# Measures of Central Tendency and Dispersion

# 3

A data set may contain many variables and observations. However, we are not always interested in each of the measured values but rather in a summary which interprets the data. Statistical functions fulfil the purpose of summarizing the data in a meaningful yet concise way.

*Example 3.0.1* Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

$$22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,$$
$$25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.$$

How do we draw conclusions from this data? Looking at the individual values gives us a feeling about the temperatures one can experience in Bangkok, but it does not provide us with a clear summary. It is evident that the average of these 31 values as "Sum of all values/Total number of observations" $(22 + 24 + \cdots + 28 + 29)/31 = 26.48$ is meaningful in the sense that we know what temperature to expect "on average". To choose the right clothing for the holidays, we may also be interested in knowing the temperature range to understand the variability in temperature, which is between 21 and 31 °C. Summarizing 31 individual values with only three numbers (26.48, 21, and 31) will provide sufficient information to plan the holidays.

In this chapter, we focus on the most important statistical concepts to summarize data: these are measures of central tendency and variability. The applications of each measure depend on the scale of the variable of interest, see Appendix D.1 for a detailed summary.



C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_3

## 3.1 Measures of Central Tendency

A natural human tendency is to make comparisons with the "average". For example, a student scoring 40 % in an examination will be happy with the result if the average score of the class is 25 %. If the average class score is 90 %, then the student may not feel happy even if he got 70 % right. Some other examples of the use of "average" values in common life are mean body height, mean temperature in July in some town, the most often selected study subject, the most popular TV show in 2015, and average income. Various statistical concepts refer to the "average" of the data, but the right choice depends upon the nature and scale of the data as well as the objective of the study. We call statistical functions which describe the average or centre of the data **location parameters** or **measures of central tendency**.

### 3.1.1 Arithmetic Mean

The **arithmetic mean** is one of the most intuitive measures of central tendency. Suppose a variable of size $n$ consists of the values $x_1, x_2, \ldots, x_n$. The arithmetic mean of this data is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.1}$$

In informal language, we often speak of "the average" or just "the mean" when using the formula (3.1).

To calculate the arithmetic mean for grouped data, we need the following frequency table:

| Class intervals $a_j$ | $a_1 = e_0 - e_1$ | $a_2 = e_1 - e_2$ | ... | $a_k = e_{k-1} - e_k$ |
|---|---|---|---|---|
| Absolute freq. $n_j$ | $n_1$ | $n_2$ | ... | $n_k$ |
| Relative freq. $f_j$ | $f_1$ | $f_2$ | ... | $f_k$ |

Note that $a_1, a_2, \ldots, a_k$ are the $k$ class intervals and each interval $a_j (j = 1, 2, \ldots, k)$ contains $n_j$ observations with $\sum_{j=1}^{k} n_j = n$. The relative frequency of the $j$th class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. The mid-value of the $j$th class interval is defined as $m_j = (e_{j-1} + e_j)/2$, which is the mean of the lower and upper limits of the interval. The **weighted arithmetic mean** for grouped data is defined as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j. \tag{3.2}$$

*Example 3.1.1* Consider again Example 3.0.1 where we looked at the temperature in Bangkok during December. The measurements were

$$22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,$$
$$25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29\,.$$

The arithmetic mean is therefore

$$\bar{x} = \frac{22 + 24 + 21 + \cdots + 28 + 29}{31} = 26.48\,^{\circ}\mathrm{C}.$$

In *R*, the arithmetic mean can be calculated using the `mean` command:

```
weather <- c(22,24,21,,30,28,29)
mean(weather)
[1] 26.48387
```

Let us assume the data in Example 3.0.1 is summarized in categories as follows:

| Class intervals | $< 20$ | $(20-25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|---|
| Absolute frequencies | $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| Relative frequencies | $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

We can calculate the (weighted) arithmetic mean as

$$\bar{x} = \sum_{j=1}^{k} f_j m_j = 0 + \frac{12}{31} \cdot 22.5 + \frac{18}{31} \cdot 27.5 + \frac{1}{31} 32.5 + 0 \approx 25.7.$$

In *R*, we use the `weighted.mean` function to obtain the result. The function requires to specify the (hypothesized) means for each group, for example the middle values of the class intervals, as well as the weights.

```
weighted.mean(c(22.5,27.5,32.5),c(12/31,18/31,1/31))
```

Interestingly, the results of the mean and the weighted mean differ. This is because we use the middle of each class as an approximation of the mean within the class. The implication is that we assume that the values are uniformly distributed within each interval. This assumption is obviously not met. If we had knowledge about the mean in each class, like in this example, we would obtain the correct result as follows:

$$\bar{x} = \sum_{j=1}^{k} f_j \bar{x}_j = 0 + \frac{12}{31} \cdot 23.83333 + \frac{18}{31} \cdot 28 + \frac{1}{31} 32.5 + 0 = 26.48387.$$

However, the weighted mean is meant to estimate the arithmetic mean in those situations where only grouped data is available. It is therefore typically used to obtain an approximation of the true mean.

**Properties of the Arithmetic Mean.**

(i) The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0. \tag{3.3}$$

(ii) If the data is linearly transformed as $y_i = a + bx_i$, where $a$ and $b$ are known constants, it holds that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}. \tag{3.4}$$

*Example 3.1.2* Recall Examples 3.0.1 and 3.1.1 where we considered the temperatures in December in Bangkok. We measured them in degrees Celsius, but someone from the USA might prefer to know them in degrees Fahrenheit. With a linear transformation, we can create a new temperature variable as

$$\text{Temperature in } °\text{F} = 32 + 1.8\,\text{Temperature in } °\text{C}.$$

Using $\bar{y} = a + b\bar{x}$, we get $\bar{y} = 32 + 1.8 \cdot 26.48 \approx 79.7\,°\text{F}$.

### 3.1.2 Median and Quantiles

The median is the value which divides the observations into two equal parts such that at least 50 % of the values are greater than or equal to the median and at least 50 % of the values are less than or equal to the median. The median is denoted by $\tilde{x}_{0.5}$; then, in terms of the empirical cumulative distribution function, the condition $F(\tilde{x}_{0.5}) = 0.5$ is satisfied. Consider the $n$ observations $x_1, x_2, \ldots, x_n$ which can be ordered as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. The calculation of the median depends on whether the number of observations $n$ is odd or even. When $n$ is odd, then $\tilde{x}_{0.5}$ is the middle ordered value. When $n$ is even, then $\tilde{x}_{0.5}$ is the arithmetic mean of the two middle ordered values:

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases} \tag{3.5}$$

*Example 3.1.3* Consider again Examples 3.0.1–3.1.2 where we evaluated the temperature in Bangkok in December. The ordered values $x_{(i)}$, $i = 1, 2, \ldots, 31$, are as follows:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| °C | 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| (i) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| °C | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| (i) | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

We have $n = 31$, and therefore $\tilde{x}_{0.5} = x_{((n+1)/2)} = x_{((31+1)/2)} = x_{(16)} = 26$. Therefore, at least 50 % of the 31 observations are greater than or equal to 26 and at least 50 % are less than or equal to 26. If one value was missing, let us say the last observation, then the median would be calculated as $\frac{1}{2}(x_{(30/2)} + x_{(30/2+1)}) = \frac{1}{2}(26 + 26) = 26$. In *R*, we would have obtained the results using the `median` command:

```
median(weather)
```

If we deal with grouped data, we can calculate the median under the assumption that the values within each class are equally distributed. Let $K_1, K_2, \ldots, K_k$ be $k$ classes with observations of size $n_1, n_2, \ldots, n_k$, respectively. First, we need to determine which class is the median class, i.e. the class that includes the median. We define the median class as the class $K_m$ for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j \geq 0.5 \tag{3.6}$$

hold. Then, we can determine the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m \tag{3.7}$$

where $e_{m-1}$ denotes the lower limit of the interval $K_m$ and $d_m$ is the width of the interval $K_m$.

