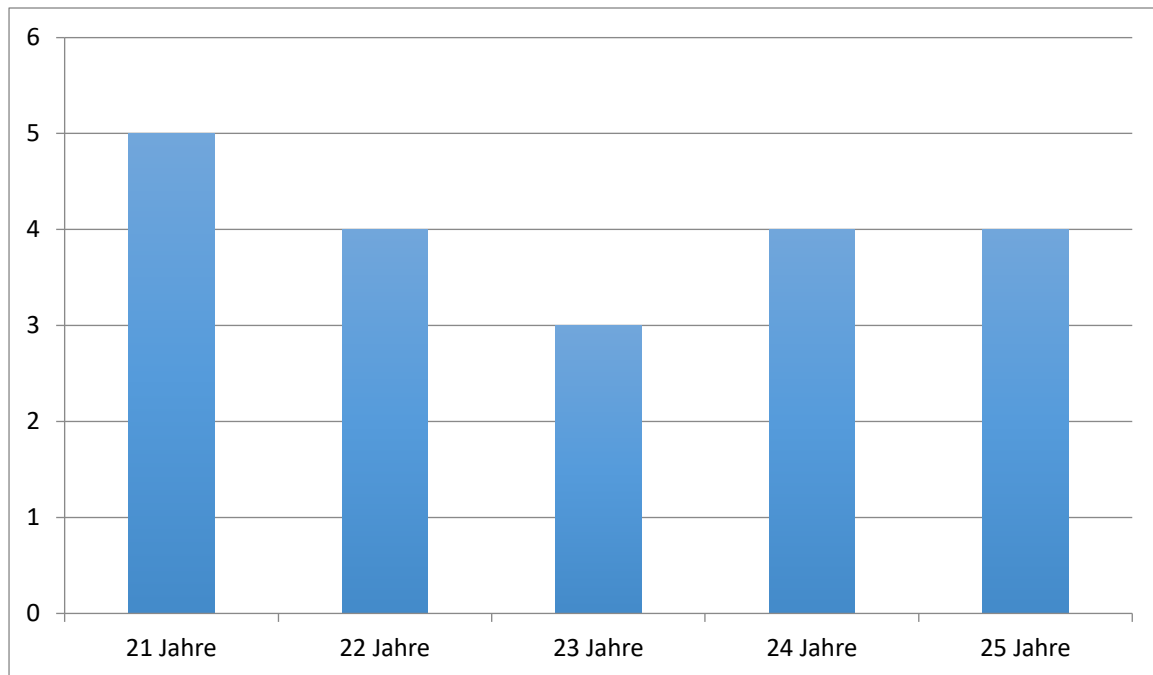
$$x_p = \frac{1}{2} (x_{(k)} + x_{(k+1)})$$

$$p_{0,25} = (x_5 + x_6) / 2 = (21+22) / 2 = \underline{21,5}$$

$$p_{0,75} = (x_{15} + x_{16}) / 2 = (24+24) / 2 = \underline{24}$$

Die drei Quartile liegen demnach bei 21,5 Jahren (unteres Quartil), 23 Jahren (Median) und 24 Jahren (oberes Quartil), der Interquartilsabstand liegt bei 2,5 Jahren (24 – 21,5).

Da kein eindeutiges Maximum existiert, ist die Bestimmung des Modus in diesem Fall nicht angezeigt. Verwendet man zum Nachrechnen dieser Übungsaufgaben eine Statistiksoftware, wird diese eventuell 21 Jahre als Modus ausgeben. Dass dies kein sinnvolles Ergebnis ist, kann man sich leicht vor Augen führen, wenn man sich das Balkendiagramm für die Verteilung anzeigen lässt.



3 Übungsaufgaben

Median und Perzentile

Im Rahmen eines Produkttests werden 20 Probandinnen und Probanden gebeten, den Geschmack eines neuartigen Joghurtprodukts auf einer Skala von 1 („hervorragend“) bis 5 („scheußlich“) zu bewerten. Der Test erbringt die folgenden Daten:

Bewertung	Anzahl an Probanden/innen
1 („hervorragend“)	5
2 („gut“)	6
3 („mittelmäßig“)	2
4 („ungenügend“)	1
5 („scheußlich“)	6

- Bestimmen Sie den Median.
- Bestimmen Sie den Interquartilsabstand.

Modus

Im Rahmen einer Qualitätsstichprobe werden 100 vom Band laufende Maschinenteile einer Genauigkeitskontrolle (Abweichung des Durchmessers von der zu erfüllenden Norm in mm) unterzogen. Die Stichprobenziehung erbringt die folgenden Daten.

Abweichung des Durchmessers von der Norm	Anzahl an Maschinenteilen in der Stichprobe
0 mm	17
[1 mm – 5 mm)	62
[5 mm – 9 mm)	12
[9 mm – 13 mm)	9

- Bestimmen Sie den Modus.

STATISTIK I – Lösung 02**Median und Modus**Median und Perzentile

Im Rahmen eines Produkttests werden 25 Probandinnen und Probanden gebeten, den Geschmack eines neuartigen Joghurtprodukts auf einer Skala von 1 („hervorragend“) bis 5 („scheußlich“) zu bewerten. Der Test erbringt die folgenden Daten:

Bewertung	Anzahl an Probanden/innen
1 („hervorragend“)	5
2 („gut“)	6
3 („mittelmäßig“)	2
4 („ungenügend“)	1
5 („scheußlich“)	6

a) Bestimmen Sie den Median.

Die geordnete Verteilung sieht wie folgt aus:

1; 1; 1; 1; 1; 2; 2; 2; 2; 2; 2; 2; 3; 3; 4; 5; 5; 5; 5; 5; 5

Bei einer geraden Anzahl von Werten wird das arithmetische Mittel der „mittigen“ Werte gebildet:

$$x_{med} = \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

$$(x_{10} + x_{11}) / 2 = (2 + 2) / 2 = \underline{2}$$

Der Median dieser Verteilung liegt bei 2.

b) Bestimmen Sie den Interquartilsabstand.

Zur Bestimmung des IQR werden das 25%-Perzentil sowie das 75%-Perzentil benötigt.

$$(0,25 * 20) = 5 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 5$$

$$(0,75 * 20) = 15 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 15$$

$$x_p = \frac{1}{2} (x_{(k)} + x_{(k+1)})$$

$$p_{0,25} = (x_5 + x_6) / 2 = (1 + 2) / 2 = \underline{1,5}$$

$$p_{0,75} = (x_{15} + x_{16}) / 2 = (5 + 5) / 2 = \underline{5}$$

Der Interquartilsabstand dieser Verteilung beträgt 3,5 (5 – 1,5).

Modus

Im Rahmen einer Qualitätsstichprobe werden 100 vom Band laufende Maschinenteile einer Genauigkeitskontrolle (Abweichung des Durchmessers von der zu erfüllenden Norm in mm) unterzogen. Die Stichprobenziehung erbringt die folgenden Daten.

Abweichung des Durchmessers von der Norm	Anzahl an Maschinenteilen in der Stichprobe
0 mm	17
[1 mm – 5 mm)	62
[5 mm – 9 mm)	12
[9 mm – 13 mm)	9

a) Bestimmen Sie den Modus.

Dass der Modus in der Klasse [1 mm – 5 mm) liegt, scheint evident zu sein. Allerdings gilt es in diesem Fall zu beachten, dass die obere Klasse nicht die gleiche Breite wie die drei unteren Klassen aufweist. Dieser Sonderfall ist im Skript nicht erfasst, kann aber leicht in der entsprechenden Fachliteratur sowie im Netz recherchiert werden. Zu bestimmen ist in dieser Situation die Klassenhöhe:

Abweichung	Abs. Häufigkeit	Klassenbreite	Klassenhöhe
0 mm	17	1	$17/1 = 17$
[1 mm – 5 mm)	62	4	$62/4 = 15,5$
[5 mm – 9 mm)	12	4	$12/4 = 3$
[9 mm – 13 mm)	9	4	$9/4 = 2,25$

Geht man von einer Gleichverteilung der Werte innerhalb der Klasse aus (was man, da keine genaueren Daten vorliegen, tun muss), ist tatsächlich 0 mm und nicht [1 mm – 5 mm) als Modus zu benennen. Zwar verfügt die Klasse [1 mm – 5 mm) über deutlich mehr Werte, ist aber auch erheblich breiter, so dass sich die 62 Werte entsprechend breit verteilen (auf 15,5 Werte pro diskretem Wert – betrachtet man die Daten sinnvollerweise als stetig, ist die Verteilung entsprechend breiter), während die 17 Werte in der oberen Klasse allein dem (diskreten) Wert 0 mm zugeordnet werden.

STATISTIK I – Übung 03

Statistische Lagemaße

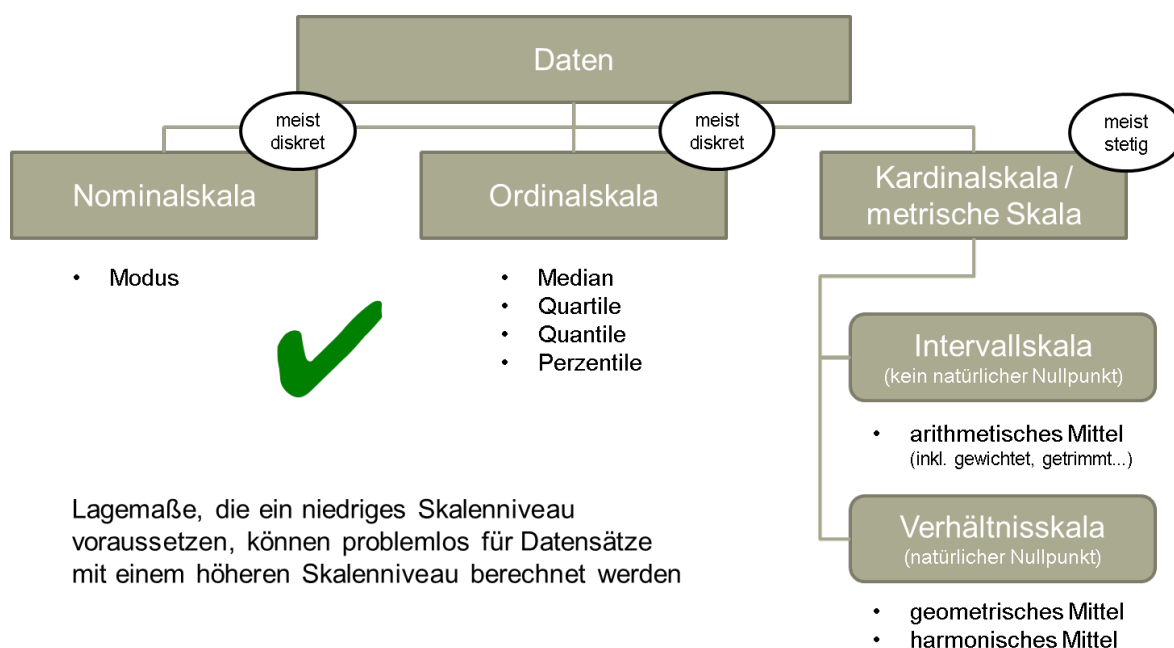
1 Kurze Wiederholung

Was sind statistische Lagemaße?

Die statistischen Lagemaße werden auch als Mittelwerte oder Maße der zentralen Tendenz bezeichnet. Sie geben Auskunft über das Zentrum einer Verteilung und sind am besten geeignet, wenn es gilt, eine Verteilung mit nur einem Parameter zusammenzufassen – wie etwa die Einkommensverteilung mit der Angabe des Durchschnittseinkommens. (Warum nur ein Parameter für die Darstellung einer Verteilung in der Regel eben nicht ausreicht, wird im nächsten Übungsabschnitt zum Thema Streuungsmaße erläutert.) Im Rahmen der Vorlesung haben wir uns insbesondere mit drei Lagemaßen – dem arithmetischen Mittel (Durchschnittswert der Verteilung), dem Median (Wert genau in der Mitte der geordneten Verteilung) und dem Modus (in der Verteilung am häufigsten auftretender Wert) – beschäftigt. Die beiden kurz betrachteten Sonderfälle des geometrischen Mittels (für relative Veränderungen) sowie des harmonischen Mittels (für Quotienten) sind nicht klausurrelevant, dafür allerdings die Berechnung beliebiger Perzentile als Verallgemeinerung des Modus. In dieser Übungseinheit wollen wir uns noch einmal allen Lagemaßen gemeinsam widmen, bevor es in der nächsten Übungseinheit dann um zwei Streuungsmaße (Interquartilsabstand und Spannweite) gehen soll.

Lagemaße und Skalenniveaus

Wie wir ebenfalls gelernt haben, entscheidet sich die Frage, welches Lagemaß für eine beliebige Verteilung berechnet werden kann, am Skalenniveau der Daten. Dabei ist zu beachten, dass die Lagemaße zwar „aufwärtskompatibel“, nicht aber „abwärtskompatibel“ sind. Liegen metrisch skalierte Daten vor, kann neben dem arithmetischen Mittel also auch der Median, oder (falls die Verteilung ein eindeutiges Maximum aufweist) der Modus berechnet werden – liegen dagegen ordinalskalierte Daten vor, ist die Berechnung des arithmetischen Mittels definitiv nicht möglich. Statistische Lagemaße, die ein niedrigeres Skalenniveau voraussetzen, können also auch auf Daten eines höheren Skalenniveaus angewandt werden – nicht jedoch umgekehrt. Die nachfolgende Grafik verdeutlicht noch einmal, welches statistische Lagemaß ab welchem Skalenniveau zum Einsatz kommen kann.



2 Beispielrechnungen

Im Rahmen einer Erhebung unter 100 Mieterinnen und Mietern wurde unter anderem nach der Größe des Badezimmers (gerundet auf ganze Zahlen) gefragt. Es ergaben sich folgende Werte:

Badgröße	Probanden/innen	Badgröße	Probanden/innen
7 m ²	13	12 m ²	7
8 m ²	8	13 m ²	23
9 m ²	12	14 m ²	5
10 m ²	6	15 m ²	6
11 m ²	12	16 m ²	8

Nachfolgend sollen das arithmetische Mittel, das um 5% getrimmte arithmetische Mittel, die Quartile (oberes/unteres Quartil und Median) sowie der Modus berechnet werden.

Berechnung des arithmetischen Mittels

$$13 + 8 + 12 + 6 + 12 + 7 + 23 + 5 + 6 + 8 = 100$$

$$(7 * 13) + (8 * 8) + (9 * 12) + (10 * 6) + (11 * 12) + (12 * 7) + (13 * 23) + (14 * 5) + (15 * 6) + (16 * 8) \\ = 91 + 64 + 108 + 60 + 132 + 84 + 299 + 70 + 90 + 128 = 1126$$

$$1126 / 100 = \underline{11,26}$$

Das arithmetische Mittel liegt bei 11,26 m².

Berechnung des um 5% getrimmten arithmetischen Mittels

Bei Trimmung eines Datensatzes der Größe $n = 100$ um 5% fallen jeweils 5 Werte ($100 * 0,05$) am oberen und unteren Ende der Verteilung weg. Im vorliegenden Fall sind also 5 x 7 m² sowie 5 x 16 m² zu streichen. Anschließend wird das arithmetische Mittel neu berechnet.

$$8 + 8 + 12 + 6 + 12 + 7 + 23 + 5 + 6 + 3 = 90$$

$$(7 * 8) + (8 * 8) + (9 * 12) + (10 * 6) + (11 * 12) + (12 * 7) + (13 * 23) + (14 * 5) + (15 * 6) + (16 * 3) \\ = 56 + 64 + 108 + 60 + 132 + 84 + 299 + 70 + 90 + 48 = 1011$$

$$1011 / 90 = \underline{11,23}$$

Das um 5% getrimmte arithmetische Mittel liegt also bei 11,23 m². Die vergleichsweise geringe Distanz zum ungetrimmten Mittel lässt vermuten, dass sich im Datensatz keine Ausreißer befinden (die sich in der Tabelle auch objektiv nicht identifizieren lassen).

