3

Kennwerte empirischer Häufigkeitsverteilungen

Im letzten Kapitel haben wir die Daten tabellarisch oder grafisch dargestellt und dabei auch den Begriff der *Häufigkeit* kennengelernt und *Häufigkeitstabellen* zusammengestellt. Wir wollen uns nun die Daten in so einer Häufigkeitstabelle etwas näher anschauen und insbesondere etwas über die Art und Weise sagen, wie die Daten *verteilt* sind; auf Seite 31 haben wir dafür bereits den Begriff **Häufigkeitsverteilung** verwendet. Mit der etwas genaueren Bezeichnung als *empirische*[1] *Häufigkeitsverteilung* wollen wir betonen, dass es sich bei unseren Daten und ihrer Verteilung um beobachtete oder gemessene Werte (einer Stichprobe) handelt und nicht um eine «theoretische» Verteilung, nach der wir die Daten *modellieren*[2].

Wie können wir nun die Daten und ihre Verteilung charakterisieren und durch aussagekräftige Kennwerte zusammenfassen?

Es gibt eine Reihe von statistischen Kennwerten, mit denen wir die Verteilung der Daten einer Stichprobe beschreiben können. Die bekanntesten sind das *arithmetische Mittel*, mit dem wir alle Daten einer Stichprobe durch einen einzigen Wert repräsentieren wollen (nämlich jenen, der «im Zentrum» der Daten steht), sowie die *Standardabweichung*, mit der wir angeben, wie stark die einzelnen Werte im Schnitt von diesem Mittel (und voneinander) abweichen. Statistiker:innen sagen, das arithmetische Mittel beschreibt die *Lage* einer Verteilung, die Standardabweichung ihre *Streuung*. Es gibt noch weitere Lage- und Streuungsmaße, und einige davon wollen wir hier angeben.

Zuvor können wir auch noch jedem beobachteten oder gemessenen Wert eine *Rangzahl* zuordnen:

[1] zum griech. $\varepsilon\mu\pi\varepsilon\iota\rho\omega\varsigma$ (empeiros): etwas aus der Erfahrung kennen

[2] Solche theoretischen Modelle werden wir in einem späteren Kapitel auch noch kennen lernen.

Rangzahl

Bevor wir zusammenfassende Kennwerte bilden, ist es manchmal für einen ersten Überblick sinnvoll, die Daten entsprechend der Größe ihrer Merkmalswerte zu sortieren und entsprechend des Ranges, den sie dabei einnehmen, zu indizieren. Üblicherweise wird dabei mit dem kleinsten Wert begonnen und dieser mit x_1 bezeichnet. Sind zwei (oder mehr) Daten gleich groß, erhalten sie nicht denselben Rang (es gibt also keine Ex aequo-Plätze), sondern der Index wird einfach weiter fortlaufend gezählt[3].

Damit Daten in eine Rangordnung gebracht werden können, müssen sie nicht unbedingt numerisch, aber zumindest Ordinaldaten sein.

Achtung bei Daten, bei denen die Chronologie eine Rolle spielt, also die zeitliche Reihenfolge, in der sie aufgetreten sind: Diese dürfen natürlich nicht der Größe nach geordnet werden, sondern müssen ihre ursprüngliche Abfolge beibehalten und werden auch in dieser Reihenfolge indiziert.

3.1 Lagekennwerte empirischer Häufigkeitsverteilungen

Minimaler und maximaler Wert

Man kann für jede Stichprobe, in der die Elemente zumindest ordinalskaliert sind, einen **Maximalwert** x_{max} und einen **Minimalwert** x_{min} angeben. Sind die Daten entsprechend ihrer Rangzahl indiziert, so ist

$$x_{\min} = x_1 \tag{3.1}$$

$$x_{\max} = x_n \tag{3.2}$$

Beispiel 3 *Gegeben ist die Kaffeemenge (Tassen pro Tag), die von einer Testperson innerhalb von vierzehn Tagen während des Lesens und Durcharbeitens dieser Unterlagen konsumiert wurde. Gesucht sind der größte und kleinste Wert.*

[3] Diese Vorgangsweise gilt nicht bei der Berechnung der Rangkorrelation, siehe Seite 88

Sowohl in *MS Excel* als auch in *LibreOffice Calc* können wir das Minimum mit dem Befehl `=MIN(Zahl1;Zahl2;...)` berechnen, wobei die Argumente `(Zahl1;Zahl2;...)` die Zahlen sind, von denen wir den kleinsten Wert wissen wollen.
Der entsprechende Befehl für das Maximum lautet in beiden Programmen: `=MAX(Zahl1;Zahl2;...)`
In *R* lauten die Funktionen `min(x)` bzw. `max(x)`, wobei als Argument (also für `x`) der Vektor eingesetzt wird, der die Daten enthält, deren Minimum bzw. Maximum wir ausrechnen wollen[a].

[a] Zur Syntax in *R* siehe die Lehrveranstaltung *PR122: Einführung in die Programmierung*, oder auch direkt in *R* durch Aufruf der «Hilfe» zu den einzelnen Funktionen, z.B. mit `?max()` (Also Eingabe eines Fragezeichens und des Funktionsnamens, ohne Argumente in den Klammern)

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

Der Einfachheit halber ordnen wir die Daten unserer Stichprobe der Größe nach:

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

Damit lassen sich Minimum und Maximum ganz leicht angeben:

$$x_{\min} = 1$$
$$x_{\max} = 6$$

Um das Beispiel in R zu rechnen, müssen wir zunächst einen Vektor bilden, der die Ausgangsdaten enthält. Wir geben ihm den Namen `bsp3`:

```
bsp3 <- c(1, 3, 1, 3, 2, 2, 5, 4, 3, 2, 3, 4, 6, 3)
```

Dann können wir das Minimum und Maximum ausrechnen und erhalten:

```
min(bsp3)
[1] 1
max(bsp3)
[1] 6
```

Modalwert

Der **Modalwert** (auch *Modus* genannt) ist jener Wert, der in einer Stichprobe am häufigsten vorkommt. In der Stichprobe $(1, 1, 3, 5, 6, 6, 6)$ ist zum Beispiel der Modalwert 6, weil der 6er dreimal vorkommt und keine andere Zahl an diese Häufigkeit herankommt.

Es kann auch mehr als einen Modalwert geben[4]. In $(1, 1, 1, 1, 3, 5, 5, 5, 5, 6)$ zum Beispiel gibt es zwei Modalwerte: 1 und 5. Beide kommen viermal vor.

Gibt es nur einen einzigen Modalwert, so spricht man auch von einer *unimodalen* Verteilung und bezeichnet den Modalwert selbst als *häufigsten Wert* oder auch als *wahrscheinlichsten Wert*: Wenn wir uns das Beispiel der Daten aus Abb.2.6 ansehen und uns 2019 ein Freund erzählt hätte, dass er gestern einen in Österreich produzierten Fisch gegessen hat, dann war das am wahrscheinlichsten eine Regenbogen- oder Lachsforelle – denn das ist der Modalwert dieser Verteilung.

Modalwerte können wir sowohl für qualitative als auch für quantitative Daten angeben. Er ist für alle Skalenniveaus möglich.