*Example 3.1.4* Recall Example 3.1.1 where we looked at the grouped temperature data:

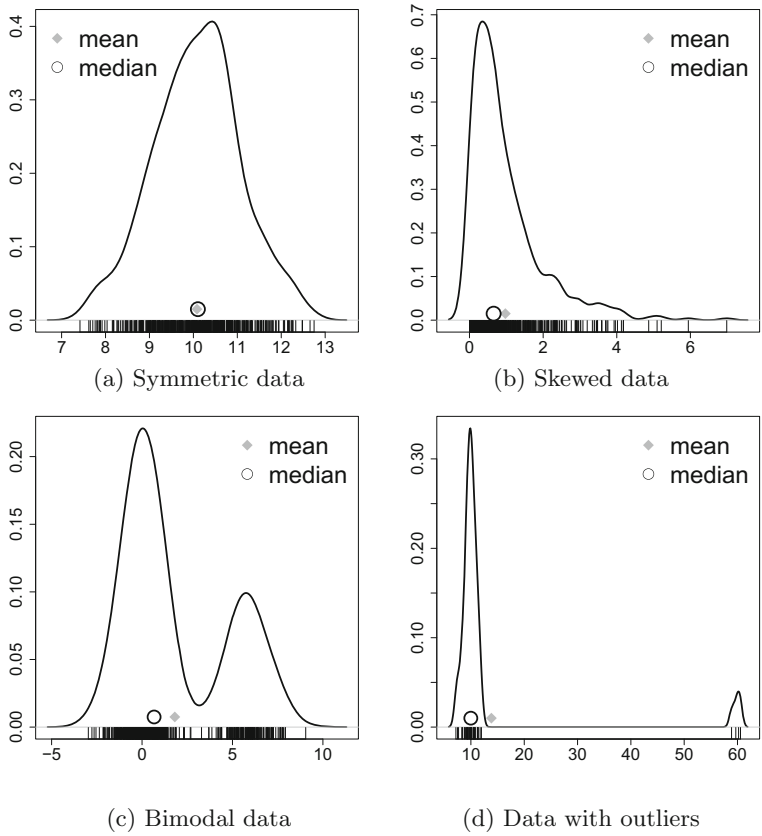
| Class intervals | <20 | (20–25] | (25, 30] | (30, 35] | >35 |
|---|---|---|---|---|---|
| $n_j$ | $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_j$ | $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |
| $\sum_j f_j$ | 0 | $\frac{12}{31}$ | $\frac{30}{31}$ | 1 | 1 |

For the third class ($m = 3$), we have

$$\sum_{j=1}^{m-1} f_j = \frac{12}{31} < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j = \frac{30}{31} \geq 0.5.$$

We can therefore calculate the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 25 + \frac{0.5 - \frac{12}{31}}{\frac{18}{31}} \cdot 5 \approx 25.97.$$




(a) Symmetric data (b) Skewed data

(c) Bimodal data (d) Data with outliers

**Fig. 3.1** Arithmetic mean and median for different data

**Comparing the Mean with the Median.** In the above examples, the mean and the median turn out to be quite similar to each other. This is because we looked at data which is symmetrically distributed around its centre, i.e. on average, we can expect 26 °C with deviations that are similar above and below the average temperature. A similar example is given in Fig. 3.1a: we see that the raw data is summarized by using ticks at the bottom of the graph and by using a kernel density estimator. The mean and the median are similar here because the distribution of the observations is symmetric around the centre. If we have skewed data (Fig. 3.1b), then the mean and the median may differ. If the data has more than one centre, such as in Fig. 3.1c, neither the median nor the mean has meaningful interpretations. If we have outliers (Fig. 3.1d), then it is wise to use the median because the mean is sensitive to outliers. These examples show that depending on the situation of interest either the mean, the median, both or neither of them can be useful.

**Quantiles.** Quantiles are a generalization of the idea of the median. The median is the value which splits the data into two equal parts. Similarly, a quantile partitions the data into other proportions. For example, a 25 %-quantile splits the data into two parts such that at least 25 % of the values are less than or equal to the quantile and at least 75 % of the values are greater than or equal to the quantile. In general, let $\alpha$ be a number between zero and one. The $(\alpha \times 100)$%-quantile, denoted as $\tilde{x}_\alpha$, is defined as the value which divides the data in proportions of $(\alpha \times 100)$% and $(1-\alpha) \times 100$ % such that at least $\alpha \times 100$ % of the values are less than or equal to the quantile and at least $(1-\alpha) \times 100$ % of the values are greater than or equal to the quantile. In terms of the empirical cumulative distribution function, we can write $F(\tilde{x}_\alpha) = \alpha$. It follows immediately that for $n$ observations, at least $n\alpha$ values are less than or equal to $\tilde{x}_\alpha$ and at least $n(1-\alpha)$ observations are greater than or equal to $\tilde{x}_\alpha$. The median is the 50 %-quantile $\tilde{x}_{0.5}$. If $\alpha$ takes the values $0.1, 0.2, \ldots, 0.9$, the quantiles are called **deciles**. If $\alpha \cdot 100$ is an integer number (e.g. $\alpha \times 100 = 95$), the quantiles are called **percentiles**, i.e. the data is divided into 100 equal parts. If $\alpha$ takes the values 0.2, 0.4, 0.6, and 0.8, the quantiles are known as **quintiles** and they divide the data into five equal parts. If $\alpha$ takes the values 0.25, 0.5, and 0.75, the quantiles are called **quartiles**.

Consider $n$ ordered observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The $\alpha \cdot 100$ %-quantile $\tilde{x}_\alpha$ is calculated as

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer number,} \\ & \text{choose } k \text{ as the smallest integer } > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{if } n\alpha \text{ is an integer.} \end{cases} \tag{3.8}$$

*Example 3.1.5* Recall Examples 3.0.1–3.1.4 where we evaluated the temperature in Bangkok in December. The ordered values $x_{(i)}$, $i = 1, 2, \ldots, 31$ are as follows:

| °C | 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (i) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| °C | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| (i) | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

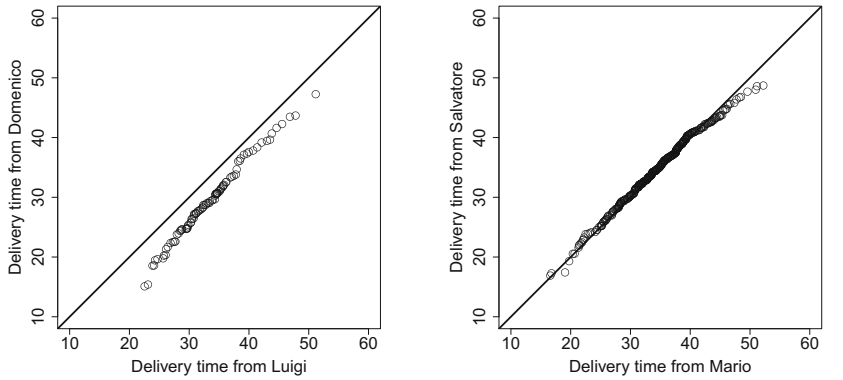
To determine the quartiles, i.e. the 25, 50, and 75 % quantiles, we calculate $n\alpha$ as $31 \cdot 0.25 = 7.75$, $31 \cdot 0.5 = 15.5$, and $31 \cdot 0.75 = 23.25$. Using (3.8), it follows that

$$\tilde{x}_{0.25} = x_{(8)} = 25, \quad \tilde{x}_{0.5} = x_{(16)} = 26,$$
$$\tilde{x}_{0.75} = x_{(24)} = 29.$$

In *R*, we obtain the same results using the `quantile` function. The `probs` argument is used to specify $\alpha$. By default, the quartiles are reported.

```
quantile(weather)
quantile(weather, probs=c(0,0.25,0.5,0.75,1))
```

However, please note that *R* offers nine different ways to obtain quantiles, each of which can be chosen by the `type` argument. See Hyndman and Fan (1996) for more details.




(a) QQ-plot for Luigi and Domenico (b) QQ-plot for Mario and Salvatore

**Fig. 3.2** QQ-plots for the pizza delivery time for different drivers

### 3.1.3 Quantile–Quantile Plots (QQ-Plots)

If we plot the quantiles of two variables against each other, we obtain a Quantile–Quantile plot (QQ-plot). This provides a simple summary of whether the distributions of the two variables are similar with respect to their location or not.