Berechnung der Quartile

Für die Berechnung des Medians bzw. der Quartile empfiehlt sich die Anlage der Tabelle der kumulierten Häufigkeitsverteilung, da sich auf diesem Wege die Bildung einer geordneten Reihe mit sämtlichen 100 Werten vermeiden lässt und man schneller zum gewünschten Ergebnis kommt. Der Median ist in dieser Tabelle aus der Zeile abzulesen, in welcher der Wert der kumulierten Häufigkeitsverteilung die 0,50 überschreitet. Ebenso sind das obere und das untere Quartil aus den Zeilen abzulesen, in denen der Wert der kumulierten Häufigkeitsverteilung die 0,25 bzw. die 0,75 überschreitet.

Badgröße	Abs. Häufigkeit	Rel. Häufigkeit	Kumulierte abs. Häufigkeit	Kumulierte rel. Häufigkeit
7 m ²	13	0,13	13	0,13
8 m ²	8	0,08	21	0,21
9 m ²	12	0,12	33	0,33
10 m ²	6	0,06	39	0,39
11 m ²	12	0,12	51	0,51
12 m ²	7	0,07	58	0,58
13 m ²	23	0,23	81	0,81
14 m ²	5	0,05	86	0,86
15 m ²	6	0,06	92	0,92
16 m ²	8	0,08	100	1,00
Summe	100	1,00	100	1,00

Aus der Tabelle lassen sich die drei Quartile nun wie folgt direkt ablesen:

Unteres Quartil: 9 m²

Mittleres Quartil: 11 m²

Oberes Quartil: 13 m²

Das mittlere Quartil ist mit dem Median identisch.

Berechnung des Modus

Der Modus ist als der Wert definiert, der in der Verteilung am häufigsten auftritt: 13 m². Da es sich um eine unimodale Verteilung handelt (der Wert 13 m² tritt 23 x auf, der nächsthäufigere Wert lediglich 12 x), kann der Modus in diesem Fall sinnvoll gebildet werden.

3 Übungsaufgaben

Die 100 Personen aus der vorangegangenen Aufgabe wurden auch zur Größe (ebenfalls gerundet) anderer Räume in ihren Wohnungen befragt. Für das Schlafzimmer ergaben sich folgende Werte:

Schlafzimmergröße	Probanden/innen	Schlafzimmergröße	Probanden/innen
12 m ²	11	17 m ²	22
13 m ²	21	18 m ²	6
14 m ²	12	19 m ²	5
15 m ²	7	20 m ²	3
16 m ²	8	21 m ²	5

- Bestimmen Sie das arithmetische Mittel.
- Bestimmen Sie das um 10% getrimmte arithmetische Mittel.
- Bestimmen Sie die drei Quartile.
- Bestimmen Sie den Modus.

STATISTIK I – Lösung 03**Statistische Lagemaße**

Die 100 Personen aus der vorangegangenen Aufgabe wurden auch zur Größe (ebenfalls gerundet) anderer Räume in ihren Wohnungen befragt. Für das Schlafzimmer ergaben sich folgende Werte:

Schlafzimmergröße	Probanden/innen	Schlafzimmergröße	Probanden/innen
12 m ²	11	17 m ²	22
13 m ²	21	18 m ²	6
14 m ²	12	19 m ²	5
15 m ²	7	20 m ²	3
16 m ²	8	21 m ²	5

a) Bestimmen Sie das arithmetische Mittel.

$$11 + 21 + 12 + 7 + 8 + 22 + 6 + 5 + 3 + 5 = 100$$

$$(12 * 11) + (13 * 21) + (14 * 12) + (15 * 7) + (16 * 8) + (17 * 22) + (18 * 6) + (19 * 5) + (20 * 3) + (21 * 5) = 132 + 273 + 168 + 105 + 128 + 374 + 108 + 95 + 60 + 105 = 1548$$

$$1548 / 100 = \underline{15,48}$$

Das arithmetische Mittel beträgt 15,48 m².

b) Bestimmen Sie das um 10% getrimmte arithmetische Mittel.

Eine Trimmung um 10% bedeutet den Wegfall der 10 größten sowie der 10 kleinsten Werte.

$$1 + 21 + 12 + 7 + 8 + 22 + 6 + 3 = 80$$

$$(12 * 1) + (13 * 21) + (14 * 12) + (15 * 7) + (16 * 8) + (17 * 22) + (18 * 6) + (19 * 3) = 12 + 273 + 168 + 105 + 128 + 374 + 108 + 57 = 1225$$

$$1225 / 80 = \underline{15,31}$$

Das um 10% getrimmte arithmetische Mittel beträgt 15,31 m².

c) Bestimmen Sie die drei Quartile.

Zimmergröße	Abs. Häufigkeit	Rel. Häufigkeit	Kumulierte abs. Häufigkeit	Kumulierte rel. Häufigkeit
12 m ²	11	0,11	11	0,11
13 m ²	21	0,21	32	0,32
14 m ²	12	0,12	44	0,44
15 m ²	7	0,07	51	0,51
16 m ²	8	0,08	59	0,59
17 m ²	22	0,22	81	0,81
18 m ²	6	0,06	87	0,87
19 m ²	5	0,05	92	0,92
20 m ²	3	0,03	95	0,95
21 m ²	5	0,05	100	1,00
Summe	100	1,00	100	1,00

Unteres Quartil = 13 m²

Mittleres Quartil = 15 m²

Oberes Quartil = 17 m²

d) Bestimmen Sie den Modus.

Da es sich um eine bimodale Verteilung handelt, ist der Modus nicht eindeutig bestimmbar.

STATISTIK I – Übung 04

Spannweite und IQR

1 Kurze Wiederholung

Was sind Dispersionsparameter?

Die sogenannten Dispersionsparameter oder statistischen Streuungsmaße geben Auskunft darüber, wie die Werte einer Verteilung um deren Zentrum (gekennzeichnet durch ein statistisches Lagemaß) streuen, d.h. ob sie tendenziell eher dicht am Zentrum oder eher weit davon entfernt liegen. Warum die Kenntnis der Streuung unter anderem für den Vergleich zweier Verteilungen von großer Bedeutung ist, soll das folgende fiktive Beispiel der Einkommensverteilung in zwei Abteilungen eines Unternehmens mit je sechs Mitarbeiterinnen und Mitarbeitern verdeutlichen.

Mitarbeiter Abt. A	Einkommen	Mitarbeiter Abt. B	Einkommen
MA 1	2.500,00 Euro	MA 1	4.130,00 Euro
MA 2	2.550,00 Euro	MA 2	1.060,00 Euro
MA 3	2.480,00 Euro	MA 3	1.110,00 Euro
MA 4	2.630,00 Euro	MA 4	5.020,00 Euro
MA 5	3.000,00 Euro	MA 5	4.000,00 Euro
MA 6	2.210,00 Euro	MA 6	1.250,00 Euro
Summe	15.370,00 Euro	Summe	16.570,00 Euro
Arithmetisches Mittel	2.561,67 Euro	Arithmetisches Mittel	2.761,67 Euro

Wie man sieht, liegen die arithmetischen Mittel beider Verteilungen mit 2.561 Euro in Abteilung A und 2.761 Euro in Abteilung B gerade einmal 200 Euro voneinander entfernt – dennoch wäre es mit Blick auf die Daten falsch, würde man aus dieser geringen Differenz den Schluss ziehen, dass die Mitarbeiterinnen und Mitarbeiter in beiden Abteilungen etwa gleich viel verdienen. Würde man in Abteilung A anheuern, könnte man in der Tat davon ausgehen, um 2.561 Euro zu verdienen. In Abteilung B liegt dagegen kein Mitarbeiter mit seinem Einkommen auch nur in der Nähe des Durchschnittswertes – hier verdient man entweder deutlich mehr oder deutlich weniger als 2.761 Euro.

Das Beispiel zeigt, dass es bei der Betrachtung von Verteilungen eben nicht nur darauf ankommt, das Zentrum der Verteilung zu kennen – vielmehr muss man auch eine Vorstellung davon haben, wie sich die Werte um dieses Zentrum herum verteilen. Auskunft hierüber geben die Dispersionsparameter, von denen wir im Rahmen der Vorlesung vier – Spannweite, Interquartilsabstand, Varianz und Standardabweichung – kennenlernen werden. In dieser Übungseinheit werden wir – neben der Theorie – zunächst die „einfacheren“ Parameter Spannweite und Interquartilsabstand abarbeiten, während wir uns in der kommenden Übungseinheit der Varianz sowie der Standardabweichung zuwenden.

Dispersionsparameter und Skalenniveaus

Ähnlich wie schon bei den statistischen Lagemaßen, hängt auch die Wahl geeigneter Dispersionsparameter vom Skalenniveau der betrachteten Verteilung ab, wobei in diesem Fall nur zwei Skalenniveaus – Ordinalskala und Kardinalskala – von Bedeutung sind. Da nominalskalierte Werte nicht quantifiziert werden können, können sie auch nicht um das Zentrum einer Verteilung (in diesem Fall also um den Modus) streuen, da das Vorhandensein einer Streuung natürlich voraussetzt, dass sich die Abstände zwischen den Werten einer Verteilung und ihrem Zentrum ermitteln lassen. Da dies bei nominalskalierten Werten nicht möglich ist, existiert kein Maß für deren Streuung. Man kann sich dies leicht vor Augen führen, indem man mit dem Geschlecht ein typisches nominalskaliertes Merkmal betrachtet. Befragt man 40 Personen, von denen 30 männlich sind, ist „männlich“ zwar der Modus – aber wie sollte in diesem Fall eine „Streuung“ der Werte um dieses Zentrum aussehen?

Die Berechnung des Interquartilsabstands verlangt mindestens ordinalskalierte Werte, da – wie wir nachfolgend noch sehen werden – das obere und untere Quartil in diesen Parameter einfließen, deren Berechnung bekanntlich ebenfalls ordinalskalierte Werte voraussetzt. Da in die Berechnung der Varianz sowie der sich aus dieser ergebenden Standardabweichung das arithmetische Mittel einfließt, können diese beiden Parameter analog nur für metrisch skalierte Werte berechnet werden. Auch die Berechnung der Spannweite ist – aus Gründen, die wir nachfolgend noch betrachten werden – nur für metrisch skalierte Werte sinnvoll. Wie bei den statistischen Lagemaßen gilt auch für die Dispersionsparameter, dass sie abwärts-, aber nicht aufwärtskompatibel sind, d.h. der Interquartilsabstand lässt sich auch für metrisch skalierte Werte berechnen, während die Berechnung von Varianz, Standardabweichung und Spannweite für ordinalskalierte Werte nicht möglich ist.

Spannweite

Bei der Spannweite handelt es sich um die Differenz zwischen dem kleinsten und dem größten Wert im Datensatz. Da die Bildung von sinnvollen Differenzen voraussetzt, dass diese quantifizierbar sind, d.h. dass mit den Abständen zwischen den Werten gerechnet werden kann, kann die Spannweite nur für metrische Daten berechnet werden – auch wenn es vielleicht kontraintuitiv erscheint, dass etwa eine Spannweite von 5 Notenpunkten bei einer Klassenarbeit keine sinnvolle Größe sein soll.

$$d_s = x_{\max} - x_{\min}$$

Intuitiv dürfte dagegen die Erkenntnis sein, dass die Spannweite kein robustes Maß für die Streuung ist, da sie von Ausreißern – soweit im Datensatz vorhanden – extrem beeinflusst wird. Tritt an einem Ende der Verteilung mindestens ein Ausreißer auf, findet dieser in jedem Fall Eingang in die Berechnung der Spannweite – gibt es an beiden Enden der Verteilung mindestens einen Ausreißer, wird die Spannweite sogar ausschließlich durch diese bestimmt. Das nachfolgende Beispiel verdeutlicht, wie extrem die Spannweite durch Ausreißer beeinflusst werden kann.

Verteilung A: [100; 120; 170; 280; 290; 300]

Spannweite: $d_s = 300 - 100 = 200$

Verteilung B: [5; 120; 170; 280; 290; 10000]

Spannweite $d_s = 10000 - 5 = 9995$

Aber auch wenn keinerlei Ausreißer im Datensatz vorhanden sind, gibt die Spannweite aufgrund der Tatsache, dass lediglich zwei Werte in ihre Berechnung einfließen, nur ein eher unvollständiges Bild der Streuung wieder, wie anhand eines zweiten Beispiels demonstriert werden soll.

Verteilung A: [100; 100; 110; 115; 118; 300]

Spannweite: $d_s = 300 - 100 = 200$

Verteilung B: [100; 140; 185; 240; 285; 300]

Spannweite: $d_s = 300 - 100 = 200$

Liegen metrisch skalierte Werte vor, ist die Berechnung von Varianz und Standardabweichung also in jedem Fall zu bevorzugen. In der Praxis werden der größte und der kleinste Wert allerdings noch für die sogenannte Fünf-Werte-Zusammenfassung benötigt, auf die später eingegangen werden soll.

Interquartilsabstand (IQR)

Der Interquartilsabstand (nachfolgend als IQR – Inter Quartile Range – abgekürzt) ist als die Differenz zwischen dem oberen und dem unteren Quartil definiert, die wir bereits im Rahmen der statistischen Lagemaße kennengelernt haben.

$$IQR = x_{0,75} - x_{0,25}$$

Damit der IQR bestimmt werden kann, muss also die Berechnung der Quartile möglich sein, was wiederum – wie eingangs bereits angesprochen – mindestens ordinalskalierte Werte voraussetzt. Der IQR kann aber auch für metrisch skalierte Werte kalkuliert werden (Abwärtskompatibilität).

Im Gegensatz zur Spannweite sowie auch zur Varianz und Standardabweichung, die wir im Rahmen der nächsten Übungseinheit besprechen werden, wird der IQR durch Ausreißer nicht beeinflusst und ist somit der einzige in dieser Vorlesung besprochene robuste Dispersionsparameter. Dies lässt sich anhand des weiter oben genutzten Beispiels zur Robustheit der Spannweite demonstrieren:

Verteilung A: [100; 120; 170; 280; 290; 300]

$$\text{IQR} = 290 - 120 = 170$$

Verteilung B: [5; 120; 170; 280; 290; 10000]

$$\text{IQR} = 290 - 120 = 170$$

Der IQR spielt eine zentrale Rolle bei der Konstruktion von Box-Plots (Höhe der Box) – einer der zwei grafischen Darstellungsformen (zusammen mit dem Stem-and-Leaf-Plot) mit Relevanz für die Klausur. Wie ein Box-Plot konstruiert wird, wird Gegenstand einer zukünftigen Übungseinheit sein.

Fünf-Werte-Zusammenfassung

Bei der sogenannten Fünf-Werte-Zusammenfassung handelt es sich um eine hochkomprimierte Darstellung der Lage sowie der Streuung einer Verteilung, bestehend aus den drei Quartilen (oberes und unteres Quartil sowie Median) sowie dem größten und dem kleinsten Wert der Verteilung.

$$[X_{\min}; X_{0,25}; X_{\text{med}}; X_{0,75}; X_{\max}]$$

2 Beispielrechnungen

Im Rahmen einer Befragung machten 30 Probanden Angaben zu ihrem Körpergewicht.

Nr.	Körpergewicht	Nr.	Körpergewicht
1	61,2 kg	16	62,2 kg
2	72,8 kg	17	79,5 kg
3	62,3 kg	18	61,0 kg
4	85,7 kg	19	55,5 kg
5	91,8 kg	20	98,9 kg
6	64,9 kg	21	94,4 kg
7	74,3 kg	22	85,3 kg
8	95,2 kg	23	93,0 kg
9	87,3 kg	24	72,8 kg
10	82,4 kg	25	88,2 kg
11	78,4 kg	26	87,4 kg
12	91,2 kg	27	68,6 kg
13	89,3 kg	28	90,5 kg
14	76,3 kg	29	85,0 kg
15	68,2 kg	30	71,9 kg

Berechnung der Spannweite

Die Spannweite berechnet sich als Differenz zwischen dem größten und dem kleinsten Wert.

$$d_s = x_{\max} - x_{\min} = 98,9 - 55,5 = \underline{43,4}$$

Die Spannweite beträgt 43,4 kg.