Beispiel 4 *Die folgende Tabelle enthält Ortsnamen, die in Österreich mehrfach vorkommen. Welches ist der häufigste Wert?*

Name	Anzahl	Name	Anzahl
Au	57	Aigen	37
Berg	41	Grub	50
Hart	38	Hof	31
Moos	40	Reith	36
Straß	31	Winkl	30

Tabelle 3.1: Die beliebtesten Ortsnamen Österreichs

Bei diesem einfachen Beispiel können wir das Ergebnis gleich durch einen Blick auf die Tabelle herauslesen. Wir könnten die Häufigkeiten auch in einem Diagramm darstellen und dann schauen, welches die höchste Säule ist. In jedem Fall kommen wir zum Ergebnis: Der häufigste in Österreich vorkommende Ortsname ist Au *(nämlich* $57\times$*).*

*(**Hinweis:** Beachte, dass der Modalwert in diesem Beispiel* Au *ist, und nicht 57.)*

Aufgabe 1 Gib zur Stichprobe des Beispiels 3 den oder die Modalwert(e) an.

[4]Wenn man die Bezeichnung *Modus* bevorzugt heißt die Mehrzahl: *Modi*.

Der *Excel*-Befehl für den Modalwert lautet
`=MODUS.VIELF((Zahl1;Zahl2;...)`. Gibt es mehrere Modalwerte, dann werden mit diesem Befehl auch alle zurückgegeben – das macht es aber notwendig, dass die Funktion als so genannte «Arrayformel» eingegeben werden (Siehe dazu die Hilfe-Funktion von Excel).
LibreOffice Calc hat einen ähnlichen Befehl:
`=MODALWERT(Zahl1;Zahl2;...)`. Gibt es hier mehrere Modalwerte, wird nur der kleinste von ihnen zurückgegeben und die anderen ignoriert!
Sowohl *Excel* als auch *Calc* setzen voraus, dass mindestens ein Wert der Stichprobe mindestens zweimal vorkommt, ansonsten gibt es eine Fehlermeldung.
In *R* gibt es keinen vordefinierten Befehl, um den Modalwert auszurechnen.

Mittelwerte

Der **arithmetische Mittelwert** (auch: das *arithmetische Mittel*) ist ein sehr gebräuchliches Maß für die Angabe des «Durchschnitts» der Verteilung von numerischen Daten. Dabei wird sprachlich die Spezifizierung «arithmetisch»[5] auch meist weggelassen und nur vom *Mittelwert* oder *Mittel* gesprochen.

Mathematisch ist das arithmetische Mittel der Quotient der Summe der Beobachtungswerte dividiert durch die Anzahl der Beobachtungswerte:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.3}$$

In *Excel* und auch in *LibreOffice Calc* wird das arithmetische Mittel mit dem Befehl `=MITTELWERT(Zahl1;Zahl2;...)` berechnet.
In *R* lautet der Befehl `mean(x)`.

Beispiel 5 *Das arithmetische Mittel der Daten aus Beispiel 3 beträgt:*

$$\bar{x} = \frac{1+3+1+3+2+2+5+4+3+2+3+4+6+3}{14} = \frac{42}{14} = 3$$

was wir auch aus R erhalten:

[5] griech. $\alpha\rho\iota\theta\mu\eta\tau\iota\kappa o\varsigma$ (arithmetikos) = im Zählen oder Rechnen geschickt

```
mean(bsp3)
[1] 3
```

Liegen die Daten in Form einer Häufigkeitstabelle vor, so erhalten wir den arithmetischen Mittelwert aus:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{m} \left(f_j \cdot x_j \right)$$

mit: (3.4)

$m \ldots$ Anzahl der unterschiedlichen Merkmalswerte

$f_j \ldots$ Häufigkeit des Auftretens dieses Merkmalswertes

Beispiel 6 *Die Daten aus Beispiel 3, in einer Häufigkeitstabelle angegeben:*

x	f	$f \cdot x$
1	2	2
2	3	6
3	5	15
4	2	8
5	1	5
6	1	6
	$n = 14$	$\Sigma = 42$

Aus n und der Summe rechts unten können wir nach Formel 3.4 das arithmetische Mittel berechnen:

$$\bar{x} = \frac{42}{14} = \underline{\underline{3}}$$

Wenn wir unsere Daten in Klassen eingeteilt haben, ist die Bildung des arithmetischen Mittels nicht mehr so einfach, weil wir ja die einzelnen Merkmalswerte, die wir für die Mittelwertsberechnung benötigen, nicht mehr zur Verfügung haben. Wir verwenden dann für das arithmetische Mittel die *Klassenmitten* als Eingangswerte in Formel 3.4. Die Klassenmitten sind jene Werte, die genau in der Mitte zwischen oberer und unterer Klassengrenze liegen.

Das arithmetische Mittel lässt sich dann nach 3.5 berechnen:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{m} \left(f_j \cdot x'_j \right)$$

mit: (3.5)

$m \ldots$ Anzahl der Klassen

$f_j \ldots$ Häufigkeit der Elemente in der j-ten Klasse

$x'_j \ldots$ Klassenmitte der j-ten Klasse

Beispiel 7 *Zur Berechnung des arithmetischen Mittels der Stichprobe aus Tabelle 2.2 ergänzen wir noch die jeweiligen Klassenmitten. Dabei unterstellen wir der letzten, an sich offenen Klasse, die gleiche Breite wie allen anderen:*

j	Klassengrenzen	Klassenmitte x'_j	f_j	$f \cdot x'_j$
1	$160 \leq x < 165$	162.5	1	162.5
2	$165 \leq x < 170$	167.5	3	502.5
3	$170 \leq x < 175$	172.5	4	690.0
4	$175 \leq x < 180$	177.5	8	1 420.0
5	$180 \leq x < 185$	182.5	3	547.5
6	$185 \leq x < 190$	187.5	4	750.0
7	$190 \leq x < 195$	192.5	0	0.0
8	$195 \leq x$	197.5	1	197.5
			$n = 24$	$\Sigma = 4\,270.0$

Tabelle 3.2: Häufigkeitstabelle einer Stichprobe mit klassifizierten Merkmalswerten

Das arithmetische Mittel ist dann

$$\bar{x} = \frac{4\,270}{24} = \underline{\underline{177.9}}$$

Hätten wir in obigem Beispiel nicht nur die klassierten Werte, sondern die Originaldaten zur Verfügung, würde vermutlich nicht genau 177.9 herauskommen. Aber das stört uns als Statistiker:in nicht wirklich. Statistik ist nicht Buchhaltung. Es geht eher darum, Modelle zu finden, wie die (messbare) Welt um uns herum *vermutlich* aussieht[6].

Neben dem arithmetischen Mittel sind in unseren Anwendungen manchmal auch das *gewichtete arithmetische Mittel* und das *geometrische Mittel* von Bedeutung:

Das **gewichtete arithmetische Mittel** wird verwendet, wenn Werte mit unterschiedlicher «Wichtigkeit» in die Berechnung des Mittels einfließen soll. Du erhältst dann unterschiedliche *Gewichte*, also Faktoren, mit denen sie multipliziert werden. Anwendung findet das zum Beispiel, wenn wir drei Noten haben (Bachelorarbeit 1, Bachelorarbeit 2 und Bachelorprüfung), für eine Gesamtbeurteilung diese drei Noten aber prozentuell nicht gleich einfließen sollen, sondern zum Beispiel im Verhältnis 20 : 20 : 60. Für die Berechnung des gewichteten arithmetischen Mittels wird dann jede Note vor dem Addieren mit ihrem Gewicht multipliziert (also mit 20 oder 60). Dividiert wird dann nicht durch die

[6] Daher verwenden wir im Zusammenhang mit Zufallsvariablen manchmal auch das Fremdwort *Stochastik*, aus dem griech. στοχαστικός (stochastikos) = *im Vermuten geschickt*

Anzahl der Elemente (in unserem Beispiel also nicht durch 3), sondern durch die Summe der Gewichte (also durch 100).