*Example 3.1.6* Consider again the pizza data which is described in Appendix A.4. We may be interested in the delivery time for different drivers to see if their performance is the same. Figure 3.2a shows a QQ-plot for the delivery time of driver Luigi and the delivery time of driver Domenico. Each point refers to the $\alpha\%$ quantile of both drivers. If the point lies on the bisection line, then they are identical and we conclude that the quantiles of the both drivers are the same. If the point is below the line, then the quantile is higher for Luigi, and if the point is above the line, then the quantile is lower for Luigi. So if all the points lie exactly on the line, we can conclude that the distributions of both the drivers are the same. We see that all the reported quantiles lie below the line, which implies that all the quantiles of Luigi have higher values than those of Domenico. This means that not only on an average, but also in general, the delivery times are higher for Luigi. If we look at two other drivers, as displayed in Fig. 3.2b, the points lie very much on the bisection line. We can therefore conclude that the delivery times of these two drivers do not differ much.

In *R*, we can generate QQ-plots by using the `qqplot` command:

```
qqplot()
```

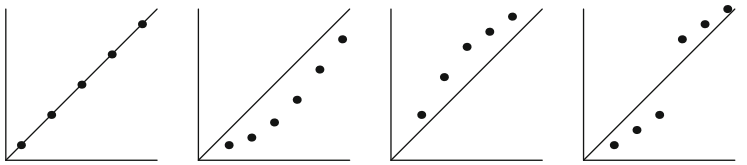



**Fig. 3.3** Different patterns for a QQ-plot

As a summary, let us consider four important patterns:

(a) If all the pairs of quantiles lie (nearly) on a straight line at an angle of 45 % from the $x$-axis, then the two samples have similar distributions (Fig. 3.3a).
(b) If the $y$-quantiles are lower than the $x$-quantiles, then the $y$-values have a tendency to be lower than the $x$-values (Fig. 3.3b).
(c) If the $x$-quantiles are lower than the $y$-quantiles, then the $x$-values have a tendency to be lower than the $y$-values (Fig. 3.3c).
(d) If the QQ-plot is like Fig. 3.3d, it indicates that there is a break point up to which the $y$-quantiles are lower than the $x$-quantiles and after that point, the $y$-quantiles are higher than the $x$-quantiles.

### 3.1.4 Mode

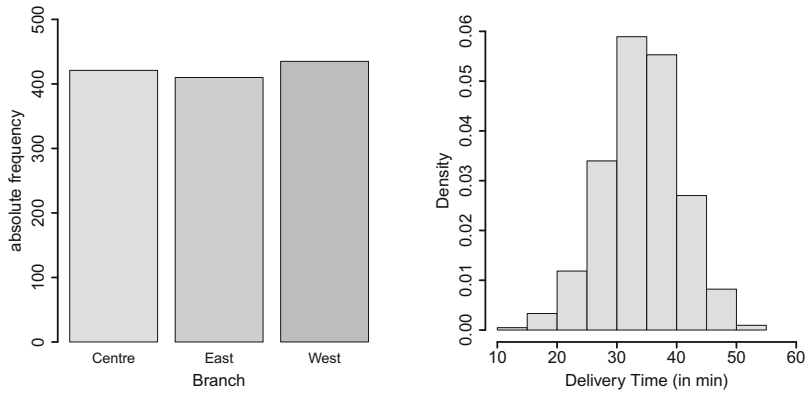
Consider a situation in which an ice cream shop owner wants to know which flavour of ice cream is the most popular among his customers. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand. To answer this type of questions, one can use the mode which is another measure of central tendency.

The mode $\bar{x}_M$ of $n$ observations $x_1, x_2, \ldots, x_n$ is the value which occurs the most compared with all other values, i.e. the value which has maximum absolute frequency. It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined. A formal definition of the mode is

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, n_2, \ldots, n_k\}. \tag{3.9}$$

The mode is typically applied to any type of variable for which the number of different values is not too large. If continuous data is summarized in groups, then the mode can be used as well.

*Example 3.1.7* Recall the pizza data set described in Appendix A.4. The pizza delivery service has three branches, in the East, West, and Centre, respectively. Suppose we want to know which branch delivers the most pizzas. We find that most of the deliveries have been made in the West, see Fig. 3.4a; therefore the mode is $\bar{x}_M =$ West. Similarly, suppose we also want to find the mode for the categorized pizza delivery time: if we group the delivery time in intervals of 5 min, then we see that the most frequent delivery time is the interval "30−35" min, see Fig. 3.4b. The mode is therefore $\bar{x}_M = [30, 35)$.




(a) Bar chart for branch (b) Histogram for grouped delivery time

**Fig. 3.4** Results from the pizza data set

### 3.1.5 Geometric Mean

Consider $n$ observations $x_1, x_2, \ldots, x_n$ which are all positive and collected on a quantitative variable. The geometric mean $\bar{x}_G$ of this data is defined as

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}. \tag{3.10}$$

The geometric mean plays an important role in fields where we are interested in products of observations, such as when we look at percentage changes in quantities. We illustrate its interpretation and use by looking at the average growth of a quantity in the sense that we allow a starting value, such as a certain amount of money or a particular population, to change over time. Suppose we have a starting value at some baseline time point 0 (zero), which may be denoted as $B_0$. At time $t$, this value may have changed and we therefore denote it as $B_t$, $t = 1, 2, \ldots, T$. The ratio of $B_t$ and $B_{t-1}$,

$$x_t = \frac{B_t}{B_{t-1}},$$

is called the $t$th growth factor. The growth rate $r_t$ is defined as

$$r_t = ((x_t - 1) \cdot 100)\,\%$$

and gives us an idea about the growth or decline of our value at time $t$. We can summarize these concepts in the following table:

| Time $t$ | Inventory $B_t$ | Growth factor $x_t$ | Growth rate $r_t$ |
|---|---|---|---|
| 0 | $B_0$ | – | – |
| 1 | $B_1$ | $x_1 = B_1/B_0$ | $((x_1 - 1) \cdot 100)\,\%$ |
| 2 | $B_2$ | $x_2 = B_2/B_1$ | $((x_2 - 1) \cdot 100)\,\%$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $T$ | $B_T$ | $x_T = B_T/B_{T-1}$ | $((x_T - 1) \cdot 100)\,\%$ |

We can calculate $B_t$ $(t = 1, 2, \ldots, T)$ by using the growth factors:

$$B_t = B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_t.$$

The average growth factor from $B_0$ to $B_T$ is the geometric mean or geometric average of the growth factors:

$$\begin{aligned} \bar{x}_G &= \sqrt[T]{x_1 \cdot x_2 \cdot \ldots \cdot x_T} \\ &= \sqrt[T]{\frac{B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_T}{B_0}} \\ &= \sqrt[T]{\frac{B_T}{B_0}}. \end{aligned} \tag{3.11}$$

Therefore, $B_t$ at time $t$ can be calculated as $B_t = B_0 \cdot \bar{x}_G^t$.

*Example 3.1.8* Suppose someone wants to deposit money, say €1000, in a bank. The bank advisor proposes a 5-year savings plan with the following plan for interest rates: 1 % in the first year, 1.5 % in the second year, 2.5 % in the third year, and 3 % in the last 2 years. Now he would like to calculate the average growth factor and average growth rate for the invested money. The concept of the geometric mean can be used as follows:

| Year | Euro | Growth factor | Growth rate (%) |
|---|---|---|---|
| 0 | 1000 | – | – |
| 1 | 1010 | 1.01 | 1.0 |
| 2 | 1025.15 | 1.015 | 1.5 |
| 3 | 1050.78 | 1.025 | 2.5 |
| 4 | 1082.30 | 1.03 | 3.0 |
| 5 | 1114.77 | 1.03 | 3.0 |

The geometric mean is calculated as

$$\bar{x}_G = (1.01 \cdot 1.015 \cdot 1.025 \cdot 1.03 \cdot 1.03)^{\frac{1}{5}} = 1.021968$$

which means that he will have on average about 2.2 % growth per year. The savings after 5 years can be calculated as

$$€\,1000 \cdot 1.021968^5 = €\,1114.77.$$

It is easy to compare two different saving plans with different growth strategies using the geometric mean.