Berechnung des Interquartilsabstands

Für die IQR-Berechnung müssen die Werte zunächst in eine geordnete Reihe gebracht werden.

Werte 1-10: 55,5; 61,0; 61,2; 62,2; 62,3; 64,9; 68,2; 68,6; 71,9; 72,8

Werte 11-20: 72,8; 74,3; 76,3; 78,4; 79,5; 82,4; 85,0; 85,3; 85,7; 87,3

Werte 21-30: 87,4; 88,2; 89,3; 90,5; 91,2; 91,8; 93,0; 94,4; 95,2; 98,9

Für die Berechnung beliebiger Perzentile (die Quartile sind bekanntlich drei Perzentilwerte) existieren – analog zur Berechnung des Median – zwei bereits in einer der Übungseinheiten zu den statistischen Lagemaßen vorgestellte Formeln. Ergibt die Multiplikation der gewünschten Perzentilgrenze p mit der Anzahl der Werte der Verteilung n (d.h. $n * p$), keinen ganzzahligen Wert, berechnet sich das gesuchte Perzentil wie folgt:

$$x_p = x_{(k)}$$

Dabei ist k die nächste auf das Ergebnis der Multiplikation ($n * p$) folgende, ganze Zahl. Ergibt ($n * p$) dagegen einen ganzzahligen Wert (in diesem Falle k), berechnet sich das gesuchte Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

In diesem Fall werden das obere ($p = 0,75$) und das untere ($p = 0,25$) Quartil benötigt, d.h.:

$(30 * 0,25) = 7,5 \rightarrow$ kein ganzzahliger Wert $\rightarrow k = 8 \rightarrow$ Der 8. Wert im Datensatz lautet 68,6

$(30 * 0,75) = 22,5 \rightarrow$ kein ganzzahliger Wert $\rightarrow k = 23 \rightarrow$ Der 23. Wert im Datensatz lautet 89,3

$$IQR = 89,3 - 68,6 = \underline{20,7}$$

Der Interquartilsabstand beträgt 20,7 kg.

Angabe der Fünf-Werte-Zusammenfassung

Vier der fünf für die Fünf-Werte-Zusammenfassung benötigten Werte (größter Wert, kleinster Wert, oberes Quartil und unteres Quartil) sind aus den vorangegangenen Rechnungen bereits bekannt – es wird demnach nur noch der Median gesucht. Da eine gerade Anzahl von Werten ($n = 30$) vorliegt, wird für die Berechnung des Medians auf folgende Formel zurückgegriffen:

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$$

$$x_{med} = \frac{1}{2}(x_{(\frac{30}{2})} + x_{(\frac{30}{2}+1)})$$

$$x_{med} = \frac{1}{2}(x_{15} + x_{16})$$

$$x_{med} = \frac{1}{2}(79,5 + 82,4)$$

$$x_{med} = 80,95$$

Wie bei jeder Perzentilberechnung kann natürlich auch analog zu oben vorgegangen werden:

$$(30 * 0,50) = 15 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 15; k+1 = 16 \rightarrow \frac{1}{2} * (79,5 + 82,4) = 80,95$$

Die Fünf-Werte-Zusammenfassung lautet somit:

[55,50 kg; 68,60 kg; 80,95 kg; 89,30 kg; 98,90 kg]

3 Übungsaufgaben

Die gleichen 30 Probanden machten außerdem Angaben zu ihrem Alter.

Nr.	Alter	Nr.	Alter
1	17 Jahre	16	33 Jahre
2	36 Jahre	17	22 Jahre
3	41 Jahre	18	62 Jahre
4	34 Jahre	19	44 Jahre
5	32 Jahre	20	41 Jahre
6	54 Jahre	21	56 Jahre
7	45 Jahre	22	62 Jahre
8	22 Jahre	23	23 Jahre
9	71 Jahre	24	86 Jahre
10	14 Jahre	25	41 Jahre
11	86 Jahre	26	65 Jahre
12	44 Jahre	27	53 Jahre
13	34 Jahre	28	35 Jahre
14	21 Jahre	29	21 Jahre
15	54 Jahre	30	19 Jahre

- a) Bestimmen Sie die Spannweite.
- b) Bestimmen Sie den Interquartilsabstand.
- c) Bestimmen Sie die Fünf-Werte-Zusammenfassung.

STATISTIK I – Lösung 04

Spannweite und IQR

In einer Erhebung wurden 30 Probanden nach ihrem Alter befragt.

Nr.	Alter	Nr.	Alter
1	17 Jahre	16	33 Jahre
2	36 Jahre	17	22 Jahre
3	41 Jahre	18	62 Jahre
4	34 Jahre	19	44 Jahre
5	32 Jahre	20	41 Jahre
6	54 Jahre	21	56 Jahre
7	45 Jahre	22	62 Jahre
8	22 Jahre	23	23 Jahre
9	71 Jahre	24	86 Jahre
10	14 Jahre	25	41 Jahre
11	86 Jahre	26	65 Jahre
12	44 Jahre	27	53 Jahre
13	34 Jahre	28	35 Jahre
14	21 Jahre	29	21 Jahre
15	54 Jahre	30	19 Jahre

a) Bestimmen Sie die Spannweite.

Die Spannweite ist definiert als Differenz zwischen dem größten und dem kleinsten Wert.

$$d_s = x_{\max} - x_{\min} = 86 - 14 = \underline{72}$$

Die Spannweite beträgt 72 Jahre.

b) Bestimmen Sie den Interquartilsabstand.

Zur Bestimmung des IQR sind das obere sowie das untere Quartil zu berechnen. Hierfür müssen die Werte zunächst in eine geordnete Reihenfolge gebracht werden:

Wert 1-10: 14; 17; 19; 21; 21; 22; 22; 23; 32; 33

Wert 11-20: 34; 34; 35; 36; 41; 41; 41; 44; 44; 45

Wert 21-30: 53; 54; 54; 56; 62; 62; 65; 71; 86; 86

$(30 \cdot 0,25) = 7,5 \rightarrow$ kein ganzzahliger Wert $\rightarrow k = 8 \rightarrow$ Der 8. Wert im Datensatz lautet 23

$(30 \cdot 0,75) = 22,5 \rightarrow$ kein ganzzahliger Wert $\rightarrow k = 23 \rightarrow$ Der 23. Wert im Datensatz lautet 54

$$IQR = 54 - 23 = \underline{31}$$

Der Interquartilsabstand beträgt 31 Jahre.

c) Bestimmen Sie die Fünf-Werte-Zusammenfassung.

Die Fünf-Werte-Zusammenfassung umfasst die drei Quartile sowie den größten und den kleinsten Wert. Lediglich der Median als mittleres Quartil muss an dieser Stelle noch berechnet werden.

$$(30 \cdot 0,50) = 15 \rightarrow \text{ganzzahliger Wert} \rightarrow k = 15; k+1 = 16 \rightarrow \frac{1}{2} \cdot (41 + 41) = 41$$

Die Fünf-Werte-Zusammenfassung lautet demnach:

[14 Jahre; 23 Jahre; 41 Jahre; 54 Jahre; 86 Jahre]

STATISTIK I – Übung 05**Varianz und Standardabweichung****1 Kurze Wiederholung**Varianz und Standardabweichung

Bei der Varianz handelt es sich um den wichtigsten und meistgenutzten Dispersionsparameter. Sie berechnet sich als die Summe der quadrierten Abweichungen der Einzelwerte einer Verteilung vom arithmetischen Mittel eben dieser Verteilung, abschließend geteilt durch die Gesamtzahl der Werte.

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

An dieser Stelle lohnt es sich, die Formel nicht nur stur anzuwenden, sondern näher zu betrachten – was wird hier eigentlich gerechnet? Zunächst einmal wird die Differenz der einzelnen Werte der Verteilung ($x_1, x_2, x_3 \dots x_n$) vom arithmetischen Mittel gebildet, die Summe aller Differenzen wird anschließend durch die Anzahl der Werte der Verteilung geteilt. Man errechnet somit quasi das arithmetische Mittel der Differenzen aller Werte vom arithmetischen Mittel – oder anders formuliert die durchschnittliche Abweichung der Werte vom Zentrum der Verteilung. Diese durchaus naheliegende und intuitive Vorgehensweise für die Bestimmung der Streuung überrascht nur in einem Detail: Die zu addierenden Differenzen werden vorab noch quadriert um zu verhindern, dass sich die positiven und die negativen Abweichungen gegenseitig neutralisieren. In der Formelsammlung findet sich neben der bekannten Formel noch diese leicht vereinfachte Berechnungsvorschrift.

$$s^2 = \left\{ \frac{1}{n} \sum_{i=1}^n x_i^2 \right\} - \bar{x}^2$$

Aufgrund des unumgänglichen Quadrierens ist allerdings auch das Ergebnis eine quadrierte Zahl, d.h. die Varianz einer Gehaltsverteilung steht nicht in der Einheit Euro, sondern in Euro². Dadurch lässt sie sich schlecht interpretieren – oder aber es besteht sogar die Gefahr einer inhaltlichen Fehlinterpretation, wenn die quadrierte Einheit „sinnvoll“ zu sein scheint (Quadratmeter). Aus diesem Grund wird bei der Berechnung der Varianz in der Regel auch noch die Standardabweichung als positive Wurzel aus der Varianz und damit „korrekt dimensionierter“ Dispersionsparameter angegeben.

$$s = +\sqrt{s^2}$$

Variationskoeffizient

Wie wir in der letzten Übungseinheit zu den Dispersionsparametern gelernt haben, spielen die Streuungsmaße unter anderem beim Vergleich von Verteilungen eine wichtige Rolle. Hierbei kann es erforderlich werden, die Standardabweichung in eine dimensionslose Größe umzuwandeln, um einen Vergleich der Streuung von Verteilungen zu ermöglichen, die in verschiedenen Dimensionen erfasst werden. Als Beispiel sei auf den Vergleich von Preisverteilungen in unterschiedlichen Währungen – etwa in Euro und mexikanische Pesos – verwiesen. Da ein Euro etwa 20 Pesos entspricht, besteht bei einem direkten Vergleich der Standardabweichungen beider Verteilungen das Risiko, die Streuung der Preisangaben in Pesos im Verhältnis zu den Preisangaben in Euro aufgrund der größeren absoluten Differenzwerte erheblich zu überschätzen. Aus diesem Grund dividiert man die Standardabweichungen beider Verteilungen vor dem Vergleich noch durch die jeweiligen arithmetischen Mittel und generiert somit den dimensionslosen Variationskoeffizienten, mit dem ein Vergleich möglich ist. Die Berechnung des Variationskoeffizienten ist nur möglich, wenn das arithmetische Mittel positiv ist.

$$v = \frac{s}{x}$$

Empirische Varianz und Stichprobenvarianz

Im Rahmen dieser Vorlesung arbeiten wir mit der sogenannten empirischen Varianz:

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

Bei der Berechnung von Varianzen mit SPSS, PSPP und anderen statistischen Analyseprogrammen wird dagegen gelegentlich die Stichprobenvarianz berechnet – allerdings nicht als solche, sondern als Varianz deklariert, was zu erheblicher Verwirrung bei Studierenden führen kann, die ihre Lösungen für Übungsaufgaben mit Hilfe von Software überprüfen möchten.

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

Die Stichprobenvarianz unterscheidet sich von der empirischen Varianz darin, dass anstatt durch die Anzahl der Werte der Verteilung durch die Anzahl an Freiheitsgraden (n-1) dividiert wird – ein Begriff, auf den in späteren Übungseinheiten noch einmal näher eingegangen werden wird.

2 Beispielrechnungen

Für die Beispielrechnung greifen wir auf die Daten aus der letzten Übungseinheit zurück – Angaben zum Körpergewicht von 30 Probandinnen und Probanden.

Nr.	Körpergewicht	Nr.	Körpergewicht
1	61,2 kg	16	62,2 kg
2	72,8 kg	17	79,5 kg
3	62,3 kg	18	61,0 kg
4	85,7 kg	19	55,5 kg
5	91,8 kg	20	98,9 kg
6	64,9 kg	21	94,4 kg
7	74,3 kg	22	85,3 kg
8	95,2 kg	23	93,0 kg
9	87,3 kg	24	72,8 kg
10	82,4 kg	25	88,2 kg
11	78,4 kg	26	87,4 kg
12	91,2 kg	27	68,6 kg
13	89,3 kg	28	90,5 kg
14	76,3 kg	29	85,0 kg
15	68,2 kg	30	71,9 kg

Berechnung der Varianz

Für die Berechnung der Varianz empfiehlt sich die Anlage einer Hilfstabelle, über die sich die beiden benötigten Größen – das arithmetische Mittel sowie die Summe der quadrierten Abstände der Werte vom arithmetischen Mittel – auf einfache Art und Weise ermitteln lassen.

Nr.	Körpergewicht	$(x_i - \bar{x})^2$
1	61,2	323,28
2	72,8	40,70
3	62,3	284,93
4	85,7	42,51
5	91,8	159,26
6	64,9	203,92
7	74,3	23,81
8	95,2	256,64
9	87,3	65,93
10	82,4	10,37
11	78,4	0,61
12	91,2	144,48
13	89,3	102,41
14	76,3	8,29
15	68,2	120,56
16	62,2	288,32
17	79,5	0,10
18	61,0	330,51
19	55,5	560,74
20	98,9	388,88
21	94,4	231,65
22	85,3	37,45
23	93,0	190,99
24	72,8	40,70
25	88,2	81,36
26	87,4	67,57
27	68,6	111,94
28	90,5	128,14
29	85,0	33,87
30	71,9	53,00
Σ	2375,5	4332,96
$\emptyset$	79,18	144,43

Die Varianz dieser Verteilung liegt bei 144,43 kg².

(Für alle Softwarenutzer: Die Stichprobenvarianz liegt bei 149,41 kg².)

Berechnung der Standardabweichung

Die Standardabweichung berechnet sich als positive Wurzel der Varianz.

$$s = +\sqrt{s^2}$$

$$s = 12,02$$

Die Standardabweichung dieser Verteilung liegt bei 12,02 kg.

(Für alle Softwarenutzer: Die Wurzel der Stichprobenvarianz beträgt 12,22 kg.)

Da hier keine unterschiedlich dimensionierten Verteilungen miteinander verglichen werden sollen (zum Beispiel eine Gewichtsverteilung in kg und eine Gewichtsverteilung in g) erübrigt sich an dieser Stelle die Bestimmung des Variationskoeffizienten.

3 Übungsaufgaben

Auch bei der Übungsaufgabe bleiben wir bei den Beispieldaten aus der vergangenen Übungseinheit – den Altersangaben der 30 schon nach ihrem Körpergewicht befragten Probandinnen und Probanden.

Nr.	Alter	Nr.	Alter
1	17 Jahre	16	33 Jahre
2	36 Jahre	17	22 Jahre
3	41 Jahre	18	62 Jahre
4	34 Jahre	19	44 Jahre
5	32 Jahre	20	41 Jahre
6	54 Jahre	21	56 Jahre
7	45 Jahre	22	62 Jahre
8	22 Jahre	23	23 Jahre
9	71 Jahre	24	86 Jahre
10	14 Jahre	25	41 Jahre
11	86 Jahre	26	65 Jahre
12	44 Jahre	27	53 Jahre
13	34 Jahre	28	35 Jahre
14	21 Jahre	29	21 Jahre
15	54 Jahre	30	19 Jahre

a) Bestimmen Sie die Varianz.

b) Bestimmen Sie die Standardabweichung.

STATISTIK I – Lösung 05**Varianz und Standardabweichung**

In einer Erhebung wurden 30 Probanden nach ihrem Alter befragt.