Etwas allgemeiner können wir angeben[7]:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i} \tag{3.6}$$

In *Excel* und *LibreOffice Calc* kann man ein gewichtetes arithmetisches Mittel mithilfe der Befehle `SUMMENPRODUKT` und `SUMME` errechnen (indem man die Summe aus den Produkten zwischen den Einzelwerten und ihren Gewichten durch die Summe der Gewichte dividiert).
In *R* gibt es einen direkten Befehl dafür: `weighted.mean(x)`

Das **geometrische Mittel** benötigt man zur Mittelung von Steigerungsfaktoren, zum Beispiel wenn sich bei der Zinseszinsrechnung der Zinsfaktor jährlich ändert und du einen durchschnittlichen Zinsfaktor angeben willst.

Das geometrische Mittel aus n Zahlen ist die n-te Wurzel des Produkts der n Zahlen:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{3.7}$$

Dabei ist zu beachten, dass das geometrische Mittel nur für positive Zahlen definiert ist. Es müssen also alle Elemente der Stichprobe durch numerische Werte ≥ 0 repräsentiert sein.

Der Umgang mit dem geometrischen Mittel ist nicht immer einfach, wie wir an folgendem Beispiel sehen.

Beispiel 8 *(Wir rechnen dieses Beispiel mit Absicht mit etwas übertriebenen, realitätsfernen Zahlen, damit wir den mathematischen Effekt sehen):*

Nehmen wir an, jemand legt 1 000 EUR in einer Veranlagungsform an, die im ersten Jahr 40% *Zinsen erbringt. Im zweiten Jahr gibt es nur noch* 10%*. Welcher durchschnittlichen jährlichen Verzinsung entspricht das?*

[7] Das w in obiger Formel steht für das englische Wort *weight* = Gewicht

Zunächst einmal rechnen wir uns das Endkapital aus:
Im ersten Jahr ergibt die Verzinsung ein Kapital von $1\,000 \cdot 1.4 = 1\,400$.
Im zweiten Jahr dann: $1\,400 \cdot 1.1 = 1\,540$ EUR.

Der durchschnittliche Zinsfaktor beträgt dann entsprechend Formel 3.7

$$\bar{x}_g = \sqrt[2]{1.4 \cdot 1.1} = \underline{\underline{1.24097}}$$

Zur Probe wenden wir jetzt jedes Jahr diesen durchschnittlichen Zinsfaktor an und erhalten tatsächlich:

$$(1\,000 \cdot 1.24097) \cdot 1.24097 = 1.540 \; EUR$$

(Hinweis: *Den genauen Wert erhältst du nur, wenn du statt des gerundeten Wertes* 1.24097 *mit dem genauen Wert* 1.240967365 . . . *rechnest, der beim Wurzelziehen herauskommt).*

Hier jetzt noch zwei Möglichkeiten, wie man dieses Beispiel ***falsch*** *rechnen könnte:*

Wenn du statt des geometrischen Mittels das arithmetische Mittel verwendest, erhältst du $\frac{1.4+1.1}{2} = 1.25$ *und damit dann:* $(1000 \cdot 1.25) \cdot 1.25 = 1562.5$ EUR.

Und wenn du statt der Zinsfaktoren 1.4 *und* 1.1 *die Zinssätze* 40% *und* 10% *in die Formel einsetzt (also die Zahlenwerte* 0.4 *und* 0.1*), erhältst du* $\bar{x}_g = \sqrt{0.4 \cdot 0.1} = 0.2$, *was überhaupt nur ein Endkapital von* 1.440 EUR *ergäbe.*

Du musst also in die Formel für das geometrische Mittel immer den ***Veränderungsfaktor*** *einsetzen, nicht den prozentuellen Veränderungswert!*

In *Excel* und auch in *LibreOffice Calc* wird das geometrische Mittel mit dem Befehl `=GEOMITTEL(Zahl1;Zahl2;...)` berechnet.
In *R* gibt es zwar keine vordefinierte Funktion dafür, wir können aber mathematisch ein wenig «tricksen» und das geometrische Mittel der Daten, die sich in `x` befinden, mit `exp(mean(log(x)))` ausrechnen.

Im Übrigen gilt: Der Wert des geometrischen Mittels ist immer kleiner als der Wert des arithmetischen Mittels derselben Werte[8].

[8] Für zwei positive Zahlen x und y lässt sich das relativ einfach zeigen, für den allgemeinen Fall von n Zahlen ist aber ungleich komplizierter und wir werden das an dieser Stelle nicht *beweisen*. Wir verlassen uns einfach darauf, dass Mathematiker:innen hier ganze Arbeit geleistet haben, siehe zum Beispiel de.wikipedia.org/wiki/Ungleichung_vom_arithmetischen_und_geometrischen_Mittel

Das letzte Mittelmaß, das wir noch anschauen wollen, ist das **harmonische Mittel**. Es wird verwendet, um den Mittelwert von Verhältniszahlen zu berechnen:

$$\bar{x}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \tag{3.8}$$

Ein einfaches Beispiel für die Verwendung des harmonischen Mittels ist die Berechnung der durchschnittlichen Geschwindigkeit für den Hin- und Rückweg einer bestimmten Strecke. Geschwindigkeiten sind ja Verhältniszahlen (nämlich das Verhältnis Weg:Zeit). Fährt man zum Beispiel mit 100 km/h von Scheibbs nach Hamburg (Entfernung = 1000 km), aber mit nur 80 km/h retour, so ist man den Gesamtweg von 2000 km nicht mit durchschnittlich $(100 + 80)/2 = 90$ km/h gefahren, sondern mit $88,89$ km/h, also jenen Wert, der dem harmonischen Mittel von 100 und 80 entspricht[9].

Ein weiteres Beispiel ist der so genannte *F1-Score*, der für die Beurteilung der Leistungsfähigkeit eines Large Language Modells verwendet wird. Der F1-Score gibt ein mittleres Maß für die beiden Kennzahlen *Precision* und *Recall* an. Diese beiden Kennzahlen sind selbst Verhältniszahlen, daher muss für den Mittelwert das harmonische Mittel verwendet werden.

In *Excel* und in *LibreOffice Calc* wird das harmonische Mittel mit dem Befehl `=HARMITTEL(Zahl1;Zahl2;...)` berechnet.
In *R* gibt es wieder keine vordefinierte Funktion dafür. Wir können uns aber zunutze machen, dass das harmonische Mittel aus mehreren Zahlen der Kehrwert des arithmetischen Mittels der Kehrwerte dieser Zahlen ist (OK, wer diesen Satz jetzt nicht verstanden hat, wendet einfach Formel 3.8 an...).