### 3.1.6 Harmonic Mean

The harmonic mean is typically used whenever different $x_i$ contribute to the mean with a different weight $w_i$, i.e. when we implicitly assume that the weight of each $x_i$ is not one. It can be calculated as

$$\bar{x}_H = \frac{w_1 + w_2 + \cdots + w_k}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_k}{x_k}} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} \frac{w_i}{x_i}}. \tag{3.12}$$

For example, when calculating the average speed, each weight relates to the relative distance travelled, $n_i/n$, with speed $x_i$. Using $w_i = n_i/n$ and $\sum_i w_i = \sum_i n_i/n = 1$, the harmonic mean can be written as

$$\bar{x}_H = \frac{1}{\sum_{i=1}^{k} \frac{w_i}{x_i}}. \tag{3.13}$$

*Example 3.1.9* Suppose an investor bought shares worth €1000 for two consecutive months. The price for a share was €50 in the first month and €200 in the second month. What is the average purchase price? The number of shares purchased in the first month is $1000/50 = 20$. The number of shares purchased in the second month is $1000/200 = 5$. The total number of shares purchased is thus $20 + 5 = 25$, and the total investment is €2000. It is evident that the average purchase price is $2000/25 = €80$. This is in fact the harmonic mean calculated as

$$\bar{x}_H = \frac{1}{\frac{0.5}{50} + \frac{0.5}{200}} = 80$$

because the weight of each purchase is $n_i/n = 1000/2000 = 0.5$. If the investment was €1200 in the first month and €800 in the second month, then we could use the harmonic mean with weights $1200/2000 = 0.6$ and $800/2000 = 0.4$, respectively, to obtain the results.

## 3.2 Measures of Dispersion

Measures of central tendency, as introduced earlier, give us an idea about the location where most of the data is concentrated. However, two different data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have different concentrations around the mean. In this case, the location measures may not be adequate enough to describe the distribution of the data. The concentration or dispersion of observations around any particular value is another property which characterizes the data and its distribution. We now introduce statistical methods which describe the **variability** or **dispersion** of data.

*Example 3.2.1* Suppose three students Christine, Andreas, and Sandro arrive at different times in the class to attend their lectures. Let us look at their arrival time in the class after or before the starting time of lecture, i.e. let us look how early or late they were (in minutes).

| Week | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Christine | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Andreas | −10 | +10 | −10 | +10 | −10 | +10 | −10 | +10 | −10 | +10 |
| Sandro | 3 | 5 | 6 | 2 | 4 | 6 | 8 | 4 | 5 | 7 |

We see that Christine always arrives on time (time difference of zero). Andreas arrives sometimes 10 min early and sometimes 10 min late. However, the arithmetic mean of both students is the same—on average, they both arrive on time! This interpretation is obviously not meaningful. The difference between both students is the variability in arrival times that cannot be measured with the mean or median. For this reason, we need to introduce measures of dispersion (variability). With the knowledge of both location and dispersion, we can give a much more nuanced comparison between the different arrival times. For example, consider the third student Sandro. He is always late; sometimes more, sometimes less. However, while on average he comes late, his behaviour is more predictable than that of Andreas. Both location and dispersion are needed to give a fair comparison.

*Example 3.2.2* Consider another example in which a supplier for the car industry needs to deliver 10 car doors with an exact width of 1.00 m. He supplies 5 doors with a width of 1.05 m and the remaining 5 doors with a width of 0.95 m. The arithmetic mean of all the 10 doors is 1.00 m. Based on the arithmetic mean, one may conclude that all the doors are good but the fact is that none of the doors are usable as they will not fit into the car. This knowledge can be summarized by a measure of dispersion.

The above examples highlight that the distribution of a variable needs to be characterized by a measure of dispersion in addition to a measure of location (central tendency). Now we introduce various measures of dispersion.

### 3.2.1 Range and Interquartile Range

Consider a variable $X$ with $n$ observations $x_1, x_2, \ldots, x_n$. Order these $n$ observations as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The range is a measure of dispersion defined as the difference between the maximum and minimum value of the data as

$$R = x_{(n)} - x_{(1)}. \tag{3.14}$$

The **interquartile range** is defined as the difference between the 75th and 25th quartiles as

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}. \tag{3.15}$$

It covers the centre of the distribution and contains 50 % of the observations.

*Remark 3.2.1* Note that the interquartile range is defined as the interval $[\tilde{x}_{0.25}; \tilde{x}_{0.75}]$ in some literature. However, in line with most of the statistical literature, we define the interquartile range to be a measure of dispersion, i.e. the difference between $\tilde{x}_{0.75}$ and $\tilde{x}_{0.25}$.

*Example 3.2.3* Recall Examples 3.0.1–3.1.5 where we looked at the temperature in Bangkok during December. The ordered values $x_{(i)}$, $i = 1, \ldots, 31$, are as follows:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| °C | 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| (i) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| °C | 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| (i) | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

We obtained the quantiles in Example 3.1.5 as $\tilde{x}_{0.25} = 25$ and $\tilde{x}_{0.75} = 29$. The interquartile range is therefore $d_Q = 29 - 25 = 4$, which means that 50 % of the data is centred between 25 and 29 °C. The range is $R = 31 - 21 = 10$ °C, meaning that the temperature is varying at most by 10 °C. In *R*, there are several ways to obtain quartiles, minimum and maximum values, e.g. by using `min`, `max`, `quantiles`, `range`, among others. All numbers can be easily obtained by the `summary` command which we recommend using.

```
summary(weather)
```

### 3.2.2 Absolute Deviation, Variance, and Standard Deviation

Another measure of dispersion is the variance. The variance is one of the most important measures in statistics and is needed throughout this book. We use the idea of "absolute deviation" to give some more background and motivation for understanding the variance as a measure of dispersion, followed by some examples.

Consider the deviations of $n$ observations around a certain value "$A$" and combine them together, for instance, via the arithmetic mean of all the deviations:

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - A). \tag{3.16}$$

This measure has the drawback that the deviations $(x_i - A)$, $i = 1, 2, \ldots, n$, can be either positive or negative and, consequently, their sum can potentially be very small or even zero. Using $D$ as a measure of variability is therefore not a good idea since $D$ may be small even for a large variability in the data.

Using absolute values of the deviations solves this problem, and we introduce the following measure of dispersion:

$$D(A) = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|. \tag{3.17}$$

It can be shown that the absolute deviation attains its minimum when $A$ corresponds to the median of the data:

$$D(\tilde{x}_{0.5}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}|. \tag{3.18}$$

We call $D(\tilde{x}_{0.5})$ the **absolute median deviation**. When $A = \bar{x}$, we speak of the **absolute mean deviation** given by

$$D(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|. \tag{3.19}$$

Another solution to avoid the positive and negative signs of deviation in (3.16) is to consider the squares of deviations $x_i - A$, rather than using the absolute value. This provides another measure of dispersion as

$$s^2(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2 \tag{3.20}$$

which is known as the **mean squared error** (MSE) with respect to $A$. The MSE is another important measure in statistics, see Chap. 9, Eq. (9.4), for details. It can be shown that $s^2(A)$ attains its minimum value when $A = \bar{x}$. This is the (sample) **variance**

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

After expanding $\tilde{s}^2$, we can write (3.21) as

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2. \tag{3.22}$$

The positive square root of the variance is called the (sample) **standard deviation**, defined as

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3.23}$$

The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations. For example, if $X$ is weight, measured in kg, then $\bar{x}$ and $\tilde{s}$ are also measured in kg, while $\tilde{s}^2$ is measured in $\text{kg}^2$ (which may be more difficult to interpret). The variance is a measure which we use in other chapters to obtain measures of association between variables and to

draw conclusions from a sample about a population of interest; however, the standard deviation is typically preferred for a descriptive summary of the dispersion of data.

The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean. A low value of the standard deviation indicates that the values are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean. If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median. In such a case, the absolute median deviation (3.18) may be preferred over the standard deviation.