Nr.	Alter	Nr.	Alter
1	17 Jahre	16	33 Jahre
2	36 Jahre	17	22 Jahre
3	41 Jahre	18	62 Jahre
4	34 Jahre	19	44 Jahre
5	32 Jahre	20	41 Jahre
6	54 Jahre	21	56 Jahre
7	45 Jahre	22	62 Jahre
8	22 Jahre	23	23 Jahre
9	71 Jahre	24	86 Jahre
10	14 Jahre	25	41 Jahre
11	86 Jahre	26	65 Jahre
12	44 Jahre	27	53 Jahre
13	34 Jahre	28	35 Jahre
14	21 Jahre	29	21 Jahre
15	54 Jahre	30	19 Jahre

a) Bestimmen Sie die Varianz.

Nr.	Alter	$(x_i - \bar{x})^2$
1	17	638,57
2	36	39,31
3	41	1,61
4	34	68,39
5	32	105,47
6	54	137,59
7	45	7,45
8	22	410,87
9	71	825,41
10	14	799,19
11	86	1912,31
12	44	2,99
13	34	68,39
14	21	452,41
15	54	137,59
16	33	85,93
17	22	410,87
18	62	389,27
19	44	2,99
20	41	1,61
21	56	188,51
22	62	389,27
23	23	371,33
24	86	1912,31
25	41	1,61
26	65	516,65

27	53	115,13
28	35	52,85
29	21	452,41
30	19	541,49
Σ	1268	11039,86
$\emptyset$	42,27	368,00

Die Varianz dieser Verteilung liegt bei 368,00 Jahren².

(Für alle Softwarenutzer: Die Stichprobenvarianz liegt bei 380,69 Jahren².)

b) Bestimmen Sie die Standardabweichung.

Die Standardabweichung berechnet sich als positive Wurzel der Varianz.

$$s = +\sqrt{s^2}$$

$$s = 19,18$$

Die Standardabweichung dieser Verteilung liegt bei 19,18 Jahren.

(Für alle Softwarenutzer: Die Wurzel der Stichprobenvarianz beträgt 19,51 Jahren.)

STATISTIK I – Übung 06

Schiefe und Wölbung

1 Kurze Wiederholung

Eine dritte Form von Verteilungsparametern?

Neben den Maßen der zentralen Tendenz (Zentrum einer Verteilung) und den Dispersionsparametern (Streuung der Werte einer Verteilung um dieses Zentrum), lassen sich Verteilungen auch – wenn dies auch weniger gebräuchlich ist – über ihre Form charakterisieren. Dies kann über die Schiefe (symmetrisch, linkssteil/rechtsschief oder rechtssteil/linksschief) sowie über die Wölbung (ähnlich der einer Normalverteilung, spitzer als die einer Normalverteilung oder flacher als die einer Normalverteilung) geschehen. Die Schiefe kann über den Momentenkoeffizienten der Schiefe sowie über den Quartilskoeffizienten der Schiefe, die Wölbung über die Kurtosis / Exzeß bestimmt werden.

Momentenkoeffizient der Schiefe

Die Formel für die Berechnung des Momentenkoeffizienten der Schiefe basiert auf der bereits bekannten Formel für die Berechnung der Varianz (d.h. potenzierte durchschnittliche Abweichung der Werte einer Verteilung von deren arithmetischem Mittel). Da die Berechnung des Momentenkoeffizienten die Berechnung des arithmetischen Mittels voraussetzt, kann dieser nur für metrische Daten ermittelt werden. Für die Berechnung des Momentenkoeffizienten g_m werden drei Formeln benötigt:

$$g_m = \frac{m_3}{s^3}$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

$$s^3 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3$$

Liegt der Momentenkoeffizient der Schiefe nahe Null, handelt es sich um eine symmetrische Verteilung. Positive Werte deuten dagegen auf eine linkssteile, negative auf eine rechtssteile Verteilung hin. Wer die Formel für s^3 näher betrachtet, wird feststellen, dass es sich um die Formel für die Varianz handelt, aus der die Wurzel gezogen (ergibt die Standardabweichung) und die anschließend mit 3 potenziert wird. Bei s^3 handelt es sich also in der Tat um die dritte Potenz der Standardabweichung s .

Quartilskoeffizient der Schiefe

Liegen lediglich ordinalskalierte Daten vor, kann der Momentenkoeffizient der Schiefe nicht berechnet werden. Möglich ist allerdings die Berechnung des Quartilskoeffizienten der Schiefe, der anstelle des arithmetischen Mittels auf die ebenfalls bereits bekannten Quartile aufbaut.

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}}$$

Die Interpretation des Quartilskoeffizienten erfolgt analog zu der des Momentenkoeffizienten: Liegt der Wert nahe Null, handelt es sich um eine symmetrische Verteilung, während positive Werte auf eine linkssteile und negative Werte auf eine rechtssteile Verteilung hindeuten. Zu beachten ist außerdem: Beträgt der Interquartilsabstand ($x_{0,75} - x_{0,25}$) Null, kann der Quartilskoeffizient der Schiefe nicht berechnet werden, da eine Division durch Null nicht möglich ist.

Kurtosis / Exzeß

Neben der Linkssteilheit/Rechtssteilheit von Verteilungen ist die Wölbung einer Verteilung ein weiteres interessantes Kriterium. Mit Hilfe der Kurtosis (auch als Exzeß bezeichnet), kann festgestellt werden, inwieweit die Wölbung einer Verteilung der Wölbung der bekannten Normalverteilung gleicht. Da die Formel voraussetzt, dass das arithmetische Mittel berechnet werden kann, lässt sich die Kurtosis – wie bereits der Momentenkoeffizient der Schiefe – nur dann berechnen, wenn metrisch skalierte Daten vorliegen.

$$g_k = \frac{m_4}{s^4} - 3$$

$$m_4 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^4$$

$$s^4 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4$$

Ergibt sich ein Wert nahe Null, entspricht die Wölbung der Verteilung der Wölbung einer Normalverteilung. Bei einem positiven Wert ist von einer „spitzeren“ Form der Verteilung, bei einem negativen Wert dagegen von einer „flacheren“ Form der Verteilung auszugehen. Die Subtraktion von 3 in der Hauptformel ist übrigens darauf zurückzuführen, dass die Normalverteilung eine Kurtosis von 3 aufweist – durch das Abziehen von 3 vom Ergebnis, ergibt sich bei völliger Gleichheit mit der Normalverteilung also ein Wert von Null und somit die Möglichkeit, das Ergebnis analog zum Momentenkoeffizienten der Schiefe zu interpretieren. Ebenso wie beim Momentenkoeffizienten der Schiefe ist die Interpretation der Kurtosis nur dann sinnvoll, wenn eine unimodale Verteilung vorliegt – und ebenso wie beim Momentenkoeffizienten findet sich auch hier in der Formel für s^4 die Varianz bzw. die Standardabweichung wieder, die hier anstelle mit 3 mit 4 potenziert wird. Wurden Varianz und Standardabweichung für die vorliegenden Daten bereits berechnet, lässt sich die Berechnung des Momentenkoeffizienten sowie der Kurtosis durch Rückgriff auf die Standardabweichung also abkürzen.

2 Beispielrechnungen

An einer Fertigungsanlage werden 20 Polymerbauteile als Zufallsstichprobe aus der laufenden Produktion entnommen und gewogen. Die (absoluten) Abweichungen von einem produktionstechnisch bedingten Idealgewicht in Gramm werden in einer Tabelle festgehalten.

Bauteil	Abweichung	Bauteil	Abweichung
1	0	11	1
2	6	12	2
3	3	13	0
4	2	14	3
5	2	15	5
6	4	16	7
7	3	17	2
8	1	18	2
9	1	19	1
10	0	20	3

Berechnung des Momentenkoeffizienten

Ein Blick auf die Formeln verrät, dass eine Hilfstabelle zur Berechnung dreier Termini erforderlich ist.

$$g_m = \frac{m_3}{s^3}$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

$$s^3 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3$$

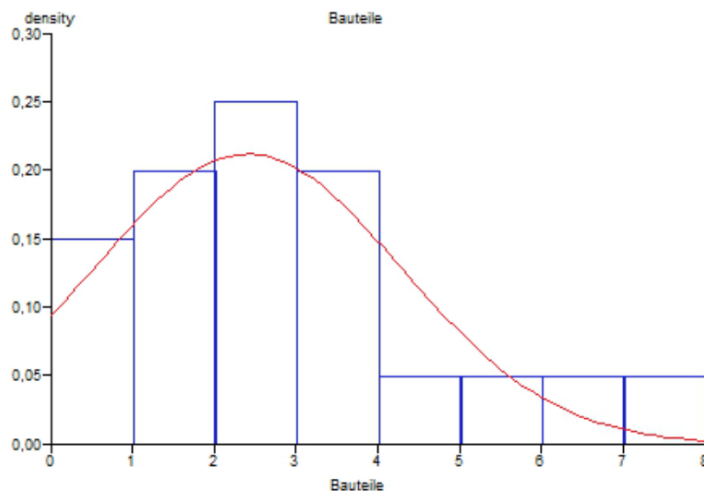
Bauteil	x_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
1	0	5,76	-13,82
2	6	12,96	46,66
3	3	0,36	0,22
4	2	0,16	-0,06
5	2	0,16	-0,06
6	4	2,56	4,10
7	3	0,36	0,22
8	1	1,96	-2,74
9	1	1,96	-2,74
10	0	5,76	-13,82
11	1	1,96	-2,74
12	2	0,16	-0,06
13	0	5,76	-13,82
14	3	0,36	0,22
15	5	6,76	17,58
16	7	21,16	97,34
17	2	0,16	-0,06
18	2	0,16	-0,06
19	1	1,96	-2,74
20	3	0,36	0,22
Σ	48	70,80	113,76
$\emptyset$	2,4	3,54	5,69

$$g_m = \frac{m_3}{s^3} = \frac{5,69}{6,66} = 0,85$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 = 5,69$$

$$s^3 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3 = \sqrt{3,54}^3 = 6,66$$

Der Momentenkoeffizient der Schiefe liegt bei 0,85 – die Verteilung ist somit leicht linkssteil. Ein Blick auf die Verteilungskurve (erstellt mit Smith's Statistical Package) bestätigt diese Interpretation.



Berechnung des Quartilkoeffizienten

Der Quartilkoeffizient der Schiefe berechnet sich aus den drei Quartilswerten.

Die geordnete Verteilung: 0; 0; 0; 1; 1; 1; 1; 2; 2; 2; 2; 2; 3; 3; 3; 3; 4; 5; 6; 7

Ergibt $(n \cdot p)$ einen ganzzahligen Wert (k), berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

$$(n \cdot p) = (20 \cdot 0,25) = 5 \rightarrow k = 5; k+1 = 6 \rightarrow x_p = (1+1)/2 = 1$$

$$(n \cdot p) = (20 \cdot 0,50) = 10 \rightarrow k = 10; k+1 = 11 \rightarrow x_p = (2+2)/2 = 2$$

$$(n \cdot p) = (20 \cdot 0,75) = 15 \rightarrow k = 15; k+1 = 16 \rightarrow x_p = (3+3)/2 = 3$$

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}} = \frac{(3 - 2) - (2 - 1)}{(3 - 1)} = \frac{0}{2} = 0$$

Der Quartilkoeffizient der Schiefe beträgt 0. Dies legt eine symmetrische Verteilung nahe und scheint zunächst im Widerspruch zum Momentenkoeffizient der Schiefe zu stehen. Tatsächlich lässt die Kurve aber beide Interpretationen zu – und das Beispiel verdeutlicht, wie problematisch bisweilen die Subjektivität der Ergebnisinterpretation (ohne definierte Grenzwerte) sein kann.

Berechnung der Kurtosis

Auch für die Berechnung der Kurtosis empfiehlt sich die Anlage einer Hilfstabelle.

Bauteil	x_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^4$
1	0	5,76	33,18
2	6	12,96	167,96
3	3	0,36	0,13
4	2	0,16	0,03
5	2	0,16	0,03
6	4	2,56	6,55
7	3	0,36	0,13
8	1	1,96	3,84
9	1	1,96	3,84
10	0	5,76	33,18
11	1	1,96	3,84

12	2	0,16	0,03
13	0	5,76	33,18
14	3	0,36	0,13
15	5	6,76	45,70
16	7	21,16	447,75
17	2	0,16	0,03
18	2	0,16	0,03
19	1	1,96	3,84
20	3	0,36	0,13
Σ	48	70,80	783,50
$\emptyset$	2,4	3,54	39,18

$$g_k = \frac{m_4}{s^4} - 3 = \frac{39,18}{12,53} - 3 = 0,13$$

$$m_4 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^4 = 39,18$$

$$s^4 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4 = \sqrt{3,54^4} = 12,53$$

Die Kurtosis liegt mit 0,13 nahe der Null – damit ist die Wölbung ähnlich der einer Normalverteilung. Auch diese Annahme wird durch die Betrachtung der mit SSP erstellten Kurve bestätigt.

3 Übungsaufgaben

Der gleichen Fertigungsanlage werden erneut 20 Polymerbauteile als Zufallsstichprobe aus der laufenden Produktion entnommen und gewogen. Die diesmal deutlich größer ausfallenden (absoluten) Abweichungen von einem Idealgewicht in Gramm werden in einer Tabelle festgehalten.

Bauteil	Abweichung	Bauteil	Abweichung
1	12	11	20
2	27	12	17
3	13	13	19
4	10	14	21
5	7	15	30
6	5	16	44
7	11	17	56
8	23	18	8
9	12	19	7
10	11	20	3

- Berechnen Sie den Momentenkoeffizienten der Schiefe.
- Berechnen Sie den Quartilkoeffizienten der Schiefe.
- Berechnen Sie die Kurtosis.

STATISTIK I – Lösung 06**Schiefe und Wölbung**

An einer Fertigungsanlage werden 20 Polymerbauteile als Zufallsstichprobe aus der laufenden Produktion entnommen und gewogen. Die (absoluten) Abweichungen von einem produktionstechnisch bedingten Idealgewicht in Gramm werden in einer Tabelle festgehalten.

Bauteil	Abweichung	Bauteil	Abweichung
1	12	11	20
2	27	12	17
3	13	13	19
4	10	14	21
5	7	15	30
6	5	16	44
7	11	17	56
8	23	18	8
9	12	19	7
10	11	20	3

a) Berechnen Sie den Momentenkoeffizienten der Schiefe.

Zunächst wird die bereits bekannte Hilfstabelle mit zwei zusätzlichen Spalten angelegt.

Bauteil	x_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
1	12	33,64	-195,11
2	27	84,64	778,69
3	13	23,04	-110,59
4	10	60,84	-474,55
5	7	116,64	-1259,71
6	5	163,84	-2097,15
7	11	46,24	-314,43
8	23	27,04	140,61
9	12	33,64	-195,11
10	11	46,24	-314,43
11	20	4,84	10,65
12	17	0,64	-0,51
13	19	1,44	1,73
14	21	10,24	32,77
15	30	148,84	1815,85
16	44	686,44	17984,73
17	56	1459,24	55742,97
18	8	96,04	-941,19
19	7	116,64	-1259,71
20	3	219,04	-3241,79
Σ	356	3379,20	66103,68
$\emptyset$	17,8	168,96	3305,18

Auf Basis der tabellierten Werte ermittelt sich der Momentenkoeffizient der Schiefe wie folgt:

$$g_m = \frac{m_3}{s^3} = \frac{3305,18}{2196,22} = 1,50$$

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 = 3305,18$$

$$s^3 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3 = \sqrt{168,96}^3 = 2196,22$$

Der Momentenkoeffizient der Schiefe liegt bei 1,50 – die Verteilung ist somit linkssteil.

b) Berechnen Sie den Quartilkoeffizienten der Schiefe.

Der Quartilkoeffizient der Schiefe berechnet sich aus den drei Quartilswerten.

Die geordnete Verteilung: 3; 5; 7; 7; 8; 10; 11; 11; 12; 12; 13; 17; 19; 20; 21; 23; 27; 30; 44; 56

Ergibt $(n \cdot p)$ einen ganzzahligen Wert (k) , berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2} (x_{(k)} + x_{(k+1)})$$

$$(n \cdot p) = (20 \cdot 0,25) = 5 \rightarrow k = 5; k+1 = 6 \rightarrow x_p = (8+10)/2 = 9$$

$$(n \cdot p) = (20 \cdot 0,50) = 10 \rightarrow k = 10; k+1 = 11 \rightarrow x_p = (12+13)/2 = 12,5$$

$$(n \cdot p) = (20 \cdot 0,75) = 15 \rightarrow k = 15; k+1 = 16 \rightarrow x_p = (21+23)/2 = 22$$

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}} = \frac{(22 - 12,5) - (12,5 - 9)}{(22 - 9)} = \frac{6}{13} = 0,46$$

Auch hier ergibt sich ein positiver Wert – die Verteilung ist somit linkssteil.

c) Berechnen Sie die Kurtosis.

Auch für die Berechnung der Kurtosis wird zunächst eine Hilfstabelle angelegt.

Bauteil	x_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^4$
1	12	33,64	1131,65
2	27	84,64	7163,93
3	13	23,04	530,84
4	10	60,84	3701,51
5	7	116,64	13604,89
6	5	163,84	26843,55
7	11	46,24	2138,14
8	23	27,04	731,16
9	12	33,64	1131,65
10	11	46,24	2138,14
11	20	4,84	23,43
12	17	0,64	0,41
13	19	1,44	2,07
14	21	10,24	104,86
15	30	148,84	22153,35
16	44	686,44	471199,87

17	56	1459,24	2129381,38
18	8	96,04	9223,68
19	7	116,64	13604,89
20	3	219,04	47978,52
Σ	356	3379,20	2752787,90
$\emptyset$	17,8	168,96	137639,40

$$g_k = \frac{m_4}{s^4} - 3 = \frac{137639,40}{28547,48} - 3 = 1,82$$

$$m_4 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^4 = 137639,40$$

$$s^4 = \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4 = \sqrt{168,96}^4 = 28547,48$$

Der positive Wert der Kurtosis weist auf einen im Vergleich zur Wölbung der Normalverteilung spitzeren Verlauf der Verteilung hin.

STATISTIK I – Übung 07**Box-Plots und Stem-and-Leaf-Diagramme****1 Kurze Wiederholung**Warum nur zwei grafische Darstellungsformen?

Im Rahmen der Vorlesungen haben wir – kurz – eine ganze Reihe grafischer Darstellungsformen für statistische Daten betrachtet – so etwa die Balken- und Kreisdiagramme, die Histogramme oder die 2D- und 3D-Streudiagramme. Aus praktischen Erwägungen und im Sinne der Eingrenzung des Klausurstoffs, müssen aber lediglich zwei dieser grafischen Darstellungsformen in der Prüfung konstruiert werden können: Der (erweiterte) Box-Plot sowie das Stem-and-Leaf-Diagramm.

(Erweiterter) Box-Plot

Der Box-Plot ist eine der wohl spannendsten grafischen Darstellungsformen, welche die deskriptive Statistik zu bieten hat. In dieser einen Grafik finden sich komprimiert Angaben zu einer Vielzahl von Verteilungsparametern wieder, die wir in den vorangegangenen Übungseinheiten betrachtet haben. So kann man neben Lagemaßen (Median, Quartilswerte) auch Streuungsmaße (Spannweite, IQR) sowie die Form der Verteilung (linksteil, symmetrisch oder rechtssteil) direkt aus dem Box-Plot ablesen – und sogar über das Vorhandensein von Ausreißern im Datensatz lässt sich auf Basis der Konstruktionsvorschrift für den Box-Plot eine Feststellung treffen. Der Box-Plot gestattet also Aussagen über Zentrum, Streuung, Form und Ausreißer einer Verteilung und ist damit die informationsdichteste grafische Darstellungsform, die wir im Rahmen der Vorlesung kennenlernen werden. Ein noch größeres Informationspotential entfaltet der Box-Plot übrigens beim grafischen Vergleich von Verteilungen durch das Nebeneinanderstellen mehrerer Box-Plots.

Bei der Konstruktion von Box-Plots wird in einfache Box-Plots (bei denen die Zäune jeweils bis zum größten sowie bis zum kleinsten Wert im Datensatz reichen) und in sogenannte erweiterte Box-Plots (bei deren Konstruktion die Grenzen der Zäune über den Interquartilsabstand berechnet und in denen Ausreißer und Extremwerte ausgewiesen werden) unterschieden. Nachfolgend wird in dieser Übungseinheit nur noch der erweiterte Box-Plot betrachtet. Ein solcher erweiterter Box-Plot besteht aus drei Komponenten: Der eigentlichen Box, den Zäunen der Box sowie möglicherweise einzuzeichnenden Ausreißern oder Extremwerten, sollten solche im Datensatz auftauchen. Die Konstruktion eines erweiterten Box-Plots erfolgt demnach ebenfalls in drei Schritten.

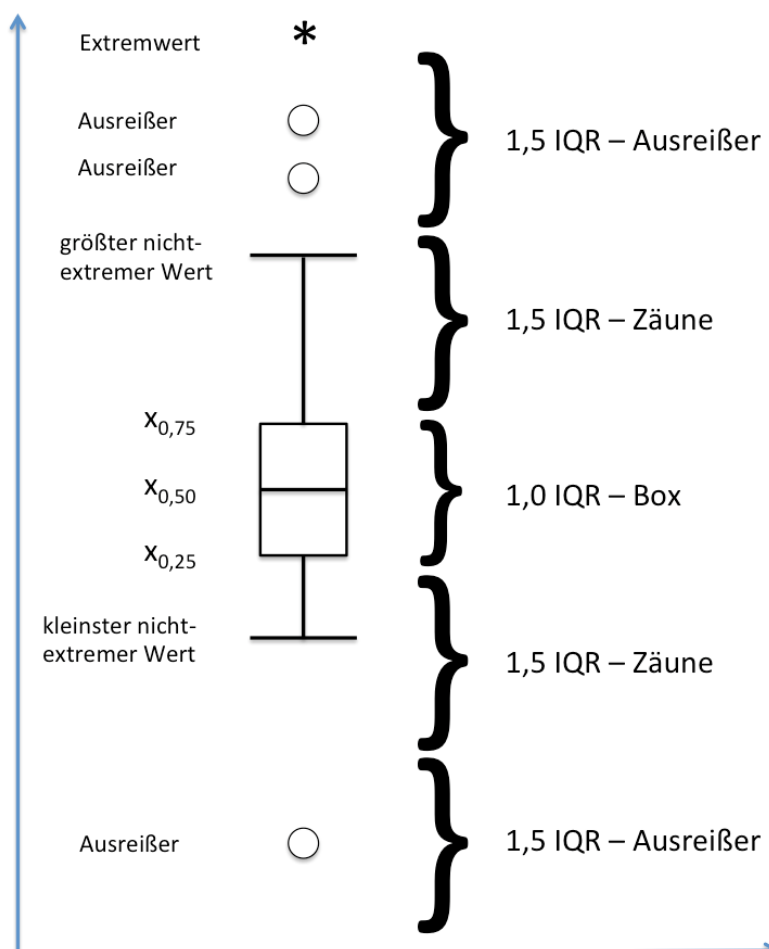
Schritt 1: Konstruktion der Box. Hierfür werden drei Werte benötigt: Das obere Quartil (obere Grenze der Box), das untere Quartil (untere Grenze der Box) sowie der Median (wird als zusätzliche Linie in die Box eingezeichnet). Sollte der Median mit einem der beiden Quartile identisch sein, wird die entsprechende Grenze mit doppelter Strichstärke gekennzeichnet. Sind die drei Quartilswerte identisch, kann keine Box konstruiert werden – in diesem Falle wird der Box durch eine dicke Linie an der Stelle $x_{0,75} = x_{0,50} = x_{0,25}$ ersetzt (Zäune und Ausreißer könnten in diesem Fall aber noch vorkommen).

Aus der Lage des Medians innerhalb der Box lässt sich übrigens eine Aussage über die Form der Verteilung herauslesen: Liegt der Median (ungefähr) in der Mitte, handelt es sich um eine symmetrische Verteilung, liegt der Median dagegen nahe der unteren Grenze der Box, so ist die Verteilung rechtsschief und linksteil. Liegt der Median nahe an der oberen Grenze der Box, so ist die Verteilung dementsprechend rechtssteil und linksschief. Der Box-Plot kann daher (etwa in der Klausur) auch als visuelle Kontrolle für die Richtigkeit der Berechnung des Momentenkoeffizienten der Schiefe oder auch des Quartilkoeffizienten der Schiefe herangezogen werden.

Schritt 2: Konstruktion der Zäune. Da die Box vom oberen zum unteren Quartil verläuft, entspricht ihre Höhe genau dem Interquartilsabstand (IQR; Inter Quartile Range – siehe Übungseinheiten zu den Streuungsmaßen). Die Berechnung des IQR liefert uns die benötigten Angaben für das Einzeichnen der Zäune. Der 1,5-fache Wert des IQR wird nämlich zum oberen Quartilswert addiert bzw. vom unteren Quartilswert subtrahiert, um die virtuellen (nicht einzuzeichnenden) Maximal- bzw. Minimalwerte für die Grenzen der Zäune zu ermitteln. Anschließend wird der größte Wert bzw. der kleinste Werte im Datensatz ermittelt, der noch in den Bereich $x_{0,75} + 1,5 \text{ IQR}$ bzw. in den Bereich $x_{0,25} - 1,5 \text{ IQR}$ fällt. Der obere bzw. der untere Zaun werden dann bis zu diesen Werten gezeichnet.

In der Praxis kann es vorkommen, dass der obere Zaun, der untere Zaun oder auch beide Zäune entfallen, da keine Werte aus dem Datensatz in den benannten Bereichen liegen. Auch kann der Fall eintreten, dass Zäune genau bis zu den Maximal- bzw. Minimalwerten reichen, weil reale Werte im Datensatz exakt an der Stelle $x_{0,75} + 1,5 \text{ IQR}$ bzw. $x_{0,25} - 1,5 \text{ IQR}$ liegen. Von beiden Fällen sollte man sich also keinesfalls irritieren lassen – insbesondere nicht in der Klausur.

Schritt 3: Identifikation von Ausreißern und Extremwerten. Liegen Werte im Datensatz oberhalb von $x_{0,75} + 1,5 \text{ IQR}$ bzw. unterhalb von $x_{0,25} - 1,5 \text{ IQR}$, handelt es sich um Ausreißer. Beim erweiterten Box-Plot wird dabei noch in Ausreißer und „extreme“ Ausreißer – die sogenannten Extremwerte - unterschieden, indem eine weitere „virtuelle“ Grenze basierend auf dem IQR errichtet wird. Werte, die zwischen $x_{0,75} + 1,5 \text{ IQR}$ und $x_{0,75} + 3 \text{ IQR}$ bzw. zwischen $x_{0,25} - 1,5 \text{ IQR}$ und $x_{0,25} - 3 \text{ IQR}$ liegen, werden als „normale“ Ausreißer mit einem Kreis markiert. Werte, die sogar noch außerhalb dieser Bereiche liegen, gelten als Extremwerte und sind mit einem Sternchen zu markieren. Sowohl die Ausreißer als auch die Extremwerte werden in der Regel noch mit der fortlaufenden Nummer des Datensatzes versehen, um diesen in nachfolgenden Untersuchungen schneller auffinden zu können.



Wie schon bei der Konstruktion der Zäune, kann es auch bei der Identifikation von Ausreißern und Extremwerten vorkommen, dass in eine oder auch beide Richtungen keine entsprechende Werte zu finden sind und daher nichts in den Box-Plot eingezeichnet wird. Auch hiervon sollte man sich – sollte ein solcher Fall in der Klausur vorkommen – also nicht irritieren lassen.

Da in der deskriptiven Statistik keine allgemeingültige Definition für Ausreißer existiert, kann es im Übrigen auch an anderer Stelle (etwa bei Unklarheiten über die Einordnung eines Wertes als Ausreißer) sinnvoll sein, auf das Konstruktionsprinzip des Box-Plots zurückzugreifen. Die Unterscheidung in Ausreißer und extreme Ausreißer ist außerhalb der Box-Plot-Konstruktion allerdings eher unüblich.

Stem-and-Leaf-Diagramm

Die zweite, ebenfalls recht spezielle Form der grafischen Darstellung von Daten, die wir im Rahmen dieser Vorlesung näher betrachten, ist das sogenannte Stem-and-Leaf- oder Stamm-Blatt-Diagramm. Während die Besonderheit des Box-Plots in der Fülle der enthaltenen Informationen (Zentrum, Streuung, Form, Ausreißer) liegt, zeichnet sich das Stem-and-Leaf-Diagramm dadurch aus, dass man aus diesem als einzige grafische Darstellungsform die Originaldaten wieder herauslesen kann.

Dies ist deshalb möglich, da das Stem-and-Leaf-Diagramm nicht gezeichnet, sondern unmittelbar mit den Zahlenwerten der Originaldaten konstruiert wird. Diese werden hierfür in einen „Stamm“ sowie in an diesen angeheftete „Blätter“ zerlegt. Liegen beispielsweise nur Werte zwischen 10 und 30 vor, so bietet es sich an, den Stamm in 10er-Schritten aufzubauen, d.h.

```
10 |
20 |
30 |
```

Die konkreten Werte 11, 18, 22, 34 und 39 würden nun wie folgt in das Diagramm eingetragen:

```
10 | 1 8
20 | 2
30 | 4 9
```

Zu notieren wäre außerdem die Stammbreite (in diesem Fall 10), die Anzahl der Werte, die jeweils einem Blatt entsprechen (in diesem Fall 1) sowie ggf. noch die Anzahl der Ausreißer (in diesem Fall 0), die allein schon aus praktischen Erwägungen – ein einzelner Ausreißer beim Wert 240 würde im vorliegenden Beispiel 20 „leere“ Stammzeilen erzeugen – nicht mit eingezeichnet werden. Da in unserem Beispielfall kein Ausreißer existiert, sieht das vollständige Diagramm also wie folgt aus:

```
10 | 1 8
20 | 2
30 | 4 9
```

Stammbreite: 10
Jedes Blatt ein Fall

Betrachtet man das Stem-and-Leaf-Diagramm genauer, so fällt auf, dass es sich im Grunde um ein auf die Seite gekipptes Histogramm handelt – immer vorausgesetzt, man verwendet eine nichtproportionale Schriftart, bei der alle Zeichen exakt die gleiche Breite aufweisen (was in diesem Dokument übrigens nicht der Fall ist). Durch Abänderungen an Stammbreite und Fallzahl lassen sich Stamm-Blatt-Diagramme – insbesondere bei Verteilungen mit vielen Werten – auf äußerst vielfältige Art und Weise konstruieren. Zur Demonstration soll uns an dieser Stelle das bereits bekannte Mini-Beispiel bei einer Stammbreite von 5 dienen:

10 | 1
 10 | 8
 20 | 2
 20 |
 30 | 4
 30 | 9

Stammbreite: 5

Jedes Blatt ein Fall

2 Beispielgrafiken

Auf dem Campus der Hochschule Harz haben wir 20 willkürlich ausgewählte Studierende nach ihrem Alter (in ganzen Jahren) befragt. Dabei ergab sich die folgende Verteilung:

Student	Alter	Student	Alter
1	24	11	21
2	22	12	24
3	23	13	22
4	32	14	26
5	28	15	26
6	62	16	28
7	31	17	31
8	36	18	22
9	22	19	21
10	22	20	26

Konstruktion des Box-Plots

Da für die Konstruktion des Box-Plots die Quartilswerte und der Interquartilsabstand berechnet werden müssen, lohnt sich im ersten Schritt das Festhalten der geordneten Verteilung:

21; 21; 22; 22; 22; 22; 22; 23; 24; 24; 26; 26; 26; 28; 28; 31; 31; 32; 36; 62

Ergibt $(n \cdot p)$ einen ganzzahligen Wert (k) , berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

$$(n \cdot p) = (20 \cdot 0,25) = 5 \rightarrow k = 5; k+1 = 6 \rightarrow x_p = (22+22)/2 = 22$$

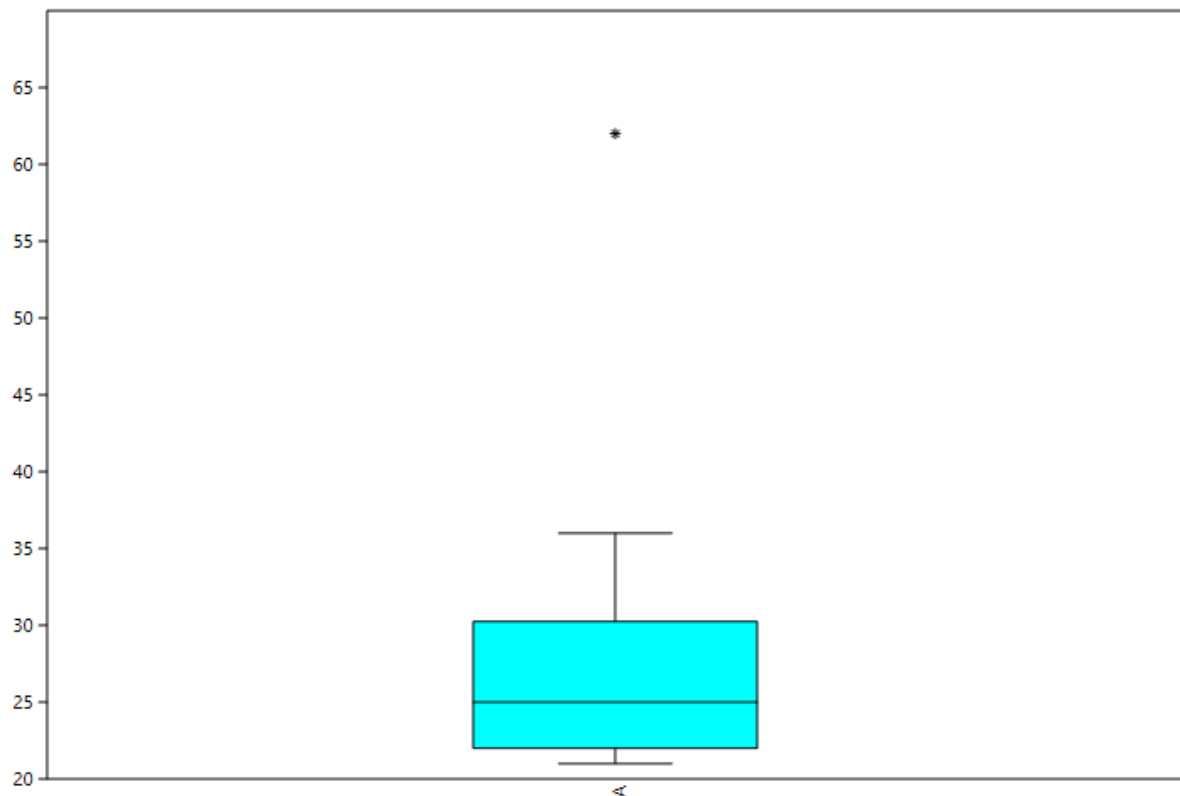
$$(n \cdot p) = (20 \cdot 0,50) = 10 \rightarrow k = 10; k+1 = 11 \rightarrow x_p = (24+26)/2 = 25$$

$$(n \cdot p) = (20 \cdot 0,75) = 15 \rightarrow k = 15; k+1 = 16 \rightarrow x_p = (28+31)/2 = 29,5$$

$$IQR = 29,5 - 22 = 7,5$$

Die Box wird im ersten Schritt also von 29,5 (obere Grenze) zu 22 (untere Grenze) gezeichnet, der Median wird bis 25 eingetragen. Da der 1,5-fache IQR bei 11,25 liegt, endet der obere Zaun beim größten Wert zwischen 29,5 und 40,75 (36), der untere Zaun beim kleinsten Wert zwischen 22 und 10,75 (21). Da die 21 bereits den kleinsten Wert der Verteilung darstellt, können im unteren Bereich der Verteilung weder Ausreißer noch Extremwerte liegen. Ausreißer im oberen Bereich der Verteilung müssten zwischen 40,75 und 52 liegen – hier finden sich in der Tabelle allerdings keine Werte. Der noch verbliebene Wert von 62 stellt damit einen Extremwert dar.

Der mit Hilfe der Software PAST berechnete erweiterte Box-Plot sieht wie folgt aus:



Konstruktion des Stem-and-Leaf-Diagramms

Auch für die Konstruktion des Stem-and-Leaf-Diagramms ist die geordnete Verteilung von Nutzen:

21; 21; 22; 22; 22; 22; 22; 23; 24; 24; 26; 26; 26; 28; 28; 31; 31; 32; 36; 62

Da alle Werte (bis auf einen) zwischen 20 und 40 liegen, führt eine Stammbreite von 10 in diesem Fall (obwohl sie durchaus nicht „falsch“ wäre) zu einem recht kurzen Diagramm, weshalb wir nachfolgend auf eine Stammbreite von 5 ausweichen wollen.

2 | 1 1 2 2 2 2 2 3 4 4

2 | 6 6 6 8 8

3 | 1 1 2

3 | 6

Stammbreite: 5

Jedes Blatt ein Fall

Ein Ausreißer (62)

3 Übungsaufgaben

Parallel zur Befragung der 20 Studierenden, wurden auch 20 willkürlich ausgewählte Professorinnen und Professoren der Hochschule Harz nach ihrem Alter befragt. Dabei ergab sich folgendes Bild:

Prof.	Alter	Prof.	Alter
1	44	11	48
2	61	12	56
3	62	13	66
4	54	14	53
5	55	15	39
6	50	16	42
7	51	17	46
8	44	18	45
9	40	19	60
10	33	20	52

- 1) Konstruieren Sie einen erweiterten Box-Plot.
- 2) Konstruieren Sie ein Stem-and-Leaf-Diagramm.

STATISTIK I – Lösung 07**Box-Plots und Stem-and-Leaf-Diagramme**

Eine Befragung von 20 willkürlich ausgewählten Professorinnen und Professoren der Hochschule Harz nach ihrem Alter ergab folgendes Bild:

Prof.	Alter	Prof.	Alter
1	44	11	48
2	61	12	56
3	62	13	66
4	54	14	53
5	55	15	39
6	50	16	42
7	51	17	46
8	44	18	45
9	40	19	60
10	33	20	52

1) Konstruieren Sie einen erweiterten Box-Plot.

Zur Berechnung der Quartile ordnen wir im ersten Schritt die Werte der Verteilung:

33; 39; 40; 42; 44; 44; 45; 46; 48; 50; 51; 52; 53; 54; 55; 56; 60; 61; 62; 66

Ergibt $(n \cdot p)$ einen ganzzahligen Wert (k) , berechnet sich das Perzentil wie folgt:

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)})$$

$$(n \cdot p) = (20 \cdot 0,25) = 5 \rightarrow k = 5; k+1 = 6 \rightarrow x_p = (44+44)/2=44$$

$$(n \cdot p) = (20 \cdot 0,50) = 10 \rightarrow k = 10; k+1 = 11 \rightarrow x_p = (50+51)/2=50,5$$

$$(n \cdot p) = (20 \cdot 0,75) = 15 \rightarrow k = 15; k+1 = 16 \rightarrow x_p = (55+56)/2=55,5$$

$$IQR = 55,5 - 44 = 11,5$$

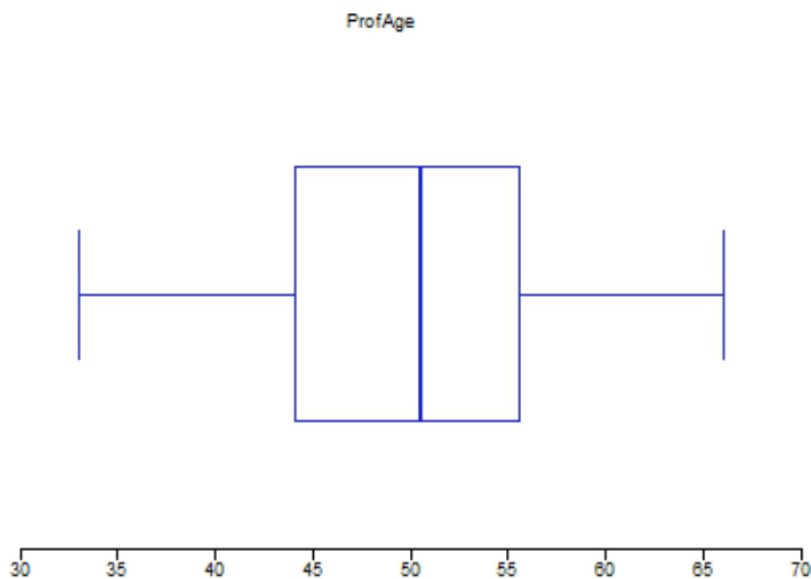
$$1,5 \text{ IQR} = 17,25$$

$$3,0 \text{ IQR} = 34,50$$

Auf Basis dieser Ergebnisse lässt sich der Box-Plot nun wie gewohnt konstruieren:

- Die Box verläuft von 44 (unteres Quartil) bis zu 55,5 (oberes Quartil).
- Der Median (mittleres Quartil) wird in diese Box beim Wert 50,5 eingezeichnet.
- Der obere Zaun könnte maximal bis zum Wert $55,5 + 17,25 = 72,75$ reichen. Der größte in dem Bereich liegende Wert (und damit die Grenze des oberen Zauns) ist die 66. Da sich im Datensatz keine größeren Werte befinden, gibt es nach oben keine Ausreißer.
- Der untere Zaun könnte minimal bis zum Wert $44 - 17,25 = 26,75$ reichen. Der kleinste in diesem Bereich liegende Wert (und damit die Grenze des unteren Zauns) ist die 33. Da sich im Datensatz keine kleineren Werte befinden, gibt es nach unten keine Ausreißer.
- Zu Übungszwecken sei an dieser Stelle noch aufgeführt, dass Ausreißer nach oben zwischen 72,75 und (einschließlich) 90, Ausreißer nach unten zwischen 26,75 und (einschließlich) 9,5 hätten liegen müssen. Werte oberhalb von 90 bzw. unterhalb von 9,5 hätten dagegen als Extremwerte gekennzeichnet werden müssen.

Die Kontrollgrafik wurde in diesem Fall nicht mit PAST, sondern mit SSP (seitlich liegend) erstellt.



2) Konstruieren Sie ein Stem-and-Leaf-Diagramm.

Für das Stem-and-Leaf-Diagramm können wir auf die geordnete Verteilung zurückgreifen:

33; 39; 40; 42; 44; 44; 45; 46; 48; 50; 51; 52; 53; 54; 55; 56; 60; 61; 62; 66

Da alle Zahlen zwischen 30 und 70 liegen (keine Ausreißer) und sich zudem kaum Werte wiederholen, bietet sich eine Stammbreite von 10 sowie die Eintragung eines Blatts je Wert der Verteilung an. Das sich ergebende Stem-and-Leaf-Diagramm sieht dann wie folgt aus:

```

3 | 3 9
4 | 0 2 4 4 5 6 8
5 | 0 1 2 3 4 5 6
6 | 0 1 2 6
  
```

Stammbreite: 10
 Ein Wert pro Blatt
 Keine Ausreißer

Zu Übungszwecken: So sieht das Stem-and-Leaf-Diagramm bei einer Stammbreite von 5 aus:

```

3 | 3
3 | 9
4 | 0 2 4 4
4 | 5 6 8
5 | 0 1 2 3 4
5 | 5 6
6 | 0 1 2
6 | 6
  
```

Stammbreite: 5
Ein Wert pro Blatt
Keine Ausreißer

STATISTIK I – Übung 08**Bravais-Pearson-Korrelationskoeffizient****1 Kurze Wiederholung**Was sind Zusammenhangsmaße?

Haben wir bislang immer nur eine Variable x (Lagemaße, Streuungsmaße, Schiefe und Wölbung, Grafiken etc.) individuell analysiert, sollen nun zwei Variablen x und y gleichzeitig betrachtet werden, um festzustellen, ob zwischen diesen ein Zusammenhang besteht. Dabei gilt, dass ein solcher Zusammenhang unterstellt werden kann – aber nicht existieren muss (siehe hierzu Aussagen zu Korrelation und Kausalität) – wenn die Variablen miteinander korrelieren (d.h. sich gleichmäßig zueinander verhalten). Hierbei wird in gleichsinnige Korrelationen (größere x -Werte fallen mit größeren y -Werten zusammen, kleinere x -Werte fallen mit kleineren y -Werten zusammen) und gegensinnige Korrelationen (größere x -Werte fallen mit kleineren y -Werten zusammen, kleinere x -Werte fallen mit größeren y -Werten zusammen) unterschieden. Die Stärke gleichsinniger oder auch gegensinniger Korrelationen (nicht aber die Stärke eines möglichen inhaltlichen Zusammenhangs) wird über sogenannte Zusammenhangsmaße wiedergegeben.

Welches Zusammenhangsmaß sich berechnen lässt, hängt – wie so vieles in der Statistik – vom Skalenniveau der betrachteten Variablen ab. Liegen nominalskalierte Variablen vor, eignet sich der χ^2 -Koeffizient, für ordinalskalierte Variablen empfiehlt sich dagegen der Rangkorrelationskoeffizient nach Spearman oder auch der Konkordanzkoeffizient nach Kendall, während sich für metrisch skalierte Daten wiederum der Bravais-Pearson-Korrelationskoeffizient anbietet. Wie in vielen Fällen ist die Berechnung der Parameter „abwärtskompatibel“, nicht aber „aufwärtskompatibel“, d.h. für metrisch skalierte Daten könnte beispielsweise auch der χ^2 -Koeffizient oder der Rangkorrelationskoeffizient nach Spearman berechnet werden, für ordinalskalierte Daten aber niemals der Bravais-Pearson-Korrelationskoeffizient. Grundsätzlich immer möglich ist auch die unterstützende grafische Zusammenhangsanalyse von Daten, wobei sich für diskrete Daten die Konstruktion gruppiert oder bedingter Balkendiagramme, für stetige Daten dagegen die Konstruktion von zwei- und dreidimensionalen Streudiagrammen sowie Scatterplot-Matrizen empfiehlt.

Korrelation und Kausalität

Von wesentlicher Bedeutung für die Interpretation von Zusammenhangsmaßen ist die Verinnerlichung der Tatsache, dass identifizierte Korrelationen zwar näher untersucht, niemals jedoch unmittelbar inhaltlich interpretiert werden sollten – auch dann nicht, wenn sich eine Interpretation im Hinblick auf einen sachlogischen Zusammenhang zwischen den betrachteten Variablen geradezu aufdrängt. Untersucht man beispielsweise den Zusammenhang zwischen Preis und Verkaufszahlen und identifiziert dabei eine starke Korrelation zwischen beiden Variablen, ist es zwar durchaus logisch davon auszugehen, dass der Preis die Verkaufszahlen beeinflusst (und nicht umgekehrt) – auf Basis eines Zusammenhangsmaßes lässt sich diese Annahme jedoch keinesfalls beweisen, da alle im Rahmen dieser Vorlesung betrachteten Zusammenhangsmaße keine Wirkungsrichtung (x beeinflusst y , y ist abhängig von x) kennen. Tatsächlich kann jede starke Korrelation zweier Variablen auf vier mögliche Ursachen zurückzuführen sein, die wir nachfolgend kurz betrachten wollen.

- **Möglichkeit 1:** Es liegt ein „echter“ inhaltlicher Zusammenhang mit einseitiger Wirkungsrichtung vor, d.h. x beeinflusst tatsächlich y bzw. y ist tatsächlich abhängig von x . Fälschlicherweise wird leider oft davon ausgegangen, dass eine starke Korrelation immer auf einen solchen monokausalen Zusammenhang zurückzuführen ist bzw. diesen sogar irgendwie „beweist“. Diesen Fehler sollte man (auch in der Klausur) nach Möglichkeit vermeiden.

- Möglichkeit 2: Es handelt sich um einen sogenannten beidseitigen Zusammenhang, d.h. x beeinflusst y während y gleichzeitig auch x beeinflusst. Dies ist etwa beim Bekanntheitsgrad eines Produktes und dessen Verkaufszahlen der Fall, da die Bekanntheit einerseits die Verkaufszahlen beeinflusst (ein stark beworbenes Produkt ist bekannter und wird häufiger angefragt), gleichzeitig aber auch die Verkaufszahlen den Bekanntheitsgrad beeinflussen (ein Produkt, das sich etwa über eine Niedrigpreisstrategie viel verkauft, wird mit der Zeit immer bekannter). Derartige beidseitige Zusammenhänge lassen sich nur mit statistischen Komplexverfahren näher untersuchen, die wir im Rahmen dieser Vorlesung nicht kennenlernen werden. Die besondere Schwierigkeit bei der Analyse beidseitiger Zusammenhänge besteht – wie man sich vermutlich leicht vorstellen kann – darin, festzustellen, welcher Anteil des Zusammenhangs sich auf welche Wirkungsrichtung ($x \rightarrow y$; $y \rightarrow x$) zurückführen lässt.

Möglichkeit 3: Es handelt sich um einen rein zufälligen Effekt – eine sogenannte Scheinkorrelation. Korreliert man nur genügend Variablen gegeneinander, wird man immer Kombinationen finden, bei denen sich allein durch Zufall eine starke Korrelation ergibt. In der Vorlesung hatten wir hierzu die Internetseite des US-Juristen Tyler Vigen kennengelernt, der seit Jahren Korrelationen (ausgedrückt über den Bravais-Pearson-Korrelationskoeffizienten) sammelt, bei denen ein inhaltlicher Zusammenhang zwischen beiden Variablen mehr als unplausibel sein dürfte. Auf <http://www.tylervigen.com/spurious-correlations> finden sich viele hervorragende Beispiele dafür, warum eine – auch sehr starke – Korrelation noch lange nicht belegt, dass zwischen den betrachteten Variablen auch ein kausaler Zusammenhang existiert – so etwa zwischen der Scheidungsrate im US-Bundesstaat Maine und dem Pro-Kopf-Konsum von Margarine ($r = 0,99258$) oder dem Pro-Kopf-Konsum von Mozzarella und der Anzahl an Promovenden in den Ingenieurwissenschaften ($r = 0,958648$).

Als Nebenbemerkung sei an dieser Stelle darauf hingewiesen, dass Untersuchungsdesigns, die auf dem wahllosen „Durchspielen“ einer Vielzahl von Korrelationsmöglichkeiten ohne die vorherige Aufstellung inhaltlich begründeter Vermutungen und Forschungsfragen basieren, aufgrund der Existenz derartiger Scheinkorrelationen abzulehnen sind. Testet man nämlich nur genügend Kombinationen von Variablen durch, werden sich letztlich immer starke oder sehr starke Korrelationen finden lassen, die jedoch nicht notwendigerweise auf einen inhaltlichen Zusammenhang zurückzuführen sind. Wir werden dieser Problematik bei der Betrachtung statistischer Testverfahren unter dem Stichwort „p-value hacking“ erneut begegnen.

- Möglichkeit 4: Es handelt sich um einen indirekten Zusammenhang zwischen x und y über eine dritte, nicht betrachtete Variable z . Das geradezu klassische Beispiel hierfür ist die über die Jahre mehrfach bestätigte, äußerst starke Korrelation zwischen der Geburtenrate und der Anzahl von Störchen in deutschen Landkreisen. Obwohl zwischen beiden Variablen – natürlich – kein unmittelbarer Zusammenhang besteht, sind sie doch über eine dritte Variable – den Grad der Urbanisierung – miteinander verbunden. Da der Urbanisierungsgrad sowohl die Geburtenquote (aus soziologischen Gründen) als auch die Storchdichte (aus ökologischen Gründen) gleichermaßen beeinflusst, korrelieren beide Variablen deutlich miteinander. Auch die Möglichkeit eines solchen Zusammenhangs sollte bei der Interpretation von Korrelationen nicht voreilig ausgeschlossen werden – bei betriebs- oder volkswirtschaftlichen Fragestellungen finden sich derartige „Hintergrundzusammenhänge“ sogar recht häufig.

Da eine starke Korrelation zweier Variablen theoretisch auf jede dieser vier Möglichkeiten zurückzuführen sein kann, sollte man sich vor der kausalen Interpretation von Korrelationen tunlichst hüten – auch wenn der inhaltliche Zusammenhang sich vermeintlich „offensichtlich“ präsentiert. Eine Korrelation ist stets nur als ein Hinweis darauf zu begreifen, welche möglichen Zusammenhänge sich für eine vertiefte Analyse mit weiterführenden statistischen Verfahren (insbesondere Tests) anbieten. In vielen Vorlesungen und Lehrbüchern wird diese wirklich wesentliche Erkenntnis mit dem bekannten Merksatz „Korrelation ist nicht gleich Kausalität“ vermittelt.

Der Bravais-Pearson-Korrelationskoeffizient

Liegen metrisch skalierte Daten (natürlich bei beiden Variablen) vor, kann – wie oben erwähnt – der Korrelationskoeffizient nach Bravais-Pearson berechnet werden. Dieser ist allerdings ausschließlich ein Maß für die Stärke einer linearen Korrelation zwischen zwei Variablen. Liegt eine andere Form des Zusammenhangs – beispielsweise ein quadratischer oder logarithmischer Zusammenhang – vor, wird dieser durch Bravais-Pearson nicht bzw. nicht vollumfänglich aufgedeckt. Bei der Interpretation des Korrelationskoeffizienten ist somit unbedingt zu berücksichtigen, dass ein niedriger Wert nicht bedeutet, dass zwischen den untersuchten Variablen keinerlei Zusammenhang besteht – er bedeutet lediglich, dass zwischen den untersuchten Variablen kein linearer Zusammenhang besteht.

Der Bravais-Pearson-Korrelationskoeffizient wird anhand folgender Formel berechnet:

$$r = \frac{\sum_{i=1}^n (x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^n (x_i^2) - n * \bar{x}^2} * \sqrt{\sum_{i=1}^n (y_i^2) - n * \bar{y}^2}}$$

Die Formel lässt erkennen, dass das Anlegen einer Hilfstabelle mit folgenden Spalten sinnvoll ist:

Nr.	x	y	x ²	y ²	(x * y)
1	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
n	...	...	...	...	...
Σ	...	...	...	...	...
∅	...	...	//	//	//

Der Korrelationskoeffizient r liegt stets zwischen -1 und +1 und wird wie folgt interpretiert:

- Bei positiven Werten liegt ein positiver Zusammenhang vor (die Wertepaare liegen auf einer steigenden Geraden), bei negativen Werten ein negativer Zusammenhang (die Wertepaare liegen auf einer fallenden Geraden). Werte nahe Null deuten darauf hin, dass keine lineare Korrelation zwischen den beiden betrachteten Variablen vorliegt.
- Liegt der Betrag (!) von r nahe 0, liegt keine (lineare) Korrelation vor, bei Werten bis 0,5 kann von einer schwachen, bei Werten zwischen 0,5 und 0,8 von einer mittelstarken und bei Werten zwischen 0,8 bis 1,0 von einer starken Korrelation gesprochen werden, wobei der maximal erreichbare Betrag von 1,0 eine perfekte Korrelation (alle Wertepaare liegen aufgereiht auf einer Geraden) anzeigen würde. Die hier aufgezeigten Interpretationsgrenzen sind dabei nicht verbindlich, sondern stellen lediglich (erfahrungswertbasierte) Vorschläge dar.

2 Beispielrechnungen

Ein Eishändler variiert ceteris paribus (unter sonst gleichbleibenden Rahmenbedingungen) die Preise einer Kugel Eis über den Verlauf von sieben Tagen und zeichnet auf, wie viele Kugeln zu dem jeweiligen Preis veräußert wurden.

Tag	Kugelpreis (in Euro)	Verkaufte Kugeln (in Stück)
1	0,50	337
2	0,55	318
3	0,60	226
4	0,65	211
5	0,70	235
6	0,75	117
7	0,80	123

Berechnen Sie den Bravais-Pearson-Korrelationskoeffizienten.

Im ersten Schritt ist die bereits vorgestellte Hilfstabelle anzulegen und auszufüllen.

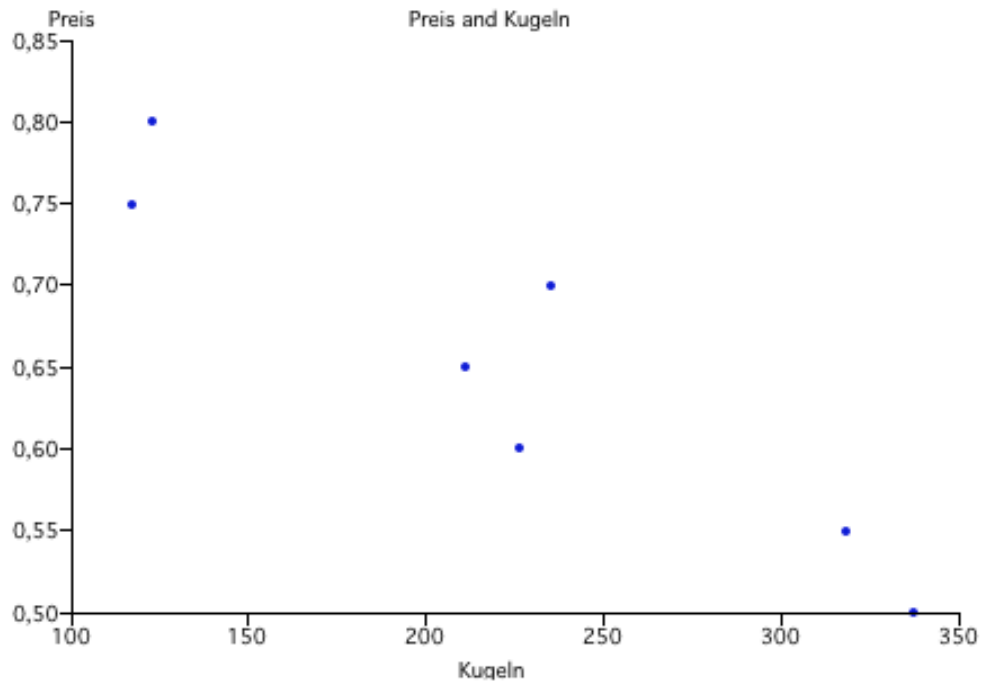
Nr.	x	y	x ²	y ²	(x * y)
1	0,50	337	0,25	113569	168,50
2	0,55	318	0,30	101124	174,90
3	0,60	226	0,36	51076	135,60
4	0,65	211	0,42	44521	137,15
5	0,70	235	0,49	55225	164,50
6	0,75	117	0,56	13689	87,75
7	0,80	123	0,64	15129	98,40
Σ	4,55	1567,00	3,03	394333	966,80
Ø	0,65	223,86	//	//	//

Die auf diese Weise errechneten Werte werden anschließend in die Formel eingetragen.

$$r = \frac{\sum_{i=1}^n (x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^n (x_i^2) - n * \bar{x}^2} * \sqrt{\sum_{i=1}^n (y_i^2) - n * \bar{y}^2}}$$

$$r = \frac{966,80 - 7 * 0,65 * 223,86}{\sqrt{3,03 - 7 * 0,65^2} * \sqrt{394333 - 7 * 223,86^2}} = -0,9372 \approx -0,94$$

Das Ergebnis weist auf eine starke negative Korrelation zwischen beiden Variablen hin – ein Ergebnis, das durch einen Blick auf das umseitige Streudiagramm (im Beispiel erstellt mit SSP) bestätigt wird.



3 Übungsaufgaben

Eine Hochschule befragt zehn Absolventen/innen fünf Jahre nach ihrem Abschluss nach ihrem aktuellen Nettogehalt und ordnet die Summen den (metrisch skalierten!) Punkten in ihrer Abschlussklausur zu. Existiert eine lineare Korrelation zwischen dem Abschneiden in der Klausur und dem Nettogehalt nach fünf Jahren Berufstätigkeit?

Student Nr.	Klausurbewertung	Nettogehalt (in EUR)
1	98	2340
2	72	3750
3	65	1890
4	71	2420
5	56	2830
6	93	3110
7	66	1970
8	82	2480
9	89	2860
10	97	3333

Berechnen und interpretieren Sie den Bravais-Pearson-Korrelationskoeffizienten.

STATISTIK I – Lösung 08**Bravais-Pearson-Korrelationskoeffizient**

Eine Hochschule befragt zehn Absolventen/innen fünf Jahre nach ihrem Abschluss nach ihrem aktuellen Nettogehalt und ordnet die Summen den (metrisch skalierten!) Punkten in ihrer Abschlussklausur zu. Existiert eine lineare Korrelation zwischen dem Abschneiden in der Klausur und dem Nettogehalt nach fünf Jahren Berufstätigkeit?

Student Nr.	Klausurbewertung	Nettogehalt (in EUR)
1	98	2340
2	72	3750
3	65	1890
4	71	2420
5	56	2830
6	93	3110
7	66	1970
8	82	2480
9	89	2860
10	97	3333

Berechnen und interpretieren Sie den Bravais-Pearson-Korrelationskoeffizienten.

Der erste Schritt besteht auch hier wieder im Anlegen der benötigten Hilfstabelle.

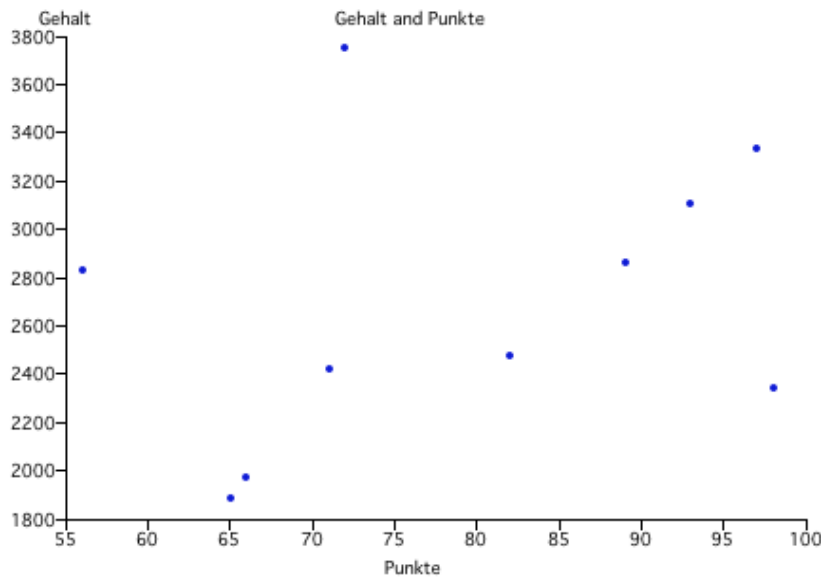
Nr.	x	y	x ²	y ²	(x * y)
1	98	2340	9604	5475600	229320
2	72	3750	5184	14062500	270000
3	65	1890	4225	3572100	122850
4	71	2420	5041	5856400	171820
5	56	2830	3136	8008900	158480
6	93	3110	8649	9672100	289230
7	66	1970	4356	3880900	130020
8	82	2480	6724	6150400	203360
9	89	2860	7921	8179600	254540
10	97	3333	9409	11108889	323301
Σ	789	26983	64249	75967389	2152921
Ø	78,90	2698,30	//	//	//

Im zweiten Schritt werden die Zahlen aus der Hilfstabelle in die entsprechende Formel eingesetzt:

$$r = \frac{\sum_{i=1}^n (x_i * y_i) - n * \bar{x} * \bar{y}}{\sqrt{\sum_{i=1}^n (x_i^2) - n * \bar{x}^2} * \sqrt{\sum_{i=1}^n (y_i^2) - n * \bar{y}^2}} = \frac{2152921 - 10 * 78,90 * 2698,30}{\sqrt{64249 - 10 * 78,90^2} * \sqrt{75967389 - 10 * 2698,30^2}} = 0,3017$$

Im dritten Schritt ist das Ergebnis von r = 0,3 nun zu interpretieren. Der Wert des Korrelationskoeffizienten deutet auf eine schwach-positive Korrelation hin – es existiert also offenkundig kein linearer

Zusammenhang zwischen den Variablen. Dass dennoch ein Zusammenhang – beispielsweise ein monotoner – bestehen könnte, zeigt der Blick auf das Streudiagramm (in diesem Fall erstellt mit SSP).



Tendenziell scheinen hier – von drei aus der Reihe fallenden Werten abgesehen – große Werte von x nämlich durchaus mit großen Werten von y einherzugehen (und umgekehrt). Ein linearer Zusammenhang ist allerdings bereits rein visuell betrachtet maximal schwach vorhanden. Eventuell würde sich an dieser Stelle also die Berechnung des Rangkorrelationskoeffizienten nach Spearman oder des Konkordanzkoeffizienten nach Kendall lohnen, die beide monotone Zusammenhänge abbilden.

Zur inhaltlichen Interpretation sei an dieser Stelle noch einmal festgestellt, dass die Aussage „Das Nettogehalt wird durch die Studienleistung beeinflusst“ auch bei einer starken linearen Korrelation falsch gewesen wäre, da der Bravais-Pearson-Korrelationskoeffizient keinerlei Aufschluss über die Wirkungsrichtung eines möglichen inhaltlichen Zusammenhangs gibt. Auch wenn die alternative Wirkungsrichtung – das Nettogehalt fünf Jahre nach dem Studium wirkt sich rückwirkend auf die Punktzahl der Abschlussklausur aus – rein logisch nicht in Frage kommt, wäre jede Interpretation, die sich inhaltlich so weit vorwagt, nicht mehr allein mit dem Bravais-Pearson-Ergebnis begründbar.

Auch in diesem Beispielfall kann (und sollte) übrigens nicht ausgeschlossen werden, dass es sich auch um eine rein zufällige Scheinkorrelation oder um eine indirekte Korrelation über eine dritte Variable z handeln könnte. Letzteres wäre im vorliegenden Fall sogar durchaus plausibel, da das Nettogehalt sich nach fünf Jahren sicherlich nicht mehr direkt an den Punkten der Abschlussklausur orientieren wird, beide Variablen aber über eine dritte Variable z (beispielsweise die Motivation oder die persönliche Leistungsfähigkeit) indirekt miteinander verbunden sein könnten.

STATISTIK I – Übung 09

Spearman und Kendall

1 Kurze Wiederholung

Was sind Rangkorrelationskoeffizienten?

Liegen keine metrisch skalierten Daten vor, ist die Berechnung des Bravais-Pearson-Korrelationskoeffizienten (welcher zudem einen linearen Zusammenhang voraussetzt) nicht möglich. In diesem Fall (sowie auch in Fällen, in denen in metrischen Daten kein linearer Zusammenhang zu vermuten ist), können alternativ der Rangkorrelationskoeffizient nach Spearman sowie der Konkordanzkoeffizient nach Kendall berechnet werden. Beide setzen lediglich ordinalskalierte Daten voraus und geben die Stärke des monotonen Zusammenhangs zwischen zwei Variablen wieder, wobei in gleichsinnige (große Werte der einen Variablen gehen tendenziell mit großen Werten der anderen Variablen einher) und gegensinnige (große Werte der einen Variablen gehen tendenziell mit kleinen Werten der anderen Variablen – und umgekehrt – einher) zu unterscheiden ist. Liegt ein linearer Zusammenhang vor, existiert stets auch ein monotoner Zusammenhang – umgekehrt kann aber durchaus ein monotoner Zusammenhang vorliegen, ohne dass auch ein linearer Zusammenhang existiert.

Der Grundgedanke hinter beiden Koeffizienten beruht auf dem Umstand, dass sich sowohl ordinale als auch metrische Daten in eine natürliche Reihenfolge bringen, d.h. ordnen lassen. Untersucht man den Grad der Korrelation zwischen zwei Reihen solcher Daten, kann man somit eine der Datenreihen ordnen und anschließend prüfen, inwiefern sich die zweite Datenreihe „mitgeordnet“ hat. Liegt ein perfekter gleichsinniger Zusammenhang vor, so ist zu erwarten, dass sich die zweite Datenreihe perfekt mitordnet, d.h. (je nach gewählter Sortierung) entweder vom kleinsten zum größten oder vom größten zum kleinsten Datensatz. Im Falle eines perfekt gegensinnigen Zusammenhangs wäre dagegen zu erwarten, dass sich die zweite Datenreihe exakt entgegengesetzt zur ersten Datenreihe mitsortiert. Alle anderen Fälle weichen mehr oder weniger stark von diesen beiden Sonderfällen ab.

Mit den Koeffizienten nach Spearman und Kendall werden wir nachfolgend zwei Wege kennenlernen, mit denen wir den Grad der Abweichung in der Mitsortierung der zweiten Datenreihe von den beiden Sonderfällen der perfekten gleichsinnigen und der perfekten gegensinnigen Korrelation auf verschiedene Arten ermitteln, und in einer leicht interpretierbaren Kennzahl ausdrücken können.

Rangkorrelationskoeffizient nach Spearman

Zur Bestimmung des Rangkorrelationskoeffizienten nach Spearman werden beide Datenreihen der Größe nach mit Rängen beziffert. Über die Summe der – im Falle eines perfekten gleichsinnigen Zusammenhangs nicht vorhandenen – quadrierten Differenzen zwischen den Rängen der ersten und der zweiten Datenreihe wird anschließend der Korrelationskoeffizient wie folgt berechnet:

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n}$$

Die Vorgehensweise lässt sich am schnellsten anhand eines einfachen Beispiels erschließen:

Werte von x	Ränge von x	Werte von y	Ränge von y	Differenzen d	d ²
30	3	90	4	-1	1
20	2	70	2	0	0
40	4	100	5	-1	1
10	1	60	1	0	0
50	5	80	3	2	4

Als Summe der quadrierten Rangdifferenzen ergibt sich hier also 6. Eingesetzt (gemeinsam mit $n = 5$) in die Formel für den Rangkorrelationskoeffizienten nach Spearman ergibt sich folgender Wert:

$$rho = 1 - \frac{6 * \sum d_i^2}{(n^2 - 1) * n} = 1 - \frac{6 * 6}{(5^2 - 1) * 5} = 1 - \frac{36}{120} = 1 - 0,3 = 0,7$$

Der Wert für „Spearman's Rho“ liegt stets zwischen -1 und +1 und ist wie folgt zu interpretieren:

- > 0 -> gleichsinniger monotoner Zusammenhang
- ≈ 0 -> kein monotoner Zusammenhang
- < 0 -> gegensinniger monotoner Zusammenhang

Ein Problem für die Aussagekraft des Koeffizienten stellt das Auftreten sogenannter verbundener Ränge (auch Rangplatzbindungen oder ties) dar: Treten Werte mehrfach auf, bekommen diese einen Durchschnittsrang (berechnet als arithmetisches Mittel der zu vergebenden Ränge) zugewiesen.

Werte von x	Ränge von x
10	2
20	4,5
10	2
10	2
20	4,5

Je mehr verbundene Ränge existieren, umso geringer ist die Aussagekraft des Rangkorrelationskoeffizienten. Eine allgemeingültige Regel, ab welchem Anteil von verbundenen Rängen der Koeffizient nicht mehr berechnet werden sollte, existiert allerdings nicht, d.h. es bleibt eine subjektive Entscheidung. Eine Konstellation wie im obigen Beispiel dürfte allerdings in jedem Fall inakzeptabel sein.

Konkordanzkoeffizient nach Kendall

Auf dem gleichen gedanklichen Prinzip wie der Rangkorrelationskoeffizient nach Spearman basiert auch der Konkordanzkoeffizient nach Kendall. Für dessen Berechnung müssen ebenfalls zunächst die Ränge beider Variablenreihen gebildet werden. Anschließend werden die Daten nach den Rängen einer der beiden Datenreihen geordnet und wiederum überprüft, inwieweit sich die zweite Datenreihe „mitsortiert“ hat. Diese Überprüfung erfolgt anhand des paarweisen Vergleichs der Ränge der zweiten Datenreihe, wobei in konkordante (natürliche Reihenfolge der Ränge wird eingehalten) und diskordante Paare (natürliche Reihenfolge der Ränge wird nicht eingehalten) unterschieden wird.

Auch dieses Prinzip lässt sich am besten anhand des bereits bekannten Beispiels illustrieren:

Werte von x	Ränge von x	Werte von y	Ränge von y	Konkordante Paare	Diskordante Paare
10	1	60	1	4	0
20	2	70	2	3	0
30	3	90	4	1	1
40	4	100	5	0	1
50	5	80	3	//	//

In der ersten Zeile wird nun der y-Rang (1) mit den folgenden Rängen (2,4,5,3) verglichen. Dabei zeigt sich, dass auf die (1) nur größere Ränge folgen, die natürliche Ordnung also bei allen vier Paarvergleichen eingehalten wird. Für diese Zeile sind somit 4 konkordante und 0 diskordante Paare zu notieren.

Eine erste Abweichung zeigt sich in der dritten Zeile. Der Rang (4) ist hier mit den Rängen (5,3) zu vergleichen, wobei nur ein Paarvergleich konkordant ($4 < 5$), der andere jedoch diskordant ($4 > 3$) ausfällt. Insgesamt ergeben sich 8 konkordante (K) und 2 diskordante (D) Paare, die neben n (hier: 5) in die Formel für den Konkordanzkoeffizienten nach Kendall einzusetzen sind:

$$\tau = \frac{2 * (K - D)}{n * (n - 1)} = \frac{2 * (8 - 2)}{5 * (5 - 1)} = \frac{12}{20} = 0,6$$

Wie beim Rangkorrelationskoeffizienten nach Spearman ergibt sich auch hier ein Wert zwischen -1 und +1, der analog zu interpretieren ist. Obwohl beide Koeffizienten auf dem gleichen Prinzip des „Ordnen und Mitordnen“ basieren, fließen in den Konkordanzkoeffizienten nach Kendall weitaus mehr Informationen ein, da nicht nur einzelne Rangpaare (über die Differenzen), sondern alle Ränge miteinander abgeglichen werden (über die Ermittlung konkordanter und diskordanter Paare).

Ein Problem ergibt sich auch bei der Berechnung von Kendall durch verbundene Ränge, die in Paarvergleichen nicht mitgezählt werden. Im nachfolgenden, bereits bekannten Beispiel, ergeben sich etwa in 10 Paarvergleichen nur 6 Werte, da 4 Paarvergleiche mit identischen Werten ausscheiden.

Werte von x	Ränge von x	Konkordante Paare	Diskordante Paare
10	2	2	0
20	4,5	0	2
10	2	1	0
10	2	1	0
20	4,5	//	//

Korrelation und Kausalität

Zur korrekten inhaltlichen Interpretation von Korrelationen wurde in den vorangegangenen Übungen zum Bravais-Pearson-Korrelationskoeffizienten bereits einiges geschrieben, das auch für die Interpretation des Rangkorrelationskoeffizienten nach Spearman sowie des Konkordanzkoeffizienten nach Kendall Gültigkeit besitzt und bei der Bearbeitung entsprechender Aufgaben beachtet werden sollte.

2 Beispielrechnungen

Ein Kinobetreiber verändert (*ceteris paribus*) an 10 Tagen die Eintrittspreise für den Abendfilm und zeichnet die Besucherzahlen auf. Es ergibt sich die folgende Tabelle:

Abend Nr.	Preis pro Karte (in Euro)	Besucher/innen pro Abend
1	8,50	836
2	9,00	741
3	9,50	798
4	10,00	740
5	10,50	685
6	11,00	673
7	11,50	710
8	12,00	625
9	12,50	531
10	13,00	330

Bestimmung des Rangkorrelationskoeffizienten nach Spearman

Nr.	x	rg(x)	y	rg(y)	d	d ²
1	8,50	1	836	10	-9	81
2	9,00	2	741	8	-6	36
3	9,50	3	798	9	-6	36
4	10,00	4	740	7	-3	9
5	10,50	5	685	5	0	0
6	11,00	6	673	4	2	4
7	11,50	7	710	6	1	1
8	12,00	8	625	3	5	25
9	12,50	9	531	2	7	49
10	13,00	10	330	1	9	81

$$\sum d^2 = 322$$

$$rho = 1 - \frac{6 \cdot \sum d_i^2}{(n^2 - 1) \cdot n} = 1 - \frac{6 \cdot 322}{(10^2 - 1) \cdot 10} = 1 - \frac{1932}{990} = -0,9515$$

Bestimmung des Konkordanzkoeffizienten nach Kendall

Nr.	x	rg(x)	y	rg(y)	K	D
1	8,50	1	836	10	0	9
2	9,00	2	741	8	1	7
3	9,50	3	798	9	0	7
4	10,00	4	740	7	0	6
5	10,50	5	685	5	1	4
6	11,00	6	673	4	1	3
7	11,50	7	710	6	0	3
8	12,00	8	625	3	0	2
9	12,50	9	531	2	0	1
10	13,00	10	330	1	//	//

$$tau = \frac{2 \cdot (K - D)}{n \cdot (n - 1)} = \frac{2 \cdot (3 - 42)}{10 \cdot (10 - 1)} = \frac{-78}{90} = -0,8667$$

3 Übungsaufgaben

Die Wiederholung des Kinopreis-Experiments in einem anderen Kino über die Dauer von 5 Tagen erbrachte folgende Ergebnisse:

Abend Nr.	Preis pro Karte (in Euro)	Besucher/innen pro Abend
1	8,50	527
2	9,00	489
3	9,50	512
4	10,00	315
5	10,50	308

- Berechnen Sie den Rangkorrelationskoeffizienten nach Spearman.
- Berechnen Sie den Konkordanzkoeffizienten nach Kendall.

STATISTIK I – Lösung 09**Spearman und Kendall**

Ein Kinobetreiber verändert (ceteris paribus) an 5 Tagen die Eintrittspreise für den Abendfilm und zeichnet die Besucherzahlen auf. Es ergibt sich die folgende Tabelle:

Abend Nr.	Preis pro Karte (in Euro)	Besucher/innen pro Abend
1	8,50	527
2	9,00	489
3	9,50	512
4	10,00	315
5	10,50	308

a) Berechnen Sie den Rangkorrelationskoeffizienten nach Spearman.

Nr.	x	rg(x)	y	rg(y)	d	d ²
1	8,50	1	527	5	-4	16
2	9,00	2	489	3	-1	1
3	9,50	3	512	4	-1	1
4	10,00	4	315	2	2	4
5	10,50	5	308	1	4	16

$$\sum d^2 = 38$$

$$n = 5$$

$$rho = 1 - \frac{6 \cdot \sum d_i^2}{(n^2 - 1) \cdot n} = 1 - \frac{6 \cdot 38}{(5^2 - 1) \cdot 5} = 1 - \frac{228}{120} = -0,9$$

b) Berechnen Sie den Konkordanzkoeffizienten nach Kendall.

Nr.	x	rg(x)	y	rg(y)	K	D
1	8,50	1	527	5	0	4
2	9,00	2	489	3	1	2
3	9,50	3	512	4	0	2
4	10,00	4	315	2	0	1
5	10,50	5	308	1	//	//

$$K = 1$$

$$D = 9$$

$$n = 5$$

$$tau = \frac{2 \cdot (K - D)}{n \cdot (n - 1)} = \frac{2 \cdot (1 - 9)}{5 \cdot (5 - 1)} = \frac{-16}{20} = -0,8$$