Beispiel 9 *Bei der Beurteilung der Leistungsfähigkeit von KI-Systemen lässt man das System eine bestimmte Anzahl von Vorhersagen machen, von denen man selbst die richtige Antwort kennt. Angenommen, es gibt nur zwei Antwortmöglichkeiten, zum Beispiel: Dieses Bild zeigt eine Katze oder nicht. Dann zählt man mit:*

$TP =$ True positive: Eine Katze wird richtigerweise als Katze erkannt
$TN =$ True negative: Es wird richtigerweise erkannt, dass keine Katze zu sehen ist
$FP =$ False positive: Es wird eine Katze erkannt, obwohl keine zu sehen ist
$FN =$ False negative: Es wird keine Katze erkannt, obwohl eine zu sehen ist

[9] Was sich leicht nachprüfen lässt: Für den Hinweg benötigt man 10 Stunden, für den Rückweg 12,5 Stunden, insgesamt also für 2000 km 22,5 Stunden, was $2000/22,5 = 88,89$ km/h ergibt.

Damit können folgende Kennwerte berechnet werden:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \text{harmonisches Mittel aus Precision und Recall}$$

Sei: $TP = 642$, $TN = 279$, $FP = 277$, $FN = 85$.

Welchen Wert hat der F1-Score?

$$\text{Precision} = \frac{642}{642 + 277} = 0,699$$

$$\text{Recall} = \frac{642}{642 + 85} = 0,883$$

$$\text{F1-Score} = \bar{x}_h = \frac{2}{\frac{1}{0,699} + \frac{1}{0,883}} = \underline{\underline{0,780}}$$

Quantile

Mittelwerte sind nicht das einzige Maß, die sich zur Angabe eines «Durchschnitts» eignen. Wir können unsere Daten auch der Größe nach ordnen (falls wir das nicht schon ohnehin gemacht haben) und die geordnete Beobachtungsreihe in zwei Teile zerlegen. Gesucht ist jener Wert, der definiert, wo wir die Stichprobe «durchschneiden» müssen, damit ein bestimmter Anteil der Beobachtungen unterhalb dieser Trennlinie liegt.

Beispiel 10 *Die schwedische Schirennläuferin Frida Hansdotter ist in ihrer Karriere (2005 - 2019) bei insgesamt 130 Slalom-Rennen im Weltcup oder bei Olympischen Spielen oder Weltmeisterschaften gestartet. Dafür hat sie eine Gesamtzeit von 3 Stunden 24 Minuten und 41,20 Sekunden benötigt. (Das inkludiert auch die Zeit der 20 Rennen, bei denen sie nur einen Durchgang ins Ziel geschafft hat (in 12 hat sie sich nicht für den 2. Durchgang qualifiziert, in 6 ist sie im 2. Durchgang ausgeschieden), nicht aber die 8 Rennen, bei denen sie bereits im 1. Durchgang ausgeschieden ist).*
In der folgenden Tabelle sind die Platzierungen der bei den letzten 50 von ihr in Angriff

date	place	position	date	place	position
16.02.2018	Pyeongchang	1	28.11.2015	Aspen	3
10.01.2017	Flachau	1	13.12.2014	Are	3
29.12.2015	Lienz	1	02.02.2019	Maribor	4
13.01.2015	Flachau	1	05.01.2019	Zagreb	4
28.01.2018	Lenzerheide	2	17.11.2018	Levi	4
07.01.2018	Kranjska Gora	2	11.11.2017	Levi	4
15.11.2017	Levi	2	03.01.2017	Zagreb	4
15.01.2016	Flachau	2	11.12.2016	Sestriere	4
13.12.2015	Are	2	04.01.2015	Zagreb	4
29.11.2015	Aspen	2	29.12.2014	Kühtai	4
21.03.2015	Meribel	2	16.03.2019	Soldeu	5
14.02.2015	Beaver Creek	2	16.02.2019	Are	5
30.11.2014	Aspen	2	08.01.2019	Flachau	5
22.12.2018	Courchevel	3	26.11.2017	Killington	5
25.11.2018	Killington	3	05.01.2016	Santa Caterina	5
17.03.2018	Are	3	15.02.2016	Crans Montana	6
10.03.2018	Ofterschwang	3	14.03.2015	Are	6
09.01.2018	Flachau	3	09.03.2019	Spindleruv Mlyn	7
03.01.2018	Zagreb	3	29.12.2016	Semmering	7
28.12.2017	Lienz	3	22.02.2015	Maribor	9
18.03.2017	Aspen	3	27.11.2016	Killington	10
18.02.2017	St. Moritz	3	06.03.2016	Jasna	10
08.01.2017	Maribor	3	29.12.2018	Semmering	DNF2
19.03.2016	St. Moritz	3	11.03.2017	Squaw Valley	DNQ2
12.01.2016	Flachau	3	12.11.2016	Levi	DNF2

Tabelle 3.3: Frida Hansdotters Platzierungen in ihren letzten 50 Slaloms im Weltcup, bei Weltmeisterschaften oder Olympischen Spielen. Für alle im folgenden durchgeführten Berechnungen, die sich auf diese Tabelle beziehen, werden wir nur $n = 47$ Werte berücksichtigen und DNQ2 und die beiden DNF2 weglassen.

genommenen Rennen gegeben. Die Daten sind nicht chronologisch sondern der Größe nach geordnet, Ordnungskriterium ist die Platzierung.

Wir können nun aus dieser geordneten Liste zum Beispiel feststellen, dass Frida Hansdotter in ihren 27 besten Rennen (der Saisonen 2014/15 bis 2018/19) nie schlechter als Dritte geworden ist oder in 40 aus 50 Rennen nie schlechter als Fünfte. In 94% der Rennen hat sie mindestens den 10. Platz erreicht (in den restlichen 3 Rennen gar keine Platzierung).

Etwas systematischer und «statistischer» betrachtet besteht unsere Aufgabe darin, eine geordnete Beobachtungsreihe so in zwei Teile zu zerlegen, dass ein bestimmter Prozentsatz der Daten vom Rest getrennt wird. Wir nennen die Stelle, an der wir diesen Trennstrich einziehen, das **p-Quantil** der Verteilung, wobei p den anteilmäßigen Umfang der abgeteilten Daten angibt.

p kann zwischen 0 und 1 liegen (bzw. zwischen 0% und 100%).

Das p-Quantil ist nun definiert als jener Wert, für den gilt: $(n \cdot p)$ aller Stichprobenelemente sind kleiner oder gleich dem p-Quantil und $(n \cdot (1-p))$ aller Elemente größer als das p-Quantil. Haben wir zum Beispiel 200 der Größe nach geordnete Daten und suchen das Quantil für $p = 0.25$, dann ist das jener Wert, bei dem unsere gesamte Stichprobe so in zwei Teile geteilt wird, dass 50 Werte unterhalb des Quantils liegen, und 150 darüber.

Wie finden wir nun den konkreten Wert des p-Quantils einer empirischen Häufigkeitsverteilung?

Um es vorweg zu nehmen: Die Bestimmung des exakten Wert ist ein wenig kompliziert: Zunächst müssen wir die zum p-Quantil zugehörige Rangzahl bestimmen:

$$i_p = p(n-1) + 1 \tag{3.9}$$

Der Wert, der an der i_p.ten Stelle liegt, ist dann das gesuchte Quantil.

Wenn i_p keine ganze Zahl ist (was ziemlich oft vorkommt), muss zwischen den Werten an der Stelle $int(i_p)$ und $(int(i_p)+1)$ linear interpoliert werden:

$$x_p = x_{int(i_p)} + (i_p - int(i_p))(x_{int(i_p)+1} - x_{int(i_p)}) \tag{3.10}$$

wobei die Funktion *int* die *Integer-Funktion* ist, das ist jene Funktion, die nur den ganzzahligen Anteil einer Zahl zurückgibt.