*Example 3.2.4* Consider again Example 3.2.1 where we evaluated the arrival times of Christine, Andreas, and Sandro in their lecture. Using the arithmetic mean, we concluded that both Andreas and Christine arrive on time, whereas Sandro is always late; however, we saw that the variation of arrival times differs substantially among the three students. To describe and quantify this variability formally, we calculate the variance and absolute median deviation:

$$\tilde{s}_C^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((0-0)^2 + \cdots + (0-0)^2) = 0$$

$$\tilde{s}_A^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((-10-0)^2 + \cdots + (10-0)^2) \approx 111.1$$

$$\tilde{s}_S^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((3-5)^2 + \cdots + (7-5)^2) \approx 3.3$$

$$D(\tilde{x}_{0.5,C}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |0-0| + \cdots + |0-0| = 0$$

$$D(\tilde{x}_{0.5,A}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |-10-0| + \cdots + |10-0| = 10$$

$$D(\tilde{x}_{0.5,S}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |3-5| + \cdots + |7-5| = 1.4.$$

We observe that the variation/dispersion/variability is the lowest for Christine and highest for Andreas. Both median absolute deviation and variance allow a comparison between the two students. If we take the square root of the variance, we obtain the standard deviation. For example, $\tilde{s}_S = \sqrt{3.3} \approx 1.8$, which means that the average difference of the observations from the arithmetic mean is 1.8.

In *R*, we can use the `var` command to calculate the variance. However, note that *R* uses $1/(n-1)$ instead of $1/n$ in calculating the variance. The idea behind the multiplication by $1/(n-1)$ in place of $1/n$ is discussed in Chap. 9, see also Theorem 9.2.1.

**Variance for Grouped Data.** The variance for grouped data can be calculated using

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(a_j - \bar{x})^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j a_j^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{j=1}^{k} n_j a_j^2 - \bar{x}^2, \tag{3.24}$$

where $a_j$ is the middle value of the $j$th interval. However, when the data is artificially grouped and the knowledge about the original ungrouped data is available, we can also use the arithmetic mean of the $j$th class:

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2. \tag{3.25}$$

The two expressions (3.24) and (3.25) represent the **variance between the different classes**, i.e. they describe the variability of the class specific means $\bar{x}_j$, weighted by the size of each class $n_j$, around the overall mean $\bar{x}$. It is evident that the variance *within* each class is not taken into account in these formulae. The variability of measurements in each class, i.e. the variability of $\forall x_i \in K_j$, is another important component to determine the overall variance in the data. It is therefore not surprising that using only the between variance $\tilde{s}_b^2$ will underestimate the total variance and therefore

$$s_b^2 \leq s^2. \tag{3.26}$$

If the data within each class is known, we can use the Theorem of Variance Decomposition (see p. 136 for the theoretical background) to determine the variance. This allows us to represent the total variance as the sum of the **variance between the different classes** and the **variance within the different classes** as

$$\tilde{s}^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j \tilde{s}_j^2}_{\text{within}}. \tag{3.27}$$

In (3.27), $\tilde{s}_j^2$ is the variance of the $j$th class:

$$\tilde{s}_j^2 = \frac{1}{n_j}\sum_{x_i \in K_j} (x_i - \bar{x}_j)^2. \tag{3.28}$$

The proof of (3.27) is given in Appendix C.1, p. 423.

*Example 3.2.5* Recall the weather data used in Examples 3.0.1–3.2.3 and the grouped data specified as follows:

| Class intervals | <20 | (20–25] | (25, 30] | (30, 35] | >35 |
|---|---|---|---|---|---|
| $n_j$ | $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $\bar{x}_j$ | – | 23.83 | 28 | 31 | – |
| $\tilde{s}_j^2$ | – | 1.972 | 2 | 0 | – |

We know that $\bar{x} = 26.48$ and $n = 31$. The first step is to calculate the mean and variances in each class using (3.28). We then obtain $\bar{x}_j$ and $s_j^2$ as listed above. The within and between variances are as follows:

$$\frac{1}{n}\sum_{j=1}^{k} n_j \tilde{s}_j^2 = \frac{1}{31}(12 \cdot 1.972 + 18 \cdot 2 + 1 \cdot 0) \approx 1.925$$

$$\frac{1}{n}\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 = \frac{1}{31}(12 \cdot [23.83 - 26.48]^2 + 18 \cdot [28 - 26.48]^2 + 1 \cdot [31 - 26.48]^2) \approx 4.71.$$

The total variance is therefore $\tilde{s}^2 \approx 6.64$. Estimating the variance using all 31 observations would yield the same results. However, it becomes clear that without knowledge about the variance within each class, we cannot reliably estimate $\tilde{s}^2$. In the above example, the variance between the classes is 3 times lower than the total variance which is a serious underestimation.

**Linear Transformations.** Let us consider a linear transformation $y_i = a + bx_i$ $(b \neq 0)$ of the original data $x_i$, $(i = 1, 2, \ldots, n)$. We get the arithmetic mean of the transformed data as $\bar{y} = a + b\bar{x}$ and for the variance:

$$\tilde{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2 \tilde{s}_x^2. \tag{3.29}$$

*Example 3.2.6* Let $x_i, i = 1, 2, \ldots, n$, denote measurements on time. These data could have been recorded and analysed in hours, but we may be interested in a summary in minutes. We can make a linear transformation $y_i = 60\, x_i$. Then, $\bar{y} = 60\bar{x}$ and $\tilde{s}_y^2 = 60^2 \tilde{s}_x^2$. If the mean and variance of the $x_i$'s have already been obtained, then the mean and variance of the $y_i$'s can be obtained directly using these transformations.

**Standardization.** A variable is called standardized if its mean is zero and its variance is 1. Standardization can be achieved by using the following transformation:

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x} x_i = a + bx_i. \tag{3.30}$$

It follows that $\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})/\tilde{s}_x = 0$ and $\tilde{s}_y^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/\tilde{s}_x^2 = 1$. There are many statistical methods which require standardization, see, for example, Sect. 10.3.1 for details in the context of statistical tests.

*Example 3.2.7* Let $X$ be a variable which measures air pollution by using the concentration of atmospheric particulate matter (in $\mu g/m^3$). Suppose we have the following 10 measurements:

30 25 12 45 50 52 38 39 45 33.

We calculate $\bar{x} = 36.9$, $\tilde{s}_x^2 = 136.09$, and $\tilde{s}_x = 11.67$. To get a standardized variable $Y$, we transform all the observations $x_i$'s as

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x} x_i = -\frac{36.9}{11.67} + \frac{1}{11.67} x_i = -3.16 + 0.086 x_i .$$

Now $y_1 = -3.16 + 0.086 \cdot 30 = -0.58$, $y_2 = -3.16 + 0.086 \cdot 25 = -1.01$, ..., are the standardized observations. The `scale` command in *R* allows standardization, and we can obtain the standardized observations corresponding to the 10 measurements as

```
air <- c(30,25,12,45,50,52,38,39,45,33)
scale(air)
```

Please note that the `scale` command uses $1/(n-1)$ for calculating the variance, as already outlined above. Thus, the results provided by `scale` are not identical to those using (3.30).

### 3.2.3 Coefficient of Variation

Consider a situation where two different variables have arithmetic means $\bar{x}_1$ and $\bar{x}_2$ with standard deviations $\tilde{s}_1$ and $\tilde{s}_2$, respectively. Suppose we want to compare the variability of hotel prices in Munich (measured in euros) and London (measured in British pounds). How can we provide a fair comparison? Since the prices are measured in different units, and therefore likely have arithmetic means which differ substantially, it does not make much sense to compare the standard deviations directly. The coefficient of variation $v$ is a measure of dispersion which uses both the standard deviation and mean and thus allows a fair comparison. It is properly defined only when all the values of a variable are measured on a ratio scale and are positive such that $\bar{x} > 0$ holds. It is defined as

$$v = \frac{s}{\bar{x}}. \tag{3.31}$$

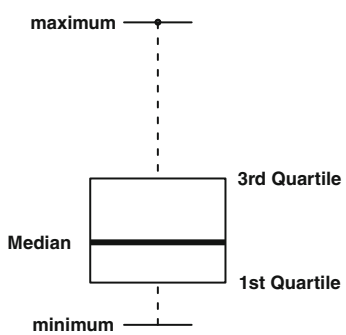
The coefficient of variation is a unit-free measure of dispersion. It is often used when the measurements of two variables are different but can be put into relation by using a linear transformation $y_i = bx_i$. It is possible to show that if all values $x_i$ of a variable $X$ are transformed into a variable $Y$ with values $y_i = b \cdot x_i$, $b > 0$, then $v$ does not change.