Zum Glück hat *Excel* die beschriebene Prozedur zur Berechnung des Quantils eingebaut: Mit der Funktion `=QUANTIL.INKL(Array;p)` (mit den Daten der Stichprobe im Datenbereich `Array`) kann es ziemlich einfach berechnet werden, ohne sich über ein i_p, eine Interpolation oder die Integer-Funktion Gedanken machen zu müssen.
In *LibreOffice Calc* heißt der Befehl: `=QUANTIL(Daten;p)`.
In *R*: `quantile(x,p)`

Beispiel 11 *Aus den Daten der Tabelle 3.3 können wir ausrechnen:* $x_{0.25} = 2$, $x_{0.5} = 3$ *und* $x_{0.75} = 4.5$

Näherungsverfahren zur Berechnung von Quantilen

Die oben beschriebene, etwas komplizierte Prozedur werden wir nur verwenden, wenn wir ein (Rechen-)Programm verwenden können. Für die Bestimmung

des Quantils «von Hand» reicht es, wenn wir folgendes **Näherungsverfahren** verwenden:

Um das p-Quantil einer Stichprobe mit n Elementen näherungsweise auszurechnen, multiplizieren wir zunächst

$$k = n \cdot p \tag{3.11}$$

Wenn k eine ganze Zahl ist, dann ist unser p-Quantil der Mittelwert zwischen x_k und dem nächsten Beobachtungswert x_{k+1}:

$$x_p = \frac{x_k + x_{k+1}}{2} \tag{3.12}$$

Wenn k nicht ganzzahlig ist, dann runden wir es auf die nächste ganze Zahl auf. Der Wert, der an dieser Stelle steht, ist dann das gesuchte p-Quantil:

$$\begin{aligned} &x_p = x_{\lceil k \rceil} \\ &\text{mit: } \lceil k \rceil = \text{kleinste ganze Zahl größer oder gleich } k \end{aligned} \tag{3.13}$$

Wir wollen das nun in einem Beispiel für einige p-Werte durchrechnen:

Gegeben sind in Tabelle 3.4 für alle[10] Nachbarländer der Nachbarländer Österreichs die Anzahl der Ausbildungsjahre, die ein Kind im Schuleintrittsalter in diesem Land im Schnitt vor sich hat. Die Werte wurden auf halbe Jahre gerundet. Die Tabelle ist bereits der Größe nach geordnet.

Beispiel 12 *Gib für die Daten der Tabelle 3.4 das Quantil für $p = 25\%$ an.*

$$k = n \cdot p = 20 \cdot 0.25 = 5$$

Das ist eine ganze Zahl, daher müssen wir den Mittelwert aus x_5 und x_6 bilden:

$$\frac{1}{2}(x_5 + x_6) = \frac{1}{2}(14 + 14.5) = 14.25$$

Das ist unser gesuchtes 0.25-Quantil, was wir auch so schreiben: $\underline{\underline{x_{0.25} = 14.25}}$.

(Der exakte Wert nach Formel (3.10) bzw. mit EXCEL ausgerechnet ergibt 14.375)

[10] Vatikanstadt ist in dieser Aufzählung nicht angegeben, weil es dort keine Schulen gibt. Kinder von Einwohnern des Vatikans gehen in der Regel im benachbarten Italien zur Schule.

Land	Schuljahre	Land	Schuljahre
Liechtenstein	12	Tschechien	15.5
San Marino	12.5	Ungarn	15.5
Luxemburg	13.5	Schweiz	15.5
Serbien	13.5	Frankreich	16
Kroatien	14	Italien	16
Rumänien	14.5	Belgien	16.5
Slowakei	14.5	Deutschland	16.5
Ukraine	15	Dänemark	17
Polen	15	Niederlande	17
Österreich	15.5	Slowenien	17

Tabelle 3.4: Expected Years of Schooling of children in years: Number of years of schooling that a child of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates persist throughout the child's life. Source: UNESCO Institute for Statistics (2012), http://stats.uis.unesco.org

Beispiel 13 *Gib für die Daten der Tabelle 3.4 das Quantil für $p = 1/3$ an.*

$$k = n \cdot p = 20 \cdot 1/3 = 6.67$$

Das ist nicht ganzzahlig, daher runden wir auf die nächste ganze Zahl auf:

$$\lceil 6.67 \rceil = 7$$

Der an siebter Stelle stehende Wert ist gleich 14.5*, daher ist* $x_{0.33} = \underline{\underline{14.5}}$.

(Der exakte Wert nach Formel (3.10) ergibt 14.667*)*

Aufgabe 2 Gib für die Daten der Tabelle 3.4 das 0.65-Quantil an (berechnet nach obigem Näherungsverfahren) und vergleiche dein Ergebnis mit dem exakten Wert aus *R*.

Einige Quantile (mit einem bestimmten p-Wert) spielen in der Datenauswertung eine besondere Rolle, daher haben sie eigene Namen bekommen:

Mediane, Quartile und Ähnliches

Wir sind manchmal daran interessiert, die geordneten Daten nicht nur an einer bestimmten Stelle zu teilen, sondern gleich in q gleich große Gruppen zu unterteilen. q kann zum Beispiel 2 sein; wir teilen dann unsere Daten in zwei gleich

große Gruppen. Mit $q = 4$ bilden wir vier Gruppen, mit $q = 10$ zehn Gruppen, mit $q = 100$ hundert Gruppen. Um diese Teilungen zu bewerkstelligen, benötigen wir jeweils $(q - 1)$ Teilungspunkte (Mit einem Punkt können wir eine Datenmenge in 2 Gruppen teilen, mit 3 Punkten in vier, mit neun Punkten in 10 Gruppen und mit neunundneunzig Punkten in 100 Gruppen).

Sehen wir zunächst uns wieder die – der Größe nach geordneten – Daten aus Beispiel 3 an:

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

Sei zunächst $q = 2$, d.h. wir wollen in zwei Gruppen aufteilen und benötigen dazu $q - 1 = 1$ Teilungspunkt. Zur Bestimmung dieses Punktes bedienen wir uns der Quantile aus dem letzten Abschnitt und bestimmen das Quantil mit

$$p = \frac{q-1}{q} = \frac{1}{2} = 0.5 \tag{3.14}$$

Das ist mittlerweile ja ziemlich einfach für uns:

Beispiel 14

$k = 14 \cdot 0.5 = 7$ ganzzahlig, daher:
$x_{0.5} = \frac{x_7 + x_8}{2} = \frac{3+3}{2} = 3$

Das 0.5-Quantil teilt unsere Stichprobe in genau zwei Teile. Der «mittelste» Datenwert ist 3. Oberhalb und unterhalb liegen je 50% der Werte. Das 0.5-Quantil wird auch **Median** (oder *Zentralwert*) genannt. Der Median ist jener Wert, der von mindestens der Hälfte der Merkmalswerte nicht unterschritten wird. (*Mindestens* deshalb, weil es auch sein könnte, dass der Median ein Merkmalswert ist, der mehrfach vorkommt).