*Example 3.2.8* If we want to compare the variability of hotel prices in two selected cities in Germany and England, we could calculate the mean prices, together with their standard deviation. Suppose a sample of prices of say 100 hotels in two selected cities in Germany and England is available and suppose we obtain the mean and standard deviations of the two cities as $x_1 = €130$, $x_2 = £230$, $s_1 = €99$, and $s_2 = £212$. Then, $v_1 = 99/130 \approx 0.72$ and $v_2 = 212/230 = 0.92$. This indicates higher variability in hotel prices in England. However, if the data distribution is skewed or bimodal, then it may be wise not to choose the arithmetic mean as a measure of central tendency and likewise the coefficient of variation.

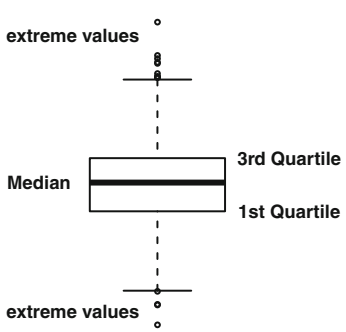
## 3.3 Box Plots

So far we have described various measures of central tendency and dispersion. It can be tedious to list those measures in summary tables. A simple and powerful graph is the **box plot** which summarizes the distribution of a continuous (or sometimes an ordinal) variable by using its median, quartiles, minimum, maximum, and extreme values.

Figure 3.5a shows a typical box plot. The vertical length of the box is the interquartile range $d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$, which shows the region that contains 50 % of the data. The bottom end of the box refers to the first quartile, and the top end of the box refers to the third quartile. The thick line in the box is the median. It becomes immediately clear that the box indicates the symmetry of the data: if the median is in the middle of the box, the data should be symmetric, otherwise it is skewed. The *whiskers* at the end of the plot mark the minimum and maximum values of the data. Looking at the box plot as a whole tells us about the data distribution and the range and variability of observations. Sometimes, it may be advisable to understand which values are extreme in the sense that they are "far away" from the centre of the distribution. In many software packages, including *R*, values are defined to be extreme if they are greater than 1.5 box lengths away from the first or third quartile. Sometimes, they are called outliers. Outliers and extreme values are defined differently in some software packages and books.




(a) Box plot without extreme values



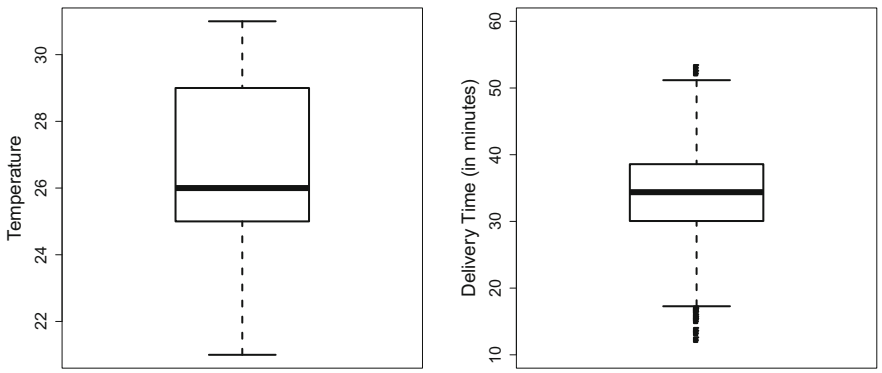

(b) Box plot with extreme values

The `boxplot` command in *R* draws a box plot. The `range` option controls whether extreme values should be plotted, and if yes, how one wants to define such values.

```
boxplot(variable, range=1.5)
```

*Example 3.3.1* Recall Examples 3.0.1–3.2.5 where we looked at the temperature in Bangkok during December. We have already calculated the median (26°C) and the quartiles (25, 29°C). The minimum and maximum values are 21°C and 31°C. The box plot for this data is shown in Fig. 3.5a. One can see that the temperature distribution is slightly skewed with more variability for lower temperatures. The interquartile range is 4, and therefore, any value $>29+4\times1.5=35$ or $<25-4\times1.5=19$ would be an extreme value. However, there are no extreme values in the data.

*Example 3.3.2* Consider again the pizza data described in Appendix A.4. We use *R* to plot the box plot for the delivery time via `boxplot(time)` (Fig. 3.5b). We see a symmetric distribution with a median delivery time of about 35 min. Most of the deliveries took between 30 and 40 min. The extreme values indicate that there were some exceptionally short and long delivery times.




(a) Boxplot for weather data (b) Boxplot for pizza data

## 3.4 Measures of Concentration

A completely different concept used to describe a quantitative variable is the idea of concentration. For a variable $X$, it summarizes the proportion of each observation with respect to the sum of all observations $\sum_{i=1}^{n} x_i$. Let us look at a simple example to demonstrate its usefulness.

**Table 3.1** Concentration of farmland: two different situations

| Farmer ($i$) | $x_i$ (Area, in hectare) |
|---|---|
| 1 | 20 |
| 2 | 20 |
| 3 | 20 |
| 4 | 20 |
| 5 | 20 |
| | $\sum_{i=1}^{5} x_i = 100$ |

| Farmer ($i$) | $x_i$ (Area, in hectare) |
|---|---|
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 100 |
| | $\sum_{i=1}^{5} x_i = 100$ |

*Example 3.4.1* Consider a village with 5 farms. Each farmer has a farm of a certain size. How can we evaluate the land distribution? Do all farmers have a similar amount of land or do one or two farmers have a big advantage because they have considerably more space?

Table 3.1 shows two different situations: in the table on the left, we see an equal distribution of land, i.e. each farmer owns 20 hectares of farmland. This means $X$ is *not* concentrated, rather it is equally distributed. A statistical function describing the concentration could return a value of zero in such a case. Consider another extreme where one farmer owns all the farmland and the others do not own anything, as shown on the right side of Table 3.1. This is an extreme concentration of land: one person owns everything and thus, we say the concentration is high. A statistical function describing the concentration could return a value of one in such a case.

### 3.4.1 Lorenz Curve

The **Lorenz curve** is a popular method to display concentrations graphically. Consider $n$ observations $x_1, x_2, \ldots, x_n$ of a variable $X$. Assume that all the observations are positive. The sum of all the observations is $\sum_{i=1}^{n} x_i = n\bar{x}$ if the data is ungrouped. First, we need to order the data: $0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. To plot the Lorenz curve, we need

$$u_i = \frac{i}{n}, \quad i = 0, \ldots, n, \tag{3.32}$$

and

$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \quad i = 1, \ldots, n; \; v_0 := 0, \tag{3.33}$$

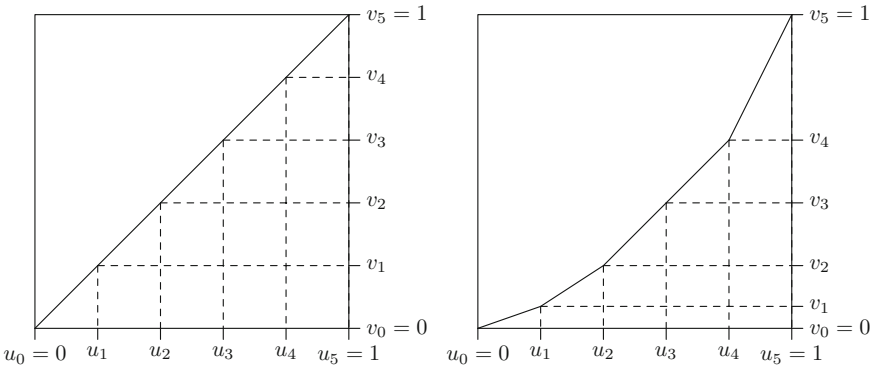



**Fig. 3.5** Lorenz curves for no concentration (*left*) and some concentration (*right*)*

where $\sum_{j=1}^{i} x_{(j)}$ is the cumulative total of observations up to the $i$th observation. The idea is that $v_i$ describe the contribution of all values $\leq i$ in comparison with the sum of all values. Plotting $u_i$ against $v_i$ for all $i$ shows how much the sum of all $x_i$, for all observations $\leq i$, contributes to the total sum. In other words, the point $(u_i, v_i)$ says that $u_i \cdot 100\,\%$ of observations contain $v_i \cdot 100\,\%$ of the sum of all $x_i$ less than or equal to $i$. Obviously, if all $x_i$ are identical, the Lorenz curve will be a straight diagonal line, also known as the identity line or **line of equality**. If the $x_i$ are of different sizes, then the Lorenz curve falls below the line of equality. This is illustrated in the following example.

*Example 3.4.2* Recall Example 3.4.1 where we looked at the distribution of farmland among 5 farmers. On the upper panel of Table 3.1, we observed an equal distribution of land among the farmers: $x_1 = 20$, $x_2 = 20$, $x_3 = 20$, $x_4 = 20$, and $x_5 = 20$. We obtain $u_1 = 1/5, u_2 = 2/5, \ldots, u_5 = 1$ and $v_1 = 20/100, v_2 = 40/100, \ldots, v_5 = 1$. This yields a Lorenz curve as displayed on the left side of Fig. 3.5: there is no concentration. We can interpret each point. For example, $(u_2, v_2) = (0.4, 0.4)$ means that 40 % of farmers own 40 % of the land.

The lower panel of Table 3.1 describes the situation with strong concentration. For this table, we obtain $u_1 = 1/5, u_2 = 2/5, \ldots, u_5 = 1$ and $v_1 = 0, v_2 = 0, \ldots, v_5 = 1$. Therefore, for example, 80 % of farmers own 0 % of the land which shows strong inequality. Most often we do not have such extreme situations. In this case, the Lorenz curve is bent towards the lower right corner of the plot, see the right side of Fig. 3.5.

We can plot the Lorenz curve in *R* using the `Lc` command in the library `ineq`. The Lorenz curve for the left table of Example 3.4.1 is plotted in *R* as follows:

```
library(ineq)
x <- c(20,20,20,20,20)
plot(Lc(x))
```

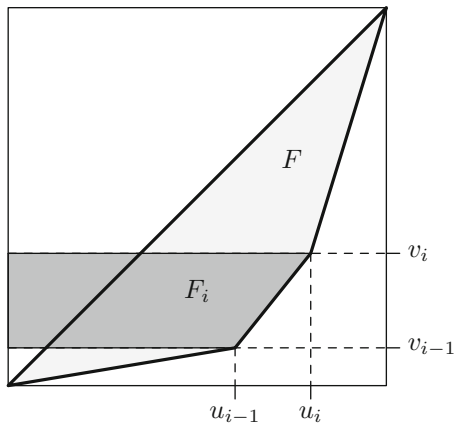



**Fig. 3.6** Lorenz curve and the Gini coefficient*

We can use the same approach as above to obtain the Lorenz curve when we have grouped data. We simply describe the contributions for each class rather than for each observation and approximate the values in each class by using its mid-point. More formally we can write:

$$\tilde{u}_i = \sum_{j=1}^{i} f_j, \quad i = 1, 2, \ldots, k; \ \tilde{u}_0 := 0 \tag{3.34}$$

and

$$\tilde{v}_i = \frac{\sum_{j=1}^{i} f_j a_j}{\sum_{j=1}^{k} f_j a_j} = \frac{\sum_{j=1}^{i} n_j a_j}{n\bar{x}}, \quad i = 1, 2, \ldots, k; \ \tilde{v}_0 := 0. \tag{3.35}$$

### 3.4.2 Gini Coefficient

We have seen in Sect. 3.4.1 that the Lorenz curve corresponds to the identity line, that is the diagonal line of equality, for no concentration. When there is some concentration, then the curve deviates from this line. The amount of deviation depends on the strength of concentration. Suppose we want to design a measure of concentration which is 0 for no concentration and 1 for perfect (i.e. extreme) concentration. We can simply measure the area between the Lorenz curve and the identity line and multiply it by 2. For no concentration, the area will be zero and hence the measure will be zero. If there is perfect concentration, then the curve will coincide with the axes, the area will be close to 0.5, and twice the area will be close to one. The measure based on such an approach is called the Gini coefficient:

$$G = 2 \cdot F. \tag{3.36}$$

Note that $F$ is the area between the curve and the bisection or diagonal line.

The Gini coefficient can be estimated by adding up the areas of the trapeziums $F_i$ as displayed in Fig. 3.6:

$$F = \sum_{i=1}^{n} F_i - 0.5,$$

where

$$F_i = \frac{u_{i-1} + u_i}{2}(v_i - v_{i-1}).$$

It can be shown that this corresponds to

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}(v_{i-1} + v_i), \tag{3.37}$$

but the proof is omitted. The same formula can be used for grouped data except that $\tilde{v}$ is used instead of $v$. Since

$$0 \leq G \leq \frac{n-1}{n}, \tag{3.38}$$

one may prefer to use the standardized Gini coefficient

$$G^+ = \frac{n}{n-1}G, \tag{3.39}$$

which takes a maximum value of 1.

*Example 3.4.3* We return to our farmland example. Suppose we have 7 farmers with farms of different sizes:

| Farmer | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Farmland size $x_i$ | 20 | 14 | 59 | 9 | 36 | 23 | 3 |

Using the ordered values, we can calculate $u_i$ and $v_i$ using (3.32) and (3.33):

| $i$ | $x_{(i)}$ | $u_i$ | $v_i$ |
|---|---|---|---|
| 1 | 3 | $\frac{1}{7} = 0.1429$ | $\frac{3}{164} = 0.0183$ |
| 2 | 9 | $\frac{2}{7} = 0.2857$ | $\frac{12}{164} = 0.0732$ |
| 3 | 14 | $\frac{3}{7} = 0.4286$ | $\frac{26}{164} = 0.1585$ |
| 4 | 20 | $\frac{4}{7} = 0.5714$ | $\frac{46}{164} = 0.2805$ |
| 5 | 23 | $\frac{5}{7} = 0.7143$ | $\frac{69}{164} = 0.4207$ |
| 6 | 36 | $\frac{6}{7} = 0.8571$ | $\frac{105}{164} = 0.6402$ |
| 7 | 59 | $\frac{7}{7} = 1.0000$ | $\frac{164}{164} = 1.0000$ |

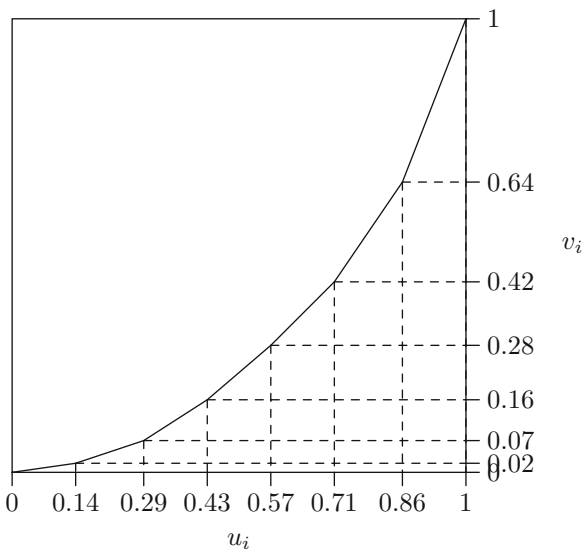



**Fig. 3.7** Lorenz curve for Example 3.4.3*

The Lorenz curve is displayed in Fig. 3.7. Using this information, it is easy to calculate the Gini coefficient:

$$G = 1 - \frac{1}{7}(0.0183 + [0.0183 + 0.0732] + [0.0732 + 0.1585] + [0.1585 + 0.2805] + [0.2805 + 0.4207] + [0.4207 + 0.6402] + [0.6402 + 1]) = 0.402$$

We know that $G = 0.4024 \leq \frac{6}{7} = \frac{n-1}{n}$. To standardize the coefficient, we therefore have to use (3.39):

$$G^+ = \frac{7}{6}G = \frac{7}{6} \cdot 0.4024 = 0.4695\,.$$

In *R*, we can obtain the non-standardized Gini Coefficient using the `ineq` function in the library `ineq`.

```
library(ineq)
farm <- c(20,14,59,9,36,23,3)
ineq(farm)
```

## 3.5 Key Points and Further Issues

**Note:**

- ✓ A summary on how to descriptively summarize data is given in Appendix D.1.
- ✓ The median is preferred over the arithmetic mean when the data distribution is skewed or there are extreme values.
- ✓ If data of a continuous variable is grouped, and the original ungrouped data is not known, additional assumptions are needed to calculate measures of central tendency and dispersion. However, in some cases, these assumptions may not be satisfied, and the formulae provided may give imprecise results.
- ✓ QQ-plots are not only descriptive summaries but can also be used to test modelling assumptions, see Chap. 11.9 for more details.
- ✓ The distribution of a continuous variable can be easily summarized using a box plot.