Wir können das auch in unseren Daten visualisieren und einzeichnen, wo wir die Stichprobe teilen müssen:

1 1 2 2 2 3 3 | 3 3 3 4 4 5 6

In Excel lautet der Befehl für den Median: `=MEDIAN(Zahl1;Zahl2;...)`.
In *R* gibt der Befehl `median(x)` den Median der Daten in `x` aus.

Für den Median können wir die Formeln 3.13 und 3.12 auch so angeben:

Wenn n eine gerade Zahl ist, dann ist der Median der Mittelwert zwischen dem Wert an der Stelle $\frac{n}{2}$ und dem nächsten Beobachtungswert an der Stelle $\frac{n}{2}+1$:

$$x_{0.5} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \tag{3.15}$$

Wenn n ungerade ist, dann ist der Median der Wert an der Stelle $\frac{n+1}{2}$:

$$x_{0.5} = x_{\frac{n+1}{2}} \tag{3.16}$$

Nachdem wir unsere Daten mit dem Median in zwei gleiche Hälften geteilt haben, wiederholen wir das und teilen die beiden Hälften wieder genau in der Mitte:

Für eine Unterteilung in $q = 4$ Teile benötigen wir $q - 1 = 3$ Teilungspunkte, die wir auch **Quartile** (oder *Viertelwerte*) nennen. Das **1. Quartil** ist das $\frac{1}{4}$-Quantil, das **2. Quartil** das $\frac{2}{4}$-Quantil und das **3. Quartil** das $\frac{3}{4}$-Quantil. Das 1. Quartil wird auch *unteres Quartil* genannt, das 3. Quartil *oberes Quartil* – und das 2. Quartil ist nichts anderes als der *Median* (siehe oben).

Oberhalb des *oberen Quartils* und unterhalb des *unteren Quartils* liegen je 25 % der Elemente, dazwischen die restlichen 50 %.

Wir verwenden wieder das Näherungsverfahren:

Beispiel 15 *Gib das nach dem Näherungsverfahren bestimmte 1., 2. und 3. Quartil der Daten aus Beispiel 3 an:*

1. Quartil:

$$\begin{aligned} &n = 14, \quad q = 4, \quad p = \tfrac{1}{4} = 0.25 \\ &k = 14 \cdot 0.25 = 3.5 \\ &\lceil 3.5 \rceil = 4 \\ &x_{0.25} = \boxed{2} \end{aligned}$$

2. Quartil:

$$\begin{aligned} &n = 14, \quad q = 4, \quad p = \tfrac{2}{4} = 0.5 \\ &k = 14 \cdot 0.5 = 7 \\ &x_{0.5} = \tfrac{x_7 + x_8}{2} = \boxed{3} \end{aligned}$$

3. Quartil:

$$n = 14, \quad q = 4, \quad p = \tfrac{3}{4} = 0.75$$
$$k = 14 \cdot 0.75 = 10.5$$
$$\lceil 10.5 \rceil = 11$$
$$x_{0.75} = \boxed{4}$$

Anm.: Wenn wir nicht das Näherungsverfahren verwenden, sondern zum Beispiel R, dann erhalten wir:

```
quantile(bsp3,0.25)
2

quantile(bsp3,0.5)
3

quantile(bsp3,0.75)
3.75
```

Manchmal wird ergänzend auch noch ein **0. Quartil** und ein **4. Quartil** angegeben: Das ist nichts anderes als das *Minimum* und das *Maximum*, die unsere Daten am Anfang und Ende «einrahmen», siehe Tab.3.5.

Quantil	**weitere Bezeichnungen**
0.00-Quantil	*0.Quartil* oder *Minimum*
0.25-Quantil	*1.Quartil* oder *unteres Quartil*
0.50-Quantil	*2.Quartil* oder *Median*
0.75-Quantil	*3.Quartil* oder *oberes Quartil*
1.00-Quantil	*4.Quartil* oder *Maximum*

Tabelle 3.5: Quartile und ihre synonymen Bezeichnungen

In *R* gibt der Befehl `summary(x)` die in Tab.3.5 angeführten Werte einer Stichprobe an, dazu auch noch das arithmetische Mittel. Sind zum Beispiel die Platzierungen aus Tab. 3.3 (ohne die letzten drei, nicht numerischen Werte DNF2 und DNQ2) im Vektor `frida` gespeichert, dann ergibt `summary(frida)` das Ergebnis:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	3.745	4.500	10.000

Aufgabe 3 Gegeben sei die Körpergröße der 7 Zwerge. Bestimme das Minimum, Maximum sowie das $1. - 3.$ Quartil sowohl näherungsweise als auch deren exakten Werte:

Name	Größe (in *cm*)
Doc	85
Grumpy	80
Happy	62
Sleepy	81
Bashful	70
Sneezy	80
Dopey	88

Aufgabe 4 Gegeben sei die Größe der 7 roten Zwerge, die innerhalb einer Entfernung von 10 Lichtjahren zur Erde liegen. Bestimme das Minimum, Maximum sowie das $1. - 3.$ Quartil sowohl «visuell» als auch deren exakten Werte:

Bezeichnung	Durchmesser (in *Tausend km*)
Luyten 726-8 A	195
Luyten 726-8 B	195
Proxima Centauri	196.4
Wolf 359	222.8
Barnards Stern	273
Ross 154	334.2
Lalande 21185	547.4

Neben Median und Quartilen sind manchmal auch noch Unterteilungen in $q = 10$ Gruppen (zu je 10%) interessant – wir nennen die Teilungspunkte dann **Dezile**, sowie jene mit $q = 100$ (also eine Einteilung in Gruppen zu je 1%), die so genannten **Perzentile**. Die Berechnung dürfte aber hoffentlich klar sein: Für das 8. Dezentil berechnen wir zum Beispiel das 0.8-Quantil, für das 5. Perzentil das 0.05-Quantil etc.

Anmerkungen zur Verwendung von Mittelwert, Median und Modalwert

Mittelwert und Median werden beide verwendet, um eine umfangreiche Datenmenge durch einen einzigen Wert möglichst gut zu repräsentieren. Im allgemeinen Sprachgebrauch sagen wir auch: wir suchen den *Durchschnitt*. Mittelwert und Median haben dabei unterschiedliche Eigenschaften, die sie – je nach Anwendungsfall – geeigneter erschienen lassen, diese Aufgabe zu erfüllen.

Sie zeigen zum Beispiel unterschiedliches *Resistenzverhalten* (Widerstandsfähigkeit) gegenüber Ausreißern:

Der Mittelwert ist sehr empfindlich gegenüber Ausreißern. Nachdem jeder einzelne Wert in seiner vollen Höhe in die Berechnung des Mittelwerts einfließt, kann jeder einzelne Wert $\bar{x}$ auch bedeutend verändern.

Der Median hingegen wird durch einzelne Ausreißer kaum verändert. Ändert sich ein Datenwert – egal um wie viel – so ändert der Median seinen Wert nur dann, wenn dieser Datenwert von der einen Hälfte der geordneten Daten in die andere Hälfte wandert.

Trittst du in Gehaltsverhandlungen mit deiner Chefin und nimmst einen «mittleren Wert» aus allen Gehältern innerhalb der Firma als Grundlage, dann verwende den arithmetischen Mittelwert, weil dann das überproportionale Gehalt deiner Chefin als «Ausreißer» den Mittelwert erhöhen wird.
Deine Chefin wird hingegen versuchen, den Median als Basis heranzuziehen, weil dann die Höhe ihres Gehalts keinen Einfluss hat.
Letztendlich ist es aber ziemlich wahrscheinlich, dass du den Modalwert erhalten wirst, also das, was die meisten kriegen …

Gut zu wissen: Der Unterschied zwischen Mittelwert, Median und Modalwert

Bei der praktischen Berechnung gibt es auch einen Unterschied zwischen Mittelwert und Median: Während für den arithmetischen Mittelwert die (ungeordnete) Urliste herangezogen werden kann, müssen zur Berechnung des Medians die Daten zuerst in eine (der Größe nach geordnete) Rangliste gebracht werden.

Ein weiterer Unterschied zwischen Median und Mittelwert ist der, dass für das arithmetische Mittel numerische Daten notwendig sind. Für den Median reicht es hingegen, wenn wir die Daten der Größe nach ordnen können. Und das geht bereits mit *ordinalskalierten* Daten, also *Rangmerkmalen*. **Hinweis:** Diese Behauptung gilt nur mit Einschränkungen: Formel 3.15 können wir nicht anwenden, wenn wir nichtnumerische ordinalskalierte Daten haben. Dann müssen wir immer, unabhängig davon ob wir eine gerade oder ungerade Anzahl von Elementen haben, Formel 3.16 verwenden.

Der Modalwert (häufigste Wert) kann als einziger auch für nominal skalierte Merkmale angegeben werden. Er gibt ein «typisches» Ergebnis an. Ein Unterschied des Modalwertes ist auch, dass es sich dabei immer um einen tatsächlich beobachteten Wert handelt. Median und Mittelwert hingegen sind errechnete

Größen, die als Beobachtung in der Stichprobe gar nicht vorkommen müssen.

3.2 Streuungskennwerte empirischer Häufigkeitsverteilungen

Lage-Kennzeichen geben noch kein vollständiges Bild der Daten und ihrer Verteilung wieder. So können zum Beispiel verschiedene Daten alle denselben Mittelwert haben, die Histogramme und Häufigkeitssummenkurven hingegen sehen alle anders aus. Offensichtlich gibt es noch ein anderes wichtiges Unterscheidungsmerkmal.

Es sind dies die so genannten **Streuungskennwerte**. Sie charakterisieren die Schwankungen der Daten und geben Auskunft darüber, wie stark die einzelnen Werte voneinander abweichen beziehungsweise wie weit sie vom Durchschnitt abweichen. Je weniger die einzelnen Merkmalswerte mit dem Durchschnitt übereinstimmen, umso wichtiger ist die Angabe von Streuungskennwerten. Einfach anzugebende Streuungsmaße sind

Spannweite und Interquartilsabstand

Die **Spannweite** (auch: *Variationsbreite*) ist die Distanz (mathematisch: die Differenz) zwischen dem größten und dem kleinsten Beobachtungswert:

$$\Delta = x_{\max} - x_{\min} \tag{3.17}$$

Der **Interquartilsabstand** (eng.: *interquartile range*) ist die Distanz zwischen dem 1. und 3. Quartil:

$$IQR = x_{0.75} - x_{0.25} \tag{3.18}$$

und beinhaltet die mittleren 50% der Daten. Zur grafischen Darstellung des Quartilsabstands dienen so genannte **Boxplots**. Dabei wird zwischen dem 1. und 3. Quartil ein Rechteck – die «Box» – gezeichnet, mit einer «Mittellinie» in der Höhe des Medians. An der Box hängt noch oben und unten je eine durch einen waagrechten Strich abgeschlossene Linie; sie werden auch «Whisker» genannt und reichen bis zum Minimum bzw. Maximum der Daten. Abb.3.1 zeigt ein Boxplot-Diagramm der Daten aus Beispiel 3.

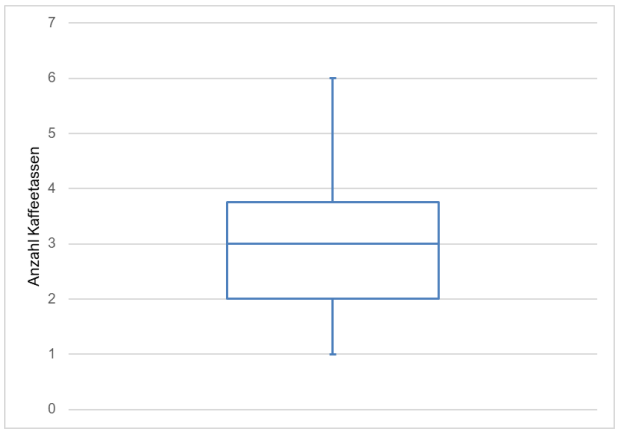

Abb. 3.1: Das Boxplot-Diagramm zu den Daten aus Beispiel 3: 50% der Daten liegen innerhalb der «Box», die beiden «Antennen» reichen vom Minimum bis zum Maximum und umfassen somit 100% der Daten

In *R* erhalten wir mit dem Befehl `range(x)` nicht die Spannweite, sondern das Minimum und das Maximum der Daten.
Der Interquartilsabstand kann mit dem Befehl `IQR(x)` berechnet werden. `IQR(frida)` ergibt zum Beispiel `2.5`

Der Quartilsabstand kann auch dazu verwendet werden, um in einer ersten Näherung *Ausreißer-Grenzen* festzulegen:

$$A_u = x_{0.25} - 1.5 \cdot IQR \tag{3.19}$$
$$A_o = x_{0.75} + 1.5 \cdot IQR \tag{3.20}$$

Datenwerte, die außerhalb des Intervalls $[A_u, A_o]$ liegen, können als extreme Werte (Ausreißer) angesehen und eventuell gestrichen werden. *Achtung:* Dies ist nur ein näherungsweises Vorgehen. Für unsere Zwecke aber meist ausreichend.

Aufgabe 5 Bestimme zu den Daten der Tabelle 3.4 die Spannweite und den Quartilsabstand. Gibt es auf Grund dieser Streuungswerte Anzeichen, dass die Daten Ausreißer enthalten?

Variationsbreite und Quartilsabstand sagen zwar schon einiges über die Verteilung der Daten aus, berücksichtigen aber nur einige wenige Werte (eben Maximum und Minimum und die Quartile). Noch informativer wäre ein Kennwert, der alle Messwerte berücksichtigt. Das machen die Folgenden:

Empirische Varianz und Standardabweichung

Die **empirische Varianz** (auch: *Stichprobenstreuung*) charakterisiert die Abweichungen der Daten von ihrem Mittelwert. Es ist die Summe der quadrierten Abweichungen der Beobachtungswerte von ihrem arithmetischen Mittelwert dividiert durch die Anzahl aller Werte minus Eins. Sie wird auch *mittlere quadratische Abweichung* genannt:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.21}$$

Die Verwendung der Quadrate in Formel (3.21) hat zwei Vorteile: Durch das Quadrieren werden alle Abweichungen positiv und können sich nicht gegenseitig aufheben; außerdem werden größere Abweichungen stärker berücksichtigt als kleinere.

Es gibt aber auch einen Nachteil: Die Varianz ist nicht sehr anschaulich und damit praktisch nicht verwendbar. Wenn wir zum Beispiel irgendeine Zufallsvariable untersuchen, die wir in Euro messen, hat die Varianz die Einheit Quadrateuro (die Einheiten werden ja mitquadriert); es ist nicht so leicht, sich darunter auch etwas vorzustellen. Besser wäre ein Maß, das in derselben Einheit wie die Messwerte angegeben werden kann:

Die **Standardabweichung** ist die positive Quadratwurzel[11] aus der Varianz:

$$s = +\sqrt{s^2} \tag{3.22}$$

Jetzt kann man sich auch leichter vor Augen halten, was dieses Maß repräsentiert: Die Standardabweichung gibt an, um wieviel ein einzelner Messwert durchschnittlich (also sozusagen «standardmäßig») vom Mittelwert abweicht. Eine geringe Standardabweichung bedeutet, die Daten liegen eher enger um den Mittelwert; eine hohe Standardabweichung weist auf eine stärkere Streuung um den Mittelwert hin.

Beispiel 16 *Berechne die Standardabweichung der Körpergröße der 7 Zwerge (Aufgabe 3):*

Dazu müssen wir zunächst den Mittelwert ausrechnen:

$$\bar{x} = \frac{85 + 80 + 62 + 81 + 70 + 80 + 88}{7} = \frac{546}{7} = 78\ cm$$

[11] normalerweise kann die Wurzel einer Zahl positiv oder negativ sein. 4 zum Beispiel hat die beiden Wurzeln $+2$ und -2. Laut Definition verwenden wir für die Standardabweichung aber nur die positive Wurzel.

und daraus die Varianz:

$$s^2 = \frac{(85-78)^2+(80-78)^2+(62-78)^2+(81-78)^2+(70-78)^2+(80-78)^2+(88-78)^2}{(7-1)} =$$
$$= \frac{7^2+2^2+(-16)^2+3^2+(-8)^2+2^2+10^2}{6} = \frac{49+4+256+9+64+4+100}{6} = \frac{486}{6} = 81$$

und schließlich die Standardabweichung:

$$s = \sqrt{81} = 9\ cm$$

Aufgabe 6 Wie groß ist die Varianz der in Beispiel 3 gegebenen Daten? (Didaktischer Hinweis: Bevor du den Rechner anwirfst: Versuch einmal, diese Aufgabe mit «Papier und Bleistift» zu lösen.).

In *Excel* wird die Standardabweichung einer Stichprobe mit der Funktion `=STABW.S(Zahl1;Zahl2;...)` berechnet.
In *LibreOffice Calc* heißt der Befehl: `=STABW(Zahl1;Zahl2;...)`
In *R* erhalten wir die Standardabweicheung mit `sd(x)` bzw. die Varianz mit `var(x)`.

Aufgabe 7 Der folgende Datensatz besteht aus elf Elementen und hat einen arithmetischen Mittelwert von 25. Außerdem wissen wir, dass die Daten völlig symmetrisch um den Mittelwert gestreut sind.

Welche Werte musst du für x_1 und x_{11} einsetzen, damit die Varianz 26 beträgt?

$$x_1, 21, 22, 23, 24, 25, 26, 27, 28, 29, x_{11}$$

Aufgabe 8 Kannst du – ohne konkret jede Kennzahl auszurechnen – «auf einen Blick» sagen und begründen, welche der drei folgenden Datensätze die größte Standardabweichung hat und welche die kleinste?

A) 0, 20, 40, 50, 60, 80, 100
B) 0, 48, 49, 50, 51, 52, 100
C) 0, 1, 2, 50, 98, 99, 100

Variationskoeffizient

Der **Variationskoeffizient** ist der Quotient der Standardabweichung dividiert durch den Betrag des arithmetischen Mittelwerts. Er wird meistens in Prozent angegeben:

$$v_x = \frac{s}{|\bar{x}|} \cdot 100\% \tag{3.23}$$

Der Variationskoeffizient ist demnach eine Art *relative Standardabweichung*. Er wird verwendet, wenn man Standardabweichungen miteinander vergleichen will.

3.3 Zentrierter, normierter und standardisierter Beobachtungswert

Manchmal müssen Werte aus unterschiedlichen Verteilungen miteinander verglichen werden. Zum Beispiel wird der IQ (Intelligenzquotient) oft auf einer Skala gemessen, die den Mittelwert 100 IQ-Punkte und die Standardabweichung 15 IQ-Punkte hat. Mitunter gibt es aber auch andere Tests, die auf einer Standardabweichung von 16 oder 24 IQ-Punkten basieren. Der unmittelbare Vergleich der Absolutwerte der Testergebnisse ist in diesem Fall nicht sehr aussagekräftig[12] und entspricht in etwa dem sprichwörtlichen Vergleich von Äpfeln und Birnen.

Besser ist es dann, die Einzelwerte zuerst zu «relativieren» und auf eine einheitliche Basis zu bringen. Diesen Vorgang nennt man *Standardisieren*. Er läuft in zwei Schritten ab: einer Zentrierung und einer Normierung.

Der **zentrierte Beobachtungswert** ist der Beobachtungswert minus des arithmetischen Mittelwerts:

$$x_i - \bar{x} \tag{3.24}$$

Zentriert man einen gesamten Datensatz, dann ist das arithmetische Mittel der zentrierten Daten gleich Null.

Der **normierte Beobachtungswert** ist der Beobachtungswert dividiert durch die Standardabweichung:

$$\frac{x_i}{s} \tag{3.25}$$

[12] Das gilt nicht nur für IQ-Tests sondern für alle Vergleich von Daten aus unterschiedlichen Verteilungen

Der **standardisierte Beobachtungswert** ist der zentrierte Beobachtungswert dividiert durch die Standardabweichung, es wird also zuerst zentriert und anschließend normiert:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.26}$$

Dieser Wert (auch als **z-Wert** bezeichnet) gibt an, «wie viele Standardabweichungen» der Messwert x_i vom Mittelwert $\bar{x}$ entfernt ist. Der *z-Wert* ist dimensionslos. Er kann positiv, negativ oder Null sein. Das Vorzeichen gibt Auskunft darüber, ob der zugehörige Messwert über- oder unterdurchschnittlich ist. Ein *z-Wert* von 2 gibt zum Beispiel an, dass der zugehörige Messwert 2 Standardabweichungen oberhalb des Mittelwertes liegt; ein *z-Wert* von -1.7 bedeutet, dass der zugehörige Messwert 1.7 Standardabweichungen unterhalb des Mittelwertes liegt; ein *z-Wert* von 0 bedeutet, dass der zugehörige Messwert genauso groß ist, wie der Mittelwert.

Bildet man von allen Messwerten die *z-Werte*, und wertet diese statistisch aus, so zeigt sich, dass der Mittelwert der *z-Werte* gleich 0 ist, und ihre Standardabweichung 1.

Aufgabe 9 Gib zu den Daten aus Aufgabe 7 die standardisierten $z-$Werte an.

Für einen besseren Überblick

hier noch einmal eine Übersicht über die Kennwerte für empirische Häufigkeitsverteilungen und für welche Skalen wir sie verwenden können:

Datenart	**Kennwerte**
Nominaldaten	Modalwert
Ordinaldaten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand
Intervallskalierte Daten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand, arithmetisches Mittel, Standardabweichung, standardisierter Beobachtungswert
Rationalskalierte Daten	Modalwert, Minimum und Maximum, Median und Quartil, Spannweite und Interquartilsabstand, arithmetisches und geometrisches Mittel, Standardabweichung, Variationskoeffizient, standardisierter Beobachtungswert

Tabelle 3.6: Übersicht: je nach Datenart mögliche Kennwerte