## 3.6 Exercises

*Exercise 3.1* A hiking enthusiast has a new app for his smartphone which summarizes his hikes by using a GPS device. Let us look at the distance hiked (in km) and maximum altitude (in m) for the last 10 hikes:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Distance | 12.5 | 29.9 | 14.8 | 18.7 | 7.6 | 16.2 | 16.5 | 27.4 | 12.1 | 17.5 |
| Altitude | 342 | 1245 | 502 | 555 | 398 | 670 | 796 | 912 | 238 | 466 |

(a) Calculate the arithmetic mean and median for both distance and altitude.
(b) Determine the first and third quartiles for both the distance and the altitude variables. Discuss the shape of the distribution given the results of (a) and (b).
(c) Calculate the interquartile range, absolute median deviation, and standard deviation for both variables. What is your conclusion about the variability of the data?
(d) One metre corresponds to approximately 3.28 ft. What is the average altitude when measured in feet rather than in metres?
(e) Draw and interpret the box plot for both distance and altitude.
(f) Assume distance is measured as only short (5–15 km), moderate (15–20 km), and long (20–30 km). Summarize the grouped data in a frequency table. Calculate the weighted arithmetic mean under the assumption that the raw data is not

known. Determine the weighted median under the assumption that the values within each class are equally distributed.

(g) What is the variance for the grouped data when the raw data is known, i.e. when one has knowledge about the variance in each class? How does it compare with the variance one obtains when the raw data is unknown?

(h) Use *R* to reproduce the results of (a), (b), (c), (e), and (f).

*Exercise 3.2* A gambler notes down his wins and losses (in €) from playing 10 games of roulette in a casino.

| Round | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Won/Lost | 200 | 600 | −200 | −200 | −200 | −100 | −100 | −400 | 0 | |

(a) Assume $\bar{x} = -$ €90 and $s =$ €294.7881. What is the result of round 10?

(b) Determine the mode and the interquartile range.

(c) A different gambler plays 33 rounds of roulette. His results are $\bar{x} =$ €12 and $s =$ €1000. Is it meaningful to compare the variability of results of the two players by using the coefficient of variation? If yes, determine the coefficients of variation; if no, why is a comparison not possible?

*Exercise 3.3* A fashion boutique has summarized its daily sales of designer socks in different groups: men's socks, women's socks, and children's socks. Unfortunately, the data for men's socks was lost. Determine the missing values.

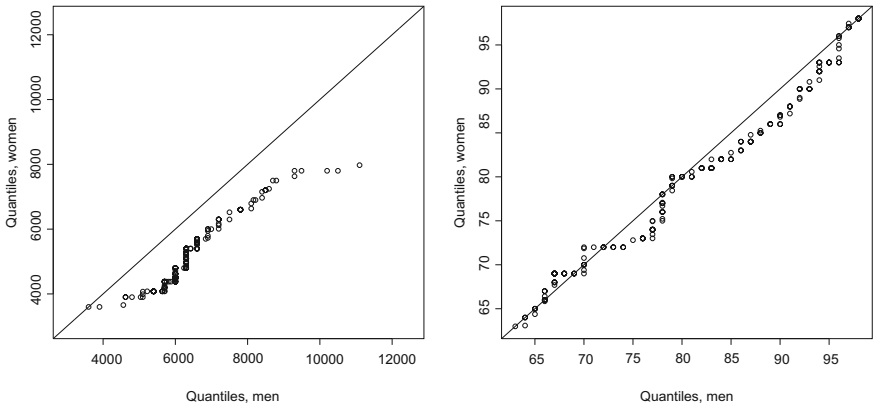
| | $n$ | Arithmetic mean in € | Standard deviation in € |
|---|---|---|---|
| Women's wear | 45 | 16 | $\sqrt{6}$ |
| Men's wear | ? | ? | ? |
| Children's wear | 20 | 7.5 | $\sqrt{3}$ |
| Total | 100 | 15 | $\sqrt{19.55}$ |

*Exercise 3.4* The number of members of a millionaires' club were as follows:

| Year | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|---|
| Members | 23 | 24 | 27 | 25 | 30 | 28 |

(a) What is the average growth rate of the membership?

(b) Based on the results of (a), how many members would one expect in 2018?




(a) for the salary (b) for length of service

**Fig. 3.8** QQ-plots

(c) The president of the club is interested in the number of members in 2025, the year when his presidency ends. Would it make sense to predict the number of members for 2025?

In 2015, the members invested €250 million on the stock market. 10 members contributed 16% of the investment sum, 8 members contributed €60 million, 8 members contributed €70 million, and another 4 members contributed the remaining amount.

(d) Draw the Lorenz curve for this data.
(e) Calculate and interpret the standardized Gini coefficient.

*Exercise 3.5* Consider the monthly salaries $Y$ (in Swiss francs) of a well-reputed software company, as well as the length of service (in months, $X$), and gender ($Z$). Figure 3.8 shows the QQ-plots for both $Y$ and $X$ given $Z$. Interpret both graphs.

*Exercise 3.6* There is no built-in function in *R* to calculate the mode of a variable. Program such a function yourself. Hint: type `?table` and `?names` to recall the functionality of these functions. Combine them in an intelligent way.

*Exercise 3.7* Consider a country in which 90 % of the wealth is owned by 20 % of the population, the so-called upper class. For simplicity, let us assume that the wealth is distributed equally within this class.

(a) Draw the Lorenz curve for this country.
(b) Now assume a revolution takes place in the country and all members of the upper class have to give away their wealth which is then distributed equally across the remaining population. Draw the Lorenz curve for this scenario.
(c) What would the curve from (b) look like if the entire upper class left the country?

*Exercise 3.8* A bus route in the mountainous regions of Romania has a length of 418 km. The manager of the bus company serving the route wants his buses to finish a trip within 8 h. The bus travels the first 180 km with an average speed of 48 km/h, the next 117 km with an average speed of 37 km/h, and the last section with an average speed of 52 km/h.

(a) What is the average speed with which the bus travels?
(b) Will the bus finish the trip in time?

*Exercise 3.9* Four friends have a start-up company which sells vegan ice cream. Their initial financial contributions are as follows:

| Person | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Contribution (in €) | 800 | 10300 | 4700 | 2220 |

(a) Calculate and draw the Lorenz curve.
(b) Determine and interpret the standardized Gini coefficient.
(c) Does $G^+$ change if each of the friends contributes only half the amount of money? If yes, how much? If no, why not?
(d) Use *R* to draw the above Lorenz curve and to calculate the Gini coefficient.

*Exercise 3.10* Recall the pizza delivery data which is described in Appendix A.4. Use *R* to read in and analyse the data.

(a) Calculate the mean, median, minimum, maximum, first quartile, and third quartile for all quantitative variables.
(b) Determine and interpret the 99 % quantile for delivery time and temperature.
(c) Write a function which calculates the absolute mean deviation. Use the function to calculate the absolute mean deviation of temperature.
(d) Scale the delivery time and calculate the mean and variance for this variable.
(e) Draw a box plot for delivery time and temperature. The box plots should not highlight extreme values.
(f) Use the `cut` command to create a new variable which summarizes delivery time in steps of 10 min. Calculate the arithmetic mean of this variable.
(g) Reproduce the QQ-plots shown in Example 3.1.6.

→ Solutions to all exercises in this chapter can be found on p. 333

*Source* Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg