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 jth class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. The mid-value of the jth 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:

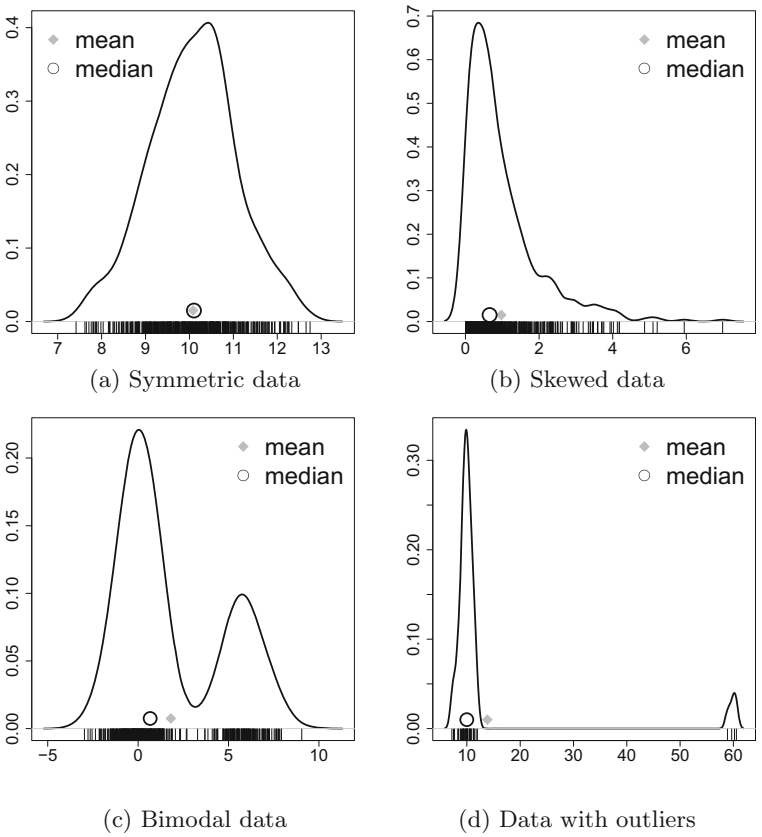
Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
n_j	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
f_j	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$
$\sum_j f_j$	0	$\frac{12}{31}$	$\frac{30}{31}$	1	1

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 α 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 α 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 α 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 α 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	

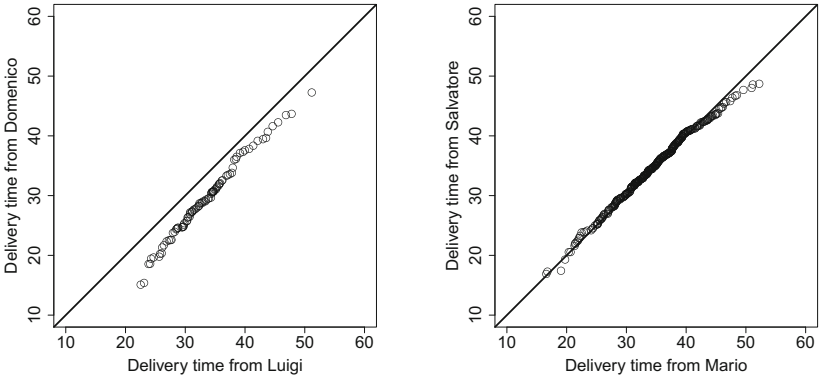
To determine the quartiles, i.e. the 25, 50, and 75 % quantiles, we calculate $n\alpha$ as $31 \cdot 0.25 = 7.75$, $31 \cdot 0.5 = 15.5$, and $31 \cdot 0.75 = 23.25$. Using (3.8), it follows that

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

In *R*, we obtain the same results using the `quantile` function. The `probs` argument is used to specify α. 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()
```

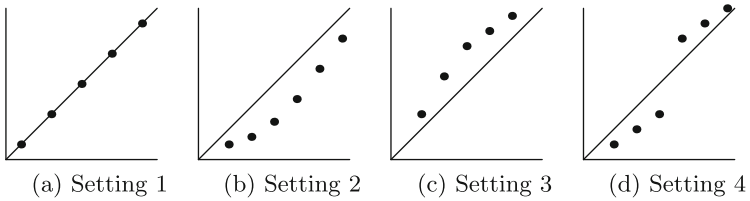


Fig. 3.3 Different patterns for a QQ-plot

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

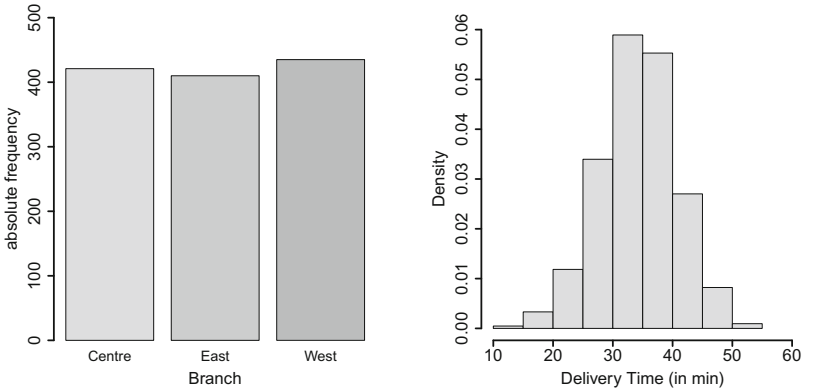
Consider a situation in which an ice cream shop owner wants to know which flavour of ice cream is the most popular among his customers. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand. To answer this type of questions, one can use the mode which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

$$30 \quad 25 \quad 12 \quad 45 \quad 50 \quad 52 \quad 38 \quad 39 \quad 45 \quad 33.$$

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

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

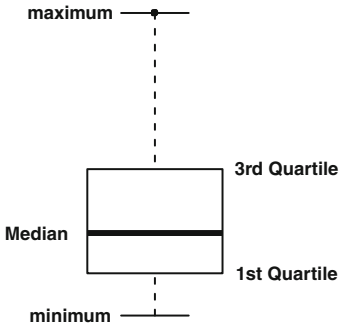
The coefficient of variation is a unit-free measure of dispersion. It is often used when the measurements of two variables are different but can be put into relation by using a linear transformation $y_i = bx_i$. It is possible to show that if all values x_i of a variable X are transformed into a variable 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.

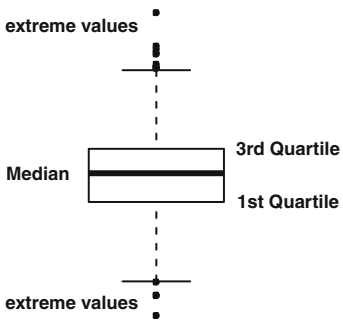
3.3 Box Plots

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



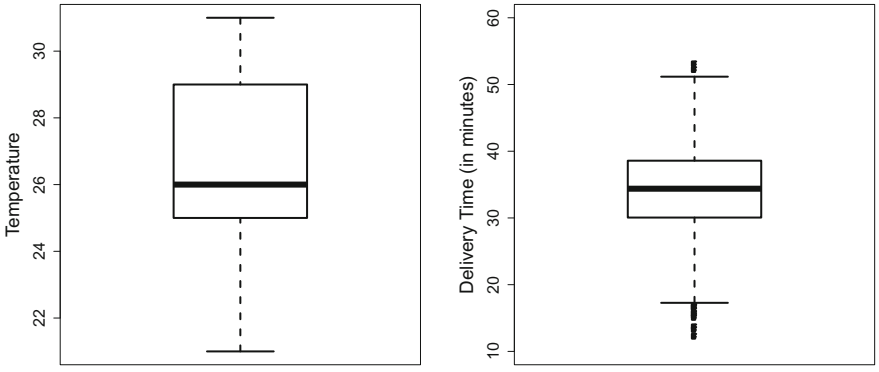
(b) Box plot with extreme values

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 $>29 + 4 \times 1.5 = 35$ or $<25 - 4 \times 1.5 = 19$ 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

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

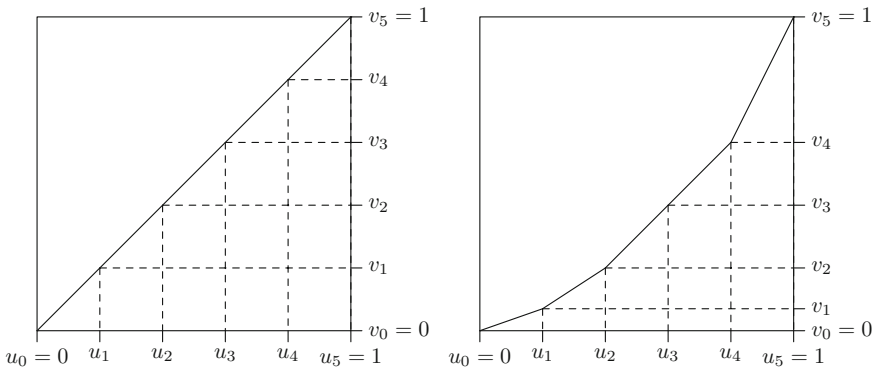


Fig. 3.5 Lorenz curves for no concentration (*left*) and some concentration (*right*)*

where $\sum_{j=1}^{i} x_{(j)}$ is the cumulative total of observations up to the ith 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))
```

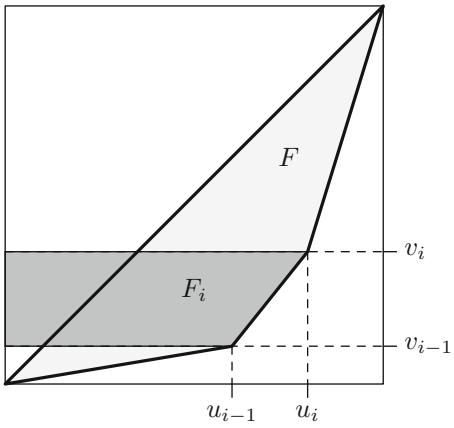


Fig. 3.6 Lorenz curve and the Gini coefficient*

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	$\frac{7}{7} = 1.0000$	$\frac{164}{164} = 1.0000$

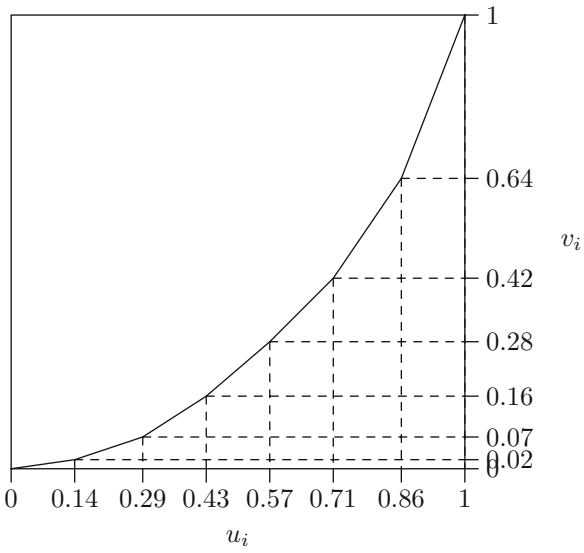


Fig. 3.7 Lorenz curve for Example 3.4.3*

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.

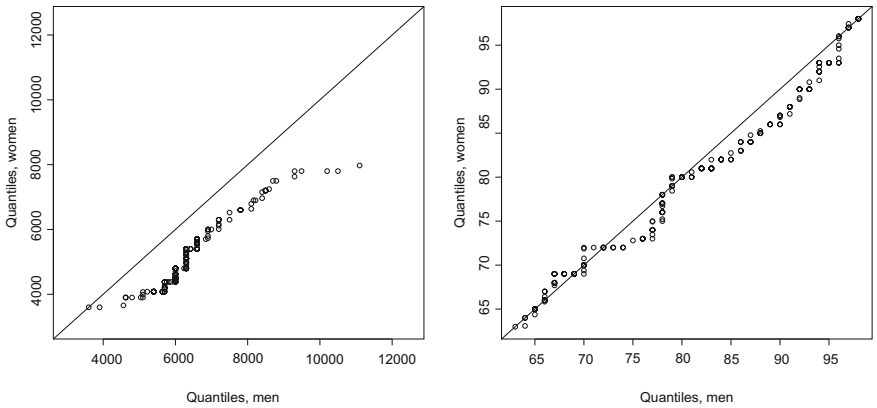
	n	Arithmetic mean in €	Standard deviation in €
Women's wear	45	16	$\sqrt{6}$
Men's wear	?	?	?
Children's wear	20	7.5	$\sqrt{3}$
Total	100	15	$\sqrt{19.55}$

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

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

Association of Two Variables

4

In Chaps. 2 and 3 we discussed how to analyse a single variable using graphs and summary statistics. However, in many situations we may be interested in the interdependence of two or more variables. For example, suppose we want to know whether male and female students in a college have any preference between the subjects mathematics and biology, i.e. if there is any evidence that male students prefer mathematics over biology and female students prefer biology over mathematics or vice versa. Suppose we choose an equal number of male and female students and ask them about their preferred subject. We expect that if there is no association between the two variables "gender of student" (male or female) and "subject" (mathematics or biology), then an equal proportion of male and female students should choose the subjects biology and mathematics, respectively. Any difference in the proportions may indicate a preference of males or females for a particular subject. Similarly, in another example, we may want to find out whether female employees of an organization are paid less than male employees or vice versa. Let us assume again that we choose an equal number of male and female employees and assume further that the salary is measured as a binary variable (low- versus high-salary group). We then expect that if there is no gender discrimination, the number of male and female employees in the lower- and higher-salary groups in the organization should be approximately equal. In both examples, the variables considered are binary and nominal (although the salary can also be seen as ordinal) and the data is summarized in terms of frequency measures. There may, however, be situations in which we are interested in associations between ordinal or continuous variables. Consider a data set in which height, weight, and age of infants are given. Usually, the height and weight of infants increase with age. Also, the height of infants increases with their weight and vice versa. Clearly, there is an interrelation or association among the three variables. In another example, two persons have to judge participants of a dance competition and rank them according to their performance. Now if we want to learn about the fairness in the judgment, we expect that both the judges give similar ranks to each candidate,

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_4

i.e. both judges give high ranks to good candidates and low ranks to not so good candidates. We are therefore interested in studying the association between the ranks given by the two judges. In all these examples, the intention lies in measuring the degree of association between two (or more) variables. We therefore need to study different ways of measuring the association level for different types of variables. In this chapter, we present measures and graphical summaries for the association of two variables—dependent on their scale.

4.1 Summarizing the Distribution of Two Discrete Variables

When both variables are discrete, then it is possible to list all combinations of values of the two variables and to count how often these combinations occur in the data. Consider the salary example in the introduction to this chapter in which both the variables were binary. There are four possible combinations of variable categories (female and low-salary group, female and high-salary group, male and low-salary group, and male and high-salary group). A complete description of the joint occurrence of these two variables can be given by counting, for each combination, the number of units for which this combination is measured. In the following, we generalize this concept to two variables where each can have an arbitrary (but fixed) number of values or categories.

4.1.1 Contingency Tables for Discrete Data

Suppose we have data on two discrete variables. This data can be described in a two-dimensional **contingency table**.

Example 4.1.1 An airline conducts a customer satisfaction survey. The survey includes questions about travel class and satisfaction levels with respect to different categories such as seat comfort, in-flight service, meals, safety, and other indicators. Consider the information on X, denoting the travel class (Economy = "E", Business = "B", First = "F"), and "Y", denoting the overall satisfaction with the flight on a scale from 1 to 4 as 1 (poor), 2 (fair), 3 (good), and 4 (very good). A possible response from 12 customers may look as follows:

	Passenger number											
i	1	2	3	4	5	6	7	8	9	10	11	12
Travel class	E	E	E	B	E	B	F	E	E	B	E	B
Satisfaction	2	4	1	3	1	2	4	3	2	4	3	3

We can calculate the absolute frequencies for each of the combination of observed values. For example, there are 2 passengers (passenger numbers 3 and 5) who were

Table 4.1 Contingency table for travel class and satisfaction

		Overall rating of flight quality				Total (row)
		Poor	Fair	Good	Very good	
Travel class	Economy business first	2 0 0	2 1 0	2 2 0	1 1 1	7 4 1
	Total (column)	2	3	4	3	12

flying in economy class and rated the flight quality as poor, there were no passengers from both business class and first class who rated the flight quality as poor; there were 2 passengers who were flying in economy class and rated the quality as fair (2), and so on. Table 4.1 is a two-dimensional table summarizing this information.

Note that we not only summarize the joint frequency distribution of the two variables but also the distributions of the individual variables. Summing up the rows and columns of the table gives the respective frequency distributions. For example, the last column of the table demonstrates that 7 passengers were flying in economy class, 4 passengers were flying in business class and 1 passenger in first class.

Now we extend this example and discuss a general framework to summarize the absolute frequencies of two discrete variables in contingency tables. We use the following notations: Let $x_1, x_2, \ldots, x_k$ be the k classes of a variable X and let $y_1, y_2, \ldots, y_l$ be the l classes of another variable Y. We assume that both X and Y are discrete variables. It is possible to summarize the absolute frequencies n_{ij} related to (x_i, y_j), $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$, in a $k \times l$ **contingency table** as shown in Table 4.2.

Table 4.2 $k \times l$ contingency table

		Y					
		y_1		y_j		y_l	Total (rows)
X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

We denote the sum of the ith row as $n_{i+} = \sum_{j=1}^{l} n_{ij}$ and the sum over the jth column as $n_{+j} = \sum_{i=1}^{k} n_{ij}$. The total number of observations is therefore

$$n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \,. \tag{4.1}$$

Remark 4.1.1 Note that it is also possible to use the relative frequencies $f_{ij} = n_{ij}/n$ instead of the absolute frequencies n_{ij} in Table 4.2, see Example 4.1.2.

4.1.2 Joint, Marginal, and Conditional Frequency Distributions

When the data on two variables are summarized in a contingency table, there are several concepts which can help us in studying the characteristics of the data. For example, how the values of both the variables behave jointly, how the values of one variable behave when another variable is kept fixed etc. These features can be studied using the concepts of joint frequency distribution, marginal frequency distribution, and conditional frequency distribution. If relative frequency is used instead of absolute frequency, then we speak of the joint relative frequency distribution, marginal relative frequency distribution, and conditional relative frequency distribution.

Definition 4.1.1 Using the notations of Table 4.2, we define the following:

The frequencies n_{ij} represent the **joint frequency distribution** of X and Y.

The frequencies n_{i+} represent the **marginal frequency distribution** of X.

The frequencies n_{+j} represent the **marginal frequency distribution** of Y.

We define $f_{i|j}^{X|Y} = n_{ij}/n_{+j}$ to be the **conditional frequency distribution** of X given $Y = y_j$.

We define $f_{j|i}^{Y|X} = n_{ij}/n_{i+}$ to be the **conditional frequency distribution** of Y given $X = x_i$

The frequencies f_{ij} represent the **joint relative frequency distribution** of X and Y.

The frequencies $f_{i+} = \sum_{j=1}^{l} f_{ij}$ represent the **marginal relative frequency distribution** of X.

The frequencies $f_{+j} = \sum_{i=1}^{k} f_{ij}$ represent the **marginal relative frequency distribution** of Y.

We define $f_{i|j}^{X|Y} = f_{ij}/f_{+j}$ to be the **conditional relative frequency distribution** of X given $Y = y_j$.

We define $f_{j|i}^{Y|X} = f_{ij}/f_{i+}$ to be the **conditional relative frequency distribution** of Y given $X = x_i$.

Table 4.3 Contingency table for travel class and satisfaction

		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total (rows)
Travel class	Economy	10	33	15	4	62
	Business	0	3	20	2	25
	First	0	0	5	8	13
	Total (columns)	10	36	40	14	100

Note that for a bivariate joint frequency distribution, there will only be two marginal (or relative) frequency distributions but possibly more than two conditional (or relative) frequency distributions.

Example 4.1.2 Recall the setup of Example 4.1.1. We now collect and evaluate the responses of 100 customers (instead of 12 passengers as in Example 4.1.1) regarding their choice of the travel class and their overall satisfaction with the flight quality.

The data is provided in Table 4.3 where each of the cell entries illustrates how many out of 100 passengers answered x_i *and* y_j: for example, the first entry "10" indicates that 10 passengers were flying in economy class *and* described the overall service quality as poor.

- The marginal frequency distributions are displayed in the last column and last row, respectively. For example, the marginal distribution of X refers to the frequency table of "travel class" (X) and tells us that 62 passengers were flying in economy class, 25 in business class, and 13 in first class. Similarly, the marginal distribution of "overall rating of flight quality" (Y) tells us that 10 passengers rated the quality as poor, 36 as fair, 40 as good, and 14 as very good.
- The conditional frequency distributions give us an idea about the behaviour of one variable when the other one is kept fixed. For example, the conditional distribution of the "overall rating of flight quality" (Y) among passengers who were flying in economy class ($f_{Y|X=\text{Economy}}$) gives $f_{1|1}^{Y|X} = 10/62 \approx 16\,\%$ which means that approximately 16 % of the customers in economy class are rating the quality as poor, $f_{2|1}^{Y|X} = 33/62 \approx 53\,\%$ of the customers in economy class are rating the quality as fair, $f_{3|1}^{Y|X} = 15/62 \approx 24\,\%$ of the customers in economy class are rating the quality as good and $f_{4|1}^{Y|X} = 4/62 \approx 7\,\%$ of the customers in economy class are rating the quality as very good. Similarly, $f_{3|2}^{Y|X} = 20/25 \approx 80\,\%$ which means that 80 % of the customers in business class are rating the quality as good and so on.
- The conditional frequency distribution of the "travel class" (X) of passengers given the "overall rating of flight quality" (Y) is obtained by $f_{X|Y=\text{Satisfaction level}}$. For example, $f_{X|Y=\text{good}}$ gives $f_{1|3}^{X|Y} = 15/40 = 37.5\,\%$ which means that 37.5 %

of the passengers who rated the flight to be good travelled in economy class, $f_{2|3}^{X|Y} = 20/40 = 50\,\%$ of the passengers who rated the flight to be good travelled in business class and $f_{3|3}^{X|Y} = 5/40 = 12.5\,\%$ of the passengers who rated the flight to be good travelled in first class.

- In total, we have 100 customers and hence

$$n = \sum_{i=1}^{k} n_{i+} = 62 + 25 + 13 = \sum_{j=1}^{l} n_{+j} = 10 + 36 + 40 + 14$$
$$= \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = 10 + 33 + 15 + 4 + +3 + 20 + 2 + 5 + 8 = 100$$

- Alternatively, we can summarize X and Y using the relative frequencies as follows:

		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total (rows)
Travel class	Economy	$\frac{10}{100}$	$\frac{33}{100}$	$\frac{15}{100}$	$\frac{4}{100}$	$\frac{62}{100}$
	Business	0	$\frac{3}{100}$	$\frac{20}{100}$	$\frac{2}{100}$	$\frac{25}{100}$
	First	0	0	$\frac{5}{100}$	$\frac{8}{100}$	$\frac{13}{100}$
	Total (columns)	$\frac{10}{100}$	$\frac{36}{100}$	$\frac{40}{100}$	$\frac{14}{100}$	1

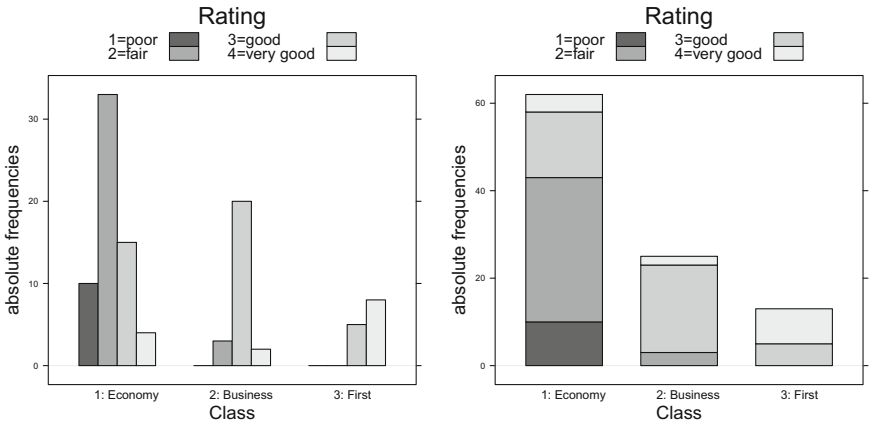
To produce the frequency table without the marginal distributions, we can use the *R* command `table(X,Y)`. To obtain the full contingency table including the marginal distributions in *R*, one can use the function `addmargins()`. For the relative frequencies, the function `prop.table()` can be used. In summary, a full contingency table is obtained by using

```
addmargins(table(X,Y))
addmargins(prop.table(table(X,Y)))
```

4.1.3 Graphical Representation of Two Nominal or Ordinal Variables

Bar charts (see Sect. 2.3.1) can be used to graphically summarize the association between two nominal or two ordinal variables. The bar chart is drawn for X and the categories of Y are represented by separated bars or stacked bars for each category of X. In this way, we summarize the joint distribution of the contingency table.

Example 4.1.3 Consider Example 4.1.2. There are 62 passengers flying in the economy class. From these 62 passengers, 10 rated the quality of the flight as poor, 33 as fair, 15 as good, and 4 as very good. This means for $X = x_1 (=$ Economy), we can



(a) Categories next to each other (b) Categories stacked

Fig. 4.1 Bar charts for travel class and rating of satisfaction

either place 4 bars next to each other, as in Fig. 4.1a, or we can stack them on top of each other, as in Fig. 4.1b. The same can be done for the other categories of X, see Fig. 4.1. Stacked and stratified bar charts are prepared in *R* by calling the library `lattice` and using the function `bar chart`. In detail, one needs to specify:

```
Class <- c(rep('1: Economy=',62),rep('2: Business',25),
rep('3: First',13))
Rating <- c(rep('1=poor',10),rep('2=fair',33),...)
library(lattice)
barchart(table(Class,Rating),horizontal=FALSE,stack=FALSE)
barchart(table(Class,Rating),horizontal=FALSE,stack=TRUE)
```

Remark 4.1.2 There are several other options in *R* to specify stratified bar charts. We refer the interested reader to Exercise 2.6 to explore how the *R* package `ggplot2` can be used to make such graphics. Sometimes it can also be useful to visualize the difference of two variables and not stack or stratify the bars, see Exercise 2.1.

Independence and Expected Frequencies An important statistical concept is **independence**. In this section, we touch upon its descriptive aspects, see Chaps. 6 (Sect. 6.5) and 7 (Sect. 7.5) for more theoretical details. Two variables are considered to be independent if the observations on one variable do not influence the observations on the other variable. For example, suppose two different persons roll a die separately; then, the outcomes of their rolls do not depend on each other. So we can say that the two observations are independent. In the context of contingency tables, two variables are independent of each other when the joint relative frequency equals the product of the marginal relative frequencies of the two variables, i.e. the

Table 4.4 Observed and expected absolute frequencies for the airline survey

		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total
Travel class	Economy	10 (6.2)	33 (22.32)	15 (24.8)	4 (8.68)	62
	Business	0 (2.5)	3 (9.0)	20 (10.0)	2 (3.5)	25
	First	0 (1.3)	0 (4.68)	5 (5.2)	8 (1.82)	13
	Total	10	36	40	14	100

following equation holds:

$$f_{ij} = f_{i+} f_{+j} \,. \tag{4.2}$$

The **expected absolute frequencies under independence** are obtained by

$$\tilde{n}_{ij} = n\, f_{ij} = n \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n} \,. \tag{4.3}$$

Note that the absolute frequencies are always integers but the expected absolute frequencies may not always be integers.

Example 4.1.4 Recall Example 4.1.2. The expected absolute frequencies for the contingency table can be calculated using (4.3). For example,

$$\tilde{n}_{11} = \frac{62 \cdot 10}{100} = 6.2, \quad \tilde{n}_{12} = \frac{62 \cdot 36}{100} = 22.32 \quad \text{etc.}$$

Table 4.4 lists both the observed absolute frequency and expected absolute frequency (in brackets).

To calculate the expected absolute frequencies in R, we can access the "expected" object returned from a χ^2-test applied to the respective contingency table as follows:

```
chisq.test(table(Class,Rating))$expected
```

A detailed motivation and explanation of this command is given in Sect. 10.8.

4.2 Measures of Association for Two Discrete Variables

When two variables are not independent, then they are associated. Their association can be weak or strong. Now we describe some popular measures of association. Measures of association describe the degree of association between two variables and can have a direction as well. Note that if variables are defined on a nominal scale, then nothing can be said about the direction of association, only about the strength.

Let us first consider a 2×2 contingency table which is a special case of a $k \times l$ contingency table, see Table 4.5.

Table 4.5 2×2 contingency table

		Y		
		y_1	y_2	Total (row)
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	Total (column)	$a+c$	$b+d$	n

Table 4.6 2×2 contingency table

		Persons		
		Not affected	Affected	Total (row)
Vaccination	Vaccinated	90	10	100
	Not vaccinated	40	60	100
	Total (column)	130	70	200

The variables X and Y are independent if

$$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n} \tag{4.4}$$

or equivalently if

$$a = \frac{(a+b)(a+c)}{n}\,. \tag{4.5}$$

Note that some other forms of the conditions (4.4)–(4.5) can also be derived in terms of a, b, c, and d.

Example 4.2.1 Suppose a vaccination against flu (influenza) is given to 200 persons. Some of the persons may get affected by flu despite the vaccination. The data is summarized in Table 4.6. Using the notations of Table 4.5, we have $a = 90, b = 10, c = 40, d = 60$, and thus, $(a+b)(a+c)/n = 100 \cdot 130/200 = 65$ which is less than $a = 90$. Hence, being affected by flu is not independent of the vaccination, i.e. whether one is vaccinated or not has an influence on getting affected by flu. In the vaccinated group, only 10 of 100 persons are affected by flu while in the group not vaccinated 60 of 100 persons are affected. Another interpretation is that if independence holds, then we would expect 65 persons to be not affected by flu in the vaccinated group but we observe 90 persons. This shows that vaccination has a protective effect.

To gain a better understanding about the strength of association between two variables, we need to develop the concept of dependence and independence further. The following three subsections illustrate this in more detail.

4.2.1 Pearson's χ^2 Statistic

We now introduce Pearson's χ^2 statistic which is used for measuring the association between variables in a contingency table and plays an important role in the construction of statistical tests, see Sect. 10.8. The χ^2 statistic or χ^2 coefficient for a $k \times l$ contingency table is given as

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}. \tag{4.6}$$

A simpler formula for 2×2 contingency tables is

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{4.7}$$

The idea behind the χ^2 coefficient is that when the relationship between two variables is stronger, then the deviations between observed and expected frequencies are expected to be higher (because the expected frequencies are calculated assuming independence) and this indicates a stronger relationship between the two variables. If observed and expected frequencies are identical or similar, then this is an indication that the association between the two variables is weak and the variables may even be independent. The χ^2 statistic for a $k \times l$ contingency table sums up all the differences between the observed and expected frequencies, squares them, and scales them with respect to the expected frequencies. The squaring of the difference makes the statistic independent of the positive and negative signs of the difference between observed and expected frequencies. The range of values for χ^2 is

$$0 \leq \chi^2 \leq n(\min(k, l) - 1). \tag{4.8}$$

Note that $\min(k, l)$ is the minimum function and simply returns the smaller of the two numbers k and l. For example, $\min(3, 4)$ returns the value 3. Consequently the values of χ^2 obtained from (4.6) can be compared with the range from (4.8). A value of χ^2 close to zero indicates a weak association and a value of χ^2 close to $n(\min(k, l) - 1)$ indicates a strong association between the two variables. Note that the range of χ^2 depends on n, k and l, i.e. the sample size and the dimension of the contingency table.

The χ^2 statistic is a *symmetric* measure in the sense that its value does not depend on which variable is defined as X and which as Y.

Example 4.2.2 Consider Examples 4.1.2 and 4.1.4. Using the values from Table 4.4, we can calculate the χ^2 statistic as

$$\chi^2 = \frac{(10 - 6.2)^2}{6.2} + \frac{(33 - 22.32)^2}{22.32} + \cdots + \frac{(8 - 1.82)^2}{1.82} = 57.95064$$

The maximum possible value for the χ^2 statistic is $100(\min(4, 3) - 1) = 200$. Thus, $\chi^2 \approx 57$ indicates a moderate association between "travel class" and "overall rating of flight quality" of the passengers. In *R*, we obtain this result as follows:

```
chisq.test(table(Class,Rating))$statistic
```

4.2.2 Cramer's V Statistic

A problem with Pearson's χ^2 coefficient is that the range of its maximum value depends on the sample size and the size of the contingency table. These values may vary in different situations. To overcome this problem, the coefficient can be standardized to lie between 0 and 1 so that it is independent of the sample size as well as the dimension of the contingency table. Since $n(\min(k, l) - 1)$ was the maximal value of the χ^2 statistic, dividing χ^2 by this maximal value automatically leads to a scaled version with maximal value 1. This idea is used by Cramer's V statistic which for a $k \times l$ contingency table is given by

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}}. \tag{4.9}$$

The closer the value of V gets to 1, the stronger the association between the two variables.

Example 4.2.3 Consider Example 4.2.2. The obtained χ^2 statistic is 57.95064. To obtain Cramer's V, we just need to calculate

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{57.95064}{100(3 - 1)}} \approx 0.54. \tag{4.10}$$

This indicates a moderate association between "travel class" and "overall rating of flight quality" because 0.54 lies in the middle of 0 and 1. In R, there are two options to calculate V: (i) to calculate the χ^2 statistic and then adjust it as in (4.9), (ii) to use the functions `assocstats` and `xtabs` contained in the package `vcd` as follows:

```
library(vcd)
assocstats(xtabs(~Class+Rating))
```

4.2.3 Contingency Coefficient C

Another option to standardize χ^2 is given by a corrected version of Pearson's contingency coefficient:

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1}} \sqrt{\frac{\chi^2}{\chi^2 + n}}, \tag{4.11}$$

with

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{and} \quad C_{\max} = \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}. \tag{4.12}$$

It always lies between 0 and 1. The closer the value of C is to 1, the stronger the association.

Example 4.2.4 We know from Example 4.2.2 that the χ^2 statistic for travel class and satisfaction level is 57.95064. To calculate C_{corr}, we need the following calculations:

$$C = \sqrt{\frac{57.95064}{57.95064 + 100}} = 0.606, \quad C_{\max} = \sqrt{\frac{\min(4,3) - 1}{\min(4,3)}} = \sqrt{\frac{2}{3}} = 0.816,$$

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \frac{0.606}{0.816} \approx 0.74\,.$$

There is a moderate to strong association between "travel class" and "overall rating of flight quality" of the passengers. We can compute C in *R* using the `vcd` package as follows:

```
library(vcd)
Cmax = sqrt((min(c(3,4))-1)/min(c(3,4)))
assocstats(xtabs(~Class+Rating))$cont/Cmax
```

4.2.4 Relative Risks and Odds Ratios

We now introduce the concepts of odds ratios and relative risks. Consider a 2×2 contingency table as introduced in Table 4.5. Now suppose we have two variables X and Y with their conditional distributions $f_{i|j}^{X|Y}$ and $f_{j|i}^{Y|X}$. In the context of a 2×2 contingency table, $f_{1|1}^{X|Y} = n_{11}/n_{+1}$, $f_{1|2}^{X|Y} = n_{12}/n_{+2}$, $f_{2|2}^{X|Y} = n_{22}/n_{+2}$, and $f_{2|1}^{X|Y} = n_{21}/n_{+1}$. The relative risks are defined as the ratio of two conditional distributions, for example

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{a/(a+c)}{b/(b+d)} \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{c/(a+c)}{d/(b+d)}\,. \tag{4.13}$$

The odds ratio is defined as the ratio of these relative risks from (4.13) as

$$OR = \frac{f_{1|1}^{X|Y}/f_{1|2}^{X|Y}}{f_{2|1}^{X|Y}/f_{2|2}^{X|Y}} = \frac{f_{1|1}^{X|Y} f_{2|2}^{X|Y}}{f_{2|1}^{X|Y} f_{1|2}^{X|Y}} = \frac{a\,d}{b\,c}\,. \tag{4.14}$$

Alternatively, the odds ratio can be defined as the ratio of the chances for "disease", a/b (number of smokers with the disease divided by the number of non-smokers with the disease), and no disease, c/d (number of smokers with no disease divided by the number of non-smokers with no disease).

The relative risks compare proportions, while the odds ratio compares odds.

Example 4.2.5 A classical example refers to the possible association of smoking with a particular disease. Consider the following data on 240 individuals:

		Smoking		Total (row)
		Yes	No	
Disease	Yes	34	66	100
	No	22	118	140
	Total (column)	56	184	240

We calculate the following relative risks:

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{34/56}{66/184} \approx 1.69 \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{22/56}{118/184} \approx 0.61\,. \tag{4.15}$$

Thus, the proportion of individuals with the disease is 1.69 times higher among smokers when compared with non-smokers. Similarly, the proportion of healthy individuals is 0.61 times smaller among smokers when compared with non-smokers.

The relative risks are calculated to compare the proportion of sick or healthy patients between smokers and non-smokers. Using these two relative risks, the odds ratio is obtained as

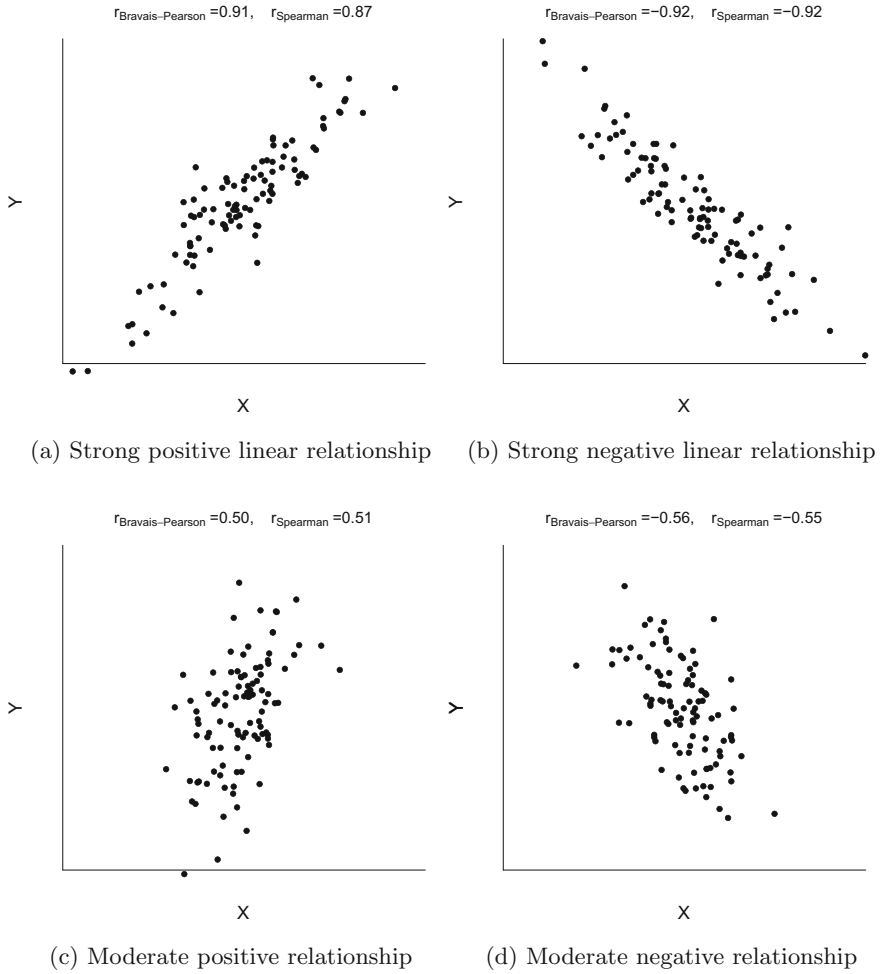
$$OR = \frac{34 \times 118}{66 \times 22} = 2.76.$$

We can interpret this outcome as follows: (i) the chances of smoking are 2.76 times higher for individuals with the disease compared with healthy individuals (follows from definition (4.14)). We can also say that (ii) the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. If we interchange either one of the "Yes" and "No" columns or the "Yes" and "No" rows, we obtain $OR = 1/2.76 \approx 0.36$, giving us further interpretations: (iii) the chances of smoking are 0.36 times lower for individuals without disease compared with individuals with the disease, and (iv) the chance of having the particular disease is 0.36 times lower for non-smokers compared with smokers. Note that all four interpretations are correct and one needs to choose the right interpretation in the light of the experimental situation and the question of interest.

4.3 Association Between Ordinal and Continuous Variables

4.3.1 Graphical Representation of Two Continuous Variables

A simple way to graphically summarize the association between two continuous variables is to plot the paired observations of the two variables in a two-dimensional coordinate system. If n paired observations for two continuous variables X and Y are available as $(x_i, y_i), i = 1, 2, \ldots, n$, then all such observations can be plotted



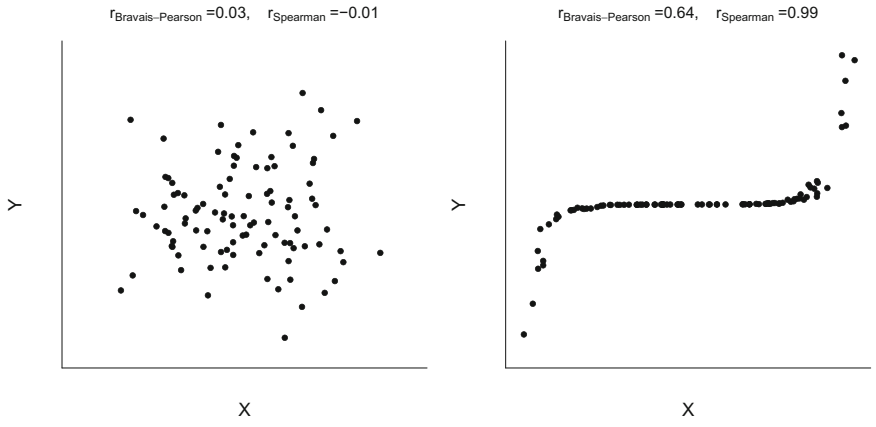
(a) Strong positive linear relationship (b) Strong negative linear relationship

(c) Moderate positive relationship (d) Moderate negative relationship

Fig. 4.2 Scatter plots

in a single graph. This graph is called a **scatter plot**. Such a plot reveals possible relationships and trends between the two variables. For example, Figs. 4.2 and 4.3 show scatter plots with six different types of association.

- Figure 4.2a shows increasing values of Y for increasing values of X. We call this relationship positive association. The relationship between X and Y is nearly linear because all the points lie around a straight line.
- Figure 4.2b shows decreasing values of Y for increasing values of X. We call this relationship negative association.
- Figure 4.2c tells us the same as Fig. 4.2a, except that the positive association is weaker.



(a) No clear relationship (b) Nonlinear relationship

Fig. 4.3 Continues Fig. 4.2—more scatter plots

- Figure 4.2d tells us the same as Fig. 4.2b, except that the negative association is weaker.
- Figure 4.3a shows that as the X-values increase, the values of Y neither increase nor decrease. This indicates that there is no clear relationship between X and Y and highlights the lack of association between X and Y.
- Figure 4.3b illustrates a nonlinear relationship between X- and Y-values.

Example 4.3.1 To explore the possible relationship between the overall number of tweets with the number of followers on Twitter, we take a sample of 10 prime ministers and heads of state in different countries as of June 2014 and obtain the following data:

Name	Tweets	Followers
Angela Merkel	25	7194
Barack Obama	11,800	43,400,000
Jacob Zuma	99	324,000
Dilma Rousseff	1934	2,330,000
Sauli Niinistö	199	39,000
Vladimir Putin	2539	189,000
Francois Hollande	4334	639,000
David Cameron	952	688,000
Enrique P. Nieto	3245	2,690,000
John Key	2468	110,000

The tweets are denoted by x_i and the followers are denoted by y_i, $i = 1, 2, \ldots, 10$. We plot paired observations (x_i, y_i) into a cartesian coordinate system. For example, we plot $(x_1, y_1) = (25, 7194)$ for Angela Merkel, $(x_2, y_2) = (11, 800, 43, 400, 000)$

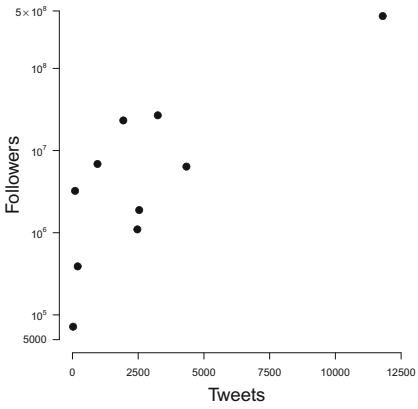


Fig. 4.4 Scatter plot between tweets and followers

for Barack Obama, and so on. Figure 4.4 shows the scatter plot for the number of tweets and the number of followers (on a log-scale).

One can see that there is a positive association between the number of tweets and the number of followers. This does, however, *not* imply a causal relationship: it is not necessarily *because* someone tweets more he/she has more followers or *because* someone has more followers he/she tweets more; the scatter plot just describes that those with more tweets have more followers. In *R*, we produce this scatter plot by the `plot` command:

```
tweets <- c(25,11800,99,...)
followers <- c(7194,43400000,...)
plot(tweets,followers)
```

4.3.2 Correlation Coefficient

Suppose two variables X and Y are measured on a continuous scale and are linearly related like $Y = a + b\,X$ where a and b are constant values. The **correlation coefficient** $r(X, Y) = r$ measures the degree of *linear* relationship between X and Y using

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \tag{4.16}$$

with

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n\tilde{s}_X^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = n\tilde{s}_Y^2, \tag{4.17}$$

and

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\,. \tag{4.18}$$

Karl Pearson (1857–1936) presented the first rigorous treatment of correlation and acknowledged Auguste Bravais (1811–1863) for ascertaining the initial mathematical formulae for correlation. This is why the correlation coefficient is also known as the **Bravais–Pearson correlation coefficient**.

The correlation coefficient is independent of the units of measurement of X and Y. For example, if someone measures the height and weight in metres and kilograms respectively and another person measures them in centimetres and grams, respectively, then the correlation coefficient between the two sets of data will be the same. The correlation coefficient is symmetric, i.e. $r(X, Y) = r(Y, X)$. The limits of r are $-1 \leq r \leq 1$. If all the points in a scatter plot lie exactly on a straight line, then the linear relationship between X and Y is perfect and $|r| = 1$, see also Exercise 4.7. If the relationship between X and Y is (i) perfectly linear and increasing, then $r = +1$ and (ii) perfectly linear and decreasing, then $r = -1$. The signs of r thus determine the direction of the association. If r is close to zero, then it indicates that the variables are independent or the relationship is not linear. Note that if the relationship between X and Y is nonlinear, then the degree of linear relationship may be low and r is then close to zero even if the variables are clearly not independent. Note that $r(X, X) = 1$ and $r(X, -X) = -1$.

Example 4.3.2 Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive linear correlation in Fig. 4.2a ($r = 0.91$), strong negative linear correlation in Fig. 4.2b ($r = -0.92$), moderate positive linear correlation in Fig. 4.2c ($r = 0.50$), moderate negative linear association in Fig. 4.2d ($r = -0.56$), no visible correlation in Fig. 4.3a ($r = 0.03$), and strong nonlinear (but not so strong linear) correlation in Fig. 4.3b ($r = 0.64$).

Example 4.3.3 In a decathlon competition, a group of athletes are competing with each other in 10 different track and field events. Suppose we are interested in how the results of the 100-m race relate to the results of the long jump competition. The correlation coefficient for the 100-m race (X, in seconds) and the long jump event (Y, in metres) for 5 athletes participating in the 2004 Olympic Games (see also Appendix A.4) are listed in Table 4.7.

To calculate the correlation coefficient, we need the following summary statistics:

$$\bar{x} = \frac{1}{5}(10.85 + 10.44 + 10.50 + 10.89 + 10.62) = 10.66$$

$$\bar{y} = \frac{1}{5}(7.84 + 7.96 + 7.81 + 7.47 + 7.74) = 7.764$$

$$S_{xx} = (10.85 - 10.66)^2 + (10.44 - 10.66)^2 + \cdots + (10.62 - 10.66)^2 = 0.1646$$

Table 4.7 Results of 100-m race and long jump of 5 athletes

i	x_i	y_i
Roman Sebrle	10.85	7.84
Bryan Clay	10.44	7.96
Dmitriy Karpov	10.50	7.81
Dean Macey	10.89	7.47
Chiel Warners	10.62	7.74

$$S_{yy} = (7.84 - 7.764)^2 + (7.96 - 7.764)^2 + \cdots + (7.74 - 7.764)^2 = 0.13332$$
$$S_{xy} = (10.85 - 10.66)(7.84 - 7.764) + \cdots + (10.62 - 10.66)(7.74 - 7.764)$$
$$= -0.1027$$

The correlation coefficient therefore is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-0.1027}{\sqrt{0.1646 \times 0.13332}} \approx -0.69\,.$$

Since -0.69 is negative, we can say that (i) there is a negative correlation between the 100-m race and the long jump event, i.e., shorter running times result in longer long jump results, and (ii) this association is moderate to strong.

In R, we can obtain the results (after attaching the data) as follows:

```
cor(X.100m,X.Long.jump, method='pearson')
```

4.3.3 Spearman's Rank Correlation Coefficient

Consider a situation where n objects are ranked with respect to two variables X and Y. For instance, the variables could represent the opinion of two different judges in a talent competition who rank the participants with respect to their performance. This means that for each judge, the worst participant (with the lowest score x_i) is assigned rank 1, the second worst participant (with the second lowest score x_i) will receive rank 2, and so on. Thus, every participant has been given two ranks by two different judges. Suppose we want to measure the degree of association between the two different judgments; that is, the two different sets of ranks. We expect that under perfect agreement, both the judges give the same judgment in the sense that they give the same ranks to each candidate. However, if they are not in perfect agreement, then there may be some variation in the ranks assigned by them. To measure the degree of agreement, or, in general, the degree of association, one can use **Spearman's rank correlation coefficient**. As the name says, this correlation coefficient uses only the ranks of the values and not the values themselves. Thus, this measure is suitable for both ordinal and continuous variables. We introduce the following notations: let $R(x_i)$ denote the rank of the ith observation on X, i.e. the rank x_i among the ordered values of X. Similarly, $R(y_i)$ denotes the rank of the ith observation of y. The difference between the two rank values is $d_i = R(x_i) - R(y_i)$. Spearman's rank correlation coefficient is defined as

$$R = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}\,. \tag{4.19}$$

The values of R lie between -1 and $+1$ and measure the degree of correlation between the ranks of X and Y. Note that it does not matter whether we choose an ascending or descending order of the ranks, the value of R remains the same. When all the observations are assigned exactly the same ranks, then $R = 1$ and when all the observations are assigned exactly the opposite ranks, then $R = -1$.

Example 4.3.4 Look again at the scatter plots in Figs. 4.2 and 4.3. We observe strong positive correlation in Fig. 4.2a ($R = 0.87$), strong negative correlation in Fig. 4.2b ($R = -0.92$), moderate positive correlation in Fig. 4.2c ($R = 0.51$), moderate negative association in Fig. 4.2d ($R = -0.55$), no visible correlation in Fig. 4.3a ($R = -0.01$), and strong nonlinear correlation in Fig. 4.3b ($R = 0.99$).

Example 4.3.5 Let us follow Example 4.3.3 a bit further and calculate Spearman's rank correlation coefficient for the first five observations of the decathlon data. Again we list the results of the 100-m race (X) and the results of the long jump competition (Y). In addition, we assign ranks to both X and Y. For example, the shortest time receives rank 1, whereas the longest time receives rank 5. Similarly, the shortest long jump result receives rank 1, the longest long jump result receives rank 5.

i	x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
Roman Sebrle	10.85	4	7.84	4	0	0
Bryan Clay	10.44	1	7.96	5	−4	16
Dmitriy Karpov	10.50	2	7.81	3	−1	1
Dean Macey	10.89	5	7.47	1	−4	16
Chiel Warners	10.62	3	7.74	2	−1	1
Total						34

Using (4.19), Spearman's rank correlation coefficient can be calculated as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 34}{5 \cdot 24} = -0.7.$$

We therefore have a moderate to strong negative association between the 100-m race and the long jump event. We now know that for the 5 athletes above longer running times relate to shorter jumping distances which in turn means that a good performance in one discipline implies a good performance in the other discipline. In *R*, we can obtain the same results by using the `cor` command:

```
cor(X.100m,X.Long.jump, method='spearman')
```

If two or more observations take the same values for x_i (or y_i), then there is a **tie**. In such situations, the respective ranks can simply be averaged, though more complicated solutions also exist (one of which is implemented in the *R* function `cor`). For example, if in Example 4.3.5 Bryan Clay's was 10.50 s instead of 10.44 s, then both Bryan Clay and Dmitriy Karpov had the same time. Instead of assigning the ranks 1 and 2 to them, we assign the ranks 1.5 to each of them.

The differences between the correlation coefficient and the rank correlation coefficient are manifold: firstly, Pearson's correlation coefficient can be used for continuous variables only, but not for nominal or ordinal variables. The rank correlation coefficient can be used for either two continuous or two ordinal variables or a combination of an ordinal and a continuous variable, but not for two nominal variables. Moreover, the rank correlation coefficient responds to any type of relationship whereas

Pearson's correlation measures the degree of a linear relationship only—see also Fig. 4.3b. Another difference between the two correlation coefficients is that Pearson uses the entire information contained in the continuous data in contrast to the rank correlation coefficient which uses only ordinal information contained in the ordered data.

4.3.4 Measures Using Discordant and Concordant Pairs

Another concept which uses ranks to measure the association between ordinal variables is based on **concordant** and **discordant** observation pairs. It is best illustrated by means of an example.

Example 4.3.6 Suppose an online book store conducts a survey on their customer's satisfaction with respect to both the timeliness of deliveries (X) and payment options (Y). Let us consider the following 2×3 contingency table with a summary of the responses of 100 customers. We assume that the categories for both variables can be ordered and ranks can be assigned to different categories, see the numbers in brackets in Table 4.8. There are 100 observation pairs (x_i, y_i) which summarize the response of the customers with respect to both X and Y. For example, there are 18 customers who were unsatisfied with the timeliness of the deliveries and complained that there are not enough payment options. If we compare two responses (x_{i_1}, y_{i_1}) and (x_{i_2}, y_{i_2}), it might be possible that one customer is more happy (or more unhappy) than the other customer with respect to both X and Y or that one customer is more happy with respect to X but more unhappy with respect to Y (or vice versa). If the former is the case, then this is a concordant observation pair; if the latter is true, then it is a discordant pair. For instance, a customer who replied "enough" and "satisfied" is more happy than a customer who replied "not enough" and "unsatisfied" because he is more happy with respect to both X and Y.

In general, a pair is

- **concordant** if $i_2 > i_1$ and $j_2 > j_1$ (or $i_2 < i_1$ and $j_2 < j_1$),
- **discordant** if $i_2 < i_1$ and $j_2 > j_1$ (or $i_2 > i_1$ and $j_2 < j_1$),
- **tied** if $i_1 = i_2$ (or $j_1 = j_2$).

Table 4.8 Payment options and timeliness survey with 100 participating customers

		Timeliness			
		Unsatisfied (1)	Satisfied (2)	Very satisfied (3)	Total
Payment options	Not enough (1)	7	11	26	44
	Enough (2)	10	15	31	56
	Total	17	26	57	100

Obviously, if we have only concordant observations, then there is a strong positive association because a higher value of X (in terms of the ranking) implies a higher value of Y. However, if we have only discordant observations, then there is a clear negative association. The measures which are introduced below simply put the number of concordant and discordant pairs into relation. This idea is reflected in **Goodman and Kruskal's γ** which is defined as

$$\gamma = \frac{K}{K+D} - \frac{D}{K+D} = \frac{K-D}{K+D}, \tag{4.20}$$

where

$$K = \sum_{i<m} \sum_{j<n} n_{ij} n_{mn}, \quad D = \sum_{i<m} \sum_{j>n} n_{ij} n_{mn}$$

describe the number of concordant and discordant observation pairs, respectively. An alternative measure is **Stuart's τ_c** given as

$$\tau_c = \frac{2\min(k,l)(K-D)}{n^2(\min(k,l)-1)}. \tag{4.21}$$

Both measures are standardized to lie between -1 and 1, where larger values indicate a stronger association and the sign indicates the direction of the association.

Example 4.3.7 Consider Example 4.3.6. A customer who replied "enough" and "satisfied" is more happy than a customer who replied "not enough" and "unsatisfied" because the observation pairs, using ranks, are (2, 2) and (1, 1) and therefore $i_2 > i_1$ and $j_2 > j_1$. There are 7×15 such pairs. Similarly those who said "not enough" and "unsatisfied" are less happy than those who said "enough" and "very satisfied" (7×31 pairs). Table 4.5 summarizes the comparisons in detail.

Table 4.5a shows that $(x_1, y_1) =$ (not enough, unsatisfied) is concordant to $(x_2, y_2) =$ (enough, satisfied) and $(x_2, y_3) =$ (enough, very satisfied) and tied to $(x_2, y_1) =$ (enough, unsatisfied), $(x_1, y_2) =$ (not enough, satisfied), and $(x_1, y_3) =$ (not enough, very satisfied). Thus for these comparisons, we have 0 discordant pairs, $(7 \times 15) + (7 \times 31)$ concordant pairs and $7 \times (10 + 11 + 26)$ tied pairs. Table 4.5b–f show how the task can be completed. While tiresome, systematically working through the table (and making sure to not count pairs more than once) yields

$$K = 7 \times (15 + 31) + 11 \times 31 = 663$$
$$D = 10 \times (11 + 26) + 15 \times 26 = 760.$$

As a visual rule of thumb, working from the top left to the bottom right yields the concordant pairs; and working from the bottom left to the top right yields the discordant pairs. It follows that $K = (663 - 760)/(663 + 760) \approx -0.07$ which indicates no clear relationship between the two variables. A similar result is obtained using τ_c which is $4 \times (760 - 663)/100^2 \approx 0.039$. This rather lengthy task can be made much quicker by using the `ord.gamma` and `ord.tau` commands from the *R* library `ryouready`:

(a)	y_1	y_2	y_3
x_1		t	t
x_2	t	c	c

(b)	y_1	y_2	y_3
x_1	t		t
x_2	d	t	c

(c)	y_1	y_2	y_3
x_1	t	t	
x_2	d	d	t

(d)	y_1	y_2	y_3
x_1	t	d	d
x_2		t	t

(e)	y_1	y_2	y_3
x_1	c	t	d
x_2	t		t

(f)	y_1	y_2	y_3
x_1	c	c	t
x_2	t	t	

Fig. 4.5 Scheme to visualize concordant (c), discordant (d), and tied (t) pairs in a 2×3 contingency table

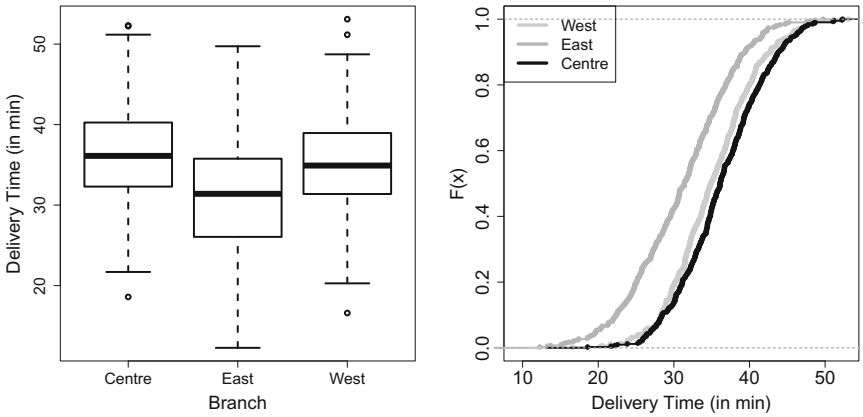
```
library(ryouready)
ex <- matrix(c(7,11,26,10,15,31),ncol=3,byrow=T)
ord.gamma(ex)
ord.tau(ex)
```

4.4 Visualization of Variables from Different Scales

If we want to jointly visualize the association between a variable X, which is either nominal or ordinal and another variable Y, which is continuous, then we can use any graph which is suitable for the continuous variable (see Chaps. 2 and 3) and produce it for each category of the nominal/ordinal variable. We recommend using stratified box plots or stratified ECDF's, as they are easy to read when summarized in a single figure; however, it is also possible to place histograms next to each other or on top of each other, or overlay kernel density plots, but we do not illustrate this here in more detail.

Example 4.4.1 Consider again our pizza delivery example (Appendix A.4). If we are interested in the pizza delivery times by branch, we may simply plot the box plots and ECDF's of delivery time by branch. Figure 4.6 shows that the shortest delivery times can be observed in the branch in the East. Producing these graphs in *R* is straightforward: The `boxplot` command can be used for two variables by separating them with the $\sim$ sign. For the ECDF, we have to produce a plot for each branch and overlay them with the "add=TRUE" option.

```
boxplot(time~branch)
plot.ecdf(time[branch=='East'])
plot.ecdf(time[branch=='West'], add=TRUE)
plot.ecdf(time[branch=='Centre'], add=TRUE)
```



(a) Using box plots (b) Using ECDF's

Fig. 4.6 Distribution of pizza delivery time stratified by branch

4.5 Key Points and Further Issues

Note:

✓ How to use different measures of association:

2 nominal variables	→	Pearson's χ^2, relative risks, odds ratio, Cramer's V, and C_{corr}
2 ordinal variables	→	Spearman's rank correlation coefficient, γ, τ_c
2 continuous variables	→	Pearson's correlation coefficient, Spearman's correlation coefficient

✓ For two variables which are measured on different scales, for example continuous/ordinal or ordinal/nominal, one should use measures of association suitable for the less informative of the two scales.

✓ Another graphical representation of both a continuous and discrete variable is stratified confidence interval plots (error plots), see Chap. 9.

4.6 Exercises

Exercise 4.1 A newspaper asks two of its staff to review the coffee quality at different trendy cafés. The coffee can be rated on a scale from 1 (miserable) to 10 (excellent). The results of the two coffee enthusiasts X and Y are as follows:

Café i	x_i	y_i
1	3	6
2	8	7
3	7	10
4	9	8
5	5	4

(a) Calculate and interpret Spearman's rank correlation coefficient.
(b) Does Spearman's R differ depending on whether ranks are assigned in a decreasing or increasing order?
(c) Suppose the coffee can only be rated as either good (>5) or bad (≤ 5). Do the chances of a good rating differ between the two journalists?

Exercise 4.2 A total of 150 customers of a petrol station are asked about their satisfaction with their car and motorbike insurance. The results are summarized below:

	Satisfied	Unsatisfied	Total
Car	33	25	58
Car (diesel engine)	29	31	60
Motorbike	12	20	32
Total	74	76	150

(a) Determine and interpret Pearson's χ^2 statistic, Cramer's V, and C_{corr}.
(b) Combine the categories "car" and "car (diesel engine)" and produce the corresponding 2×2 table. Calculate χ^2 as efficiently as possible and give a meaningful interpretation of the odds ratio.
(c) Compare the results from (a) and (b).

Exercise 4.3 There has been a big debate about the usefulness of speed limits on public roads. Consider the following table which lists the speed limits for country roads (in miles/h) and traffic deaths (per 100 million km) for different countries in 1986 when the debate was particularly serious:

(a) Draw the scatter plot for the two variables.
(b) Calculate the Bravais–Pearson and Spearman correlation coefficients.

Country	Speed limit	Traffic deaths
Denmark	55	4.1
Japan	55	4.7
Canada	60	4.3
Netherlands	60	5.1
Italy	75	6.1

(c) What are the effects on the correlation coefficients if the speed limit is given in km/h rather than miles/h (1 mile/h $\approx$ 1.61 km/h)?
(d) Consider one more observation: the speed limit for England was 70 miles/h and the death rate was 3.1.

(i) Add this observation to the scatter plot.
(ii) Calculate the Bravais–Pearson correlation coefficient given this additional observation.

Exercise 4.4 The famous passenger liner *Titanic* hit an iceberg in 1912 and sank. A total of 337 passengers travelled in first class, 285 in second class, and 721 in third class. In addition, there were 885 staff members on board. Not all passengers could be rescued. Only the following were rescued: 135 from the first class, 160 from the second class, 541 from the third class and 674 staff.

(a) Determine and interpret the contingency table for the variables "travel class" and "rescue status".
(b) Use a contingency table to summarize the conditional relative frequency distributions of rescue status given travel class. Could there be an association of the two variables?
(c) What would the contingency table from (a) look like under the independence assumption? Calculate Cramer's V statistic. Is there any association between travel class and rescue status?
(d) Combine the categories "first class" and "second class" as well as "third class" and "staff". Create a contingency table based on these new categories. Determine and interpret Cramer's V, the odds ratio, and relative risks of your choice.
(e) Given the results from (a) to (d), what are your conclusions?

Exercise 4.5 To study the association of the monthly average temperature (in °C, X) and hotel occupation (in %, Y), we consider data from three cities: Polenca (Mallorca, Spain) as a summer holiday destination, Davos (Switzerland) as a winter skiing destination, and Basel (Switzerland) as a business destination.

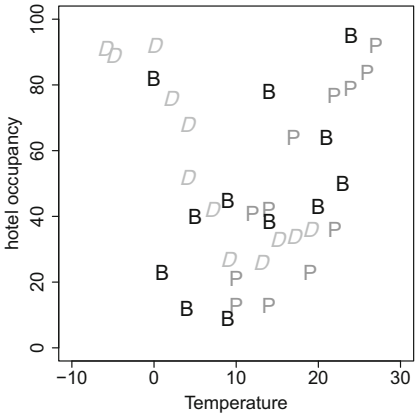


Fig. 4.7 Temperature and hotel occupancy for the different cities

Months	Davos		Polenca		Basel	
	X	Y	X	Y	X	Y
Jan	−6	91	10	13	1	23
Feb	−5	89	10	21	0	82
Mar	2	76	14	42	5	40
Apr	4	52	17	64	9	45
May	7	42	22	79	14	39
Jun	15	36	24	81	20	43
Jul	17	37	26	86	23	50
Aug	19	39	27	92	24	95
Sep	13	26	22	36	21	64
Oct	9	27	19	23	14	78
Nov	4	68	14	13	9	9
Dec	0	92	12	41	4	12

(a) Calculate the Bravais–Pearson correlation coefficient. The following summary statistics are available: $\sum_{i=1}^{36} x_i y_i = 22,776, \bar{x} = 12.22, \bar{y} = 51.28, \tilde{s}_x^2 = 76.95$, and $\tilde{s}_y^2 = 706.98$.

(b) Interpret the scatter plot in Fig. 4.7 which visualizes temperature and hotel occupancy for Davos (D), Polenca (P), and Basel (B).

(c) Use R to calculate the correlation coefficient separately for each city. Interpret the results and discuss the use of the correlation coefficient if more than two variables are available.

Exercise 4.6 Consider a neighbourhood survey on the use of a local park. Respondents were asked whether the park may be used for summer music concerts and whether dog owners should put their dogs on a lead. The results are summarized in the following contingency table:

		Put dogs on a lead			
		Agree	No opinion	Disagree	Total
Use for concerts	Agree	82	4	0	86
	No opinion	8	43	9	60
	Disagree	0	2	10	12
	Total	90	49	19	158

(a) Calculate and interpret Goodman and Kruskal's γ.
(b) Now ignore the ordinal structure of the data and calculate Cramer's V.
(c) Create the contingency table which is obtained when the categories "no opinion" and "agree" are combined.
(d) What is the relative risk of disagreement with summer concerts depending on the opinion about using leads?
(e) Calculate the odds ratio and offer two interpretations of it.
(f) Determine γ for the table calculated in (c).
(g) What is your final interpretation and what may be the best measure to use in this example?

Exercise 4.7 Consider n observations for which $y_i = a + bx_i$, $b > 0$, holds. Show that $r = 1$.

Exercise 4.8 Make yourself familiar with the Olympic decathlon data described in Appendix A.4. Read in and attach the data in *R*.

(a) Use *R* to calculate and interpret the Bravais–Pearson correlation coefficient between the results of the discus and the shot-put events.
(b) There are 10 continuous variables. How many different correlation coefficients can you calculate? How would you summarize them?
(c) Apply the `cor` command to the whole data and interpret the output.
(d) Omit the two rows which contain missing data and interpret the output again.

Exercise 4.9 We are interested in the pizza delivery data which is described in Appendix A.4.

(a) Read in the data and create two new binary variables which describe whether a pizza was hot (>65 °C) and the delivery time was short (<30 min). Create a contingency table for the two new variables.
(b) Calculate and interpret the odds ratio for the contingency table from (a).

(c) Use Cramer's V, Stuart's τ_c, Goodman and Kruskal's γ, and a stacked bar chart to explore the association between the categorical time and temperature variables.
(d) Draw a scatter plot for the continuous time and temperature variables. Determine both the Bravais–Pearson and Spearman correlation coefficients.
(e) Use methods of your choice to explore the relationship between temperature and driver, operator, number of ordered pizzas and bill. Is it clear which of the variables influence the pizza temperature?

→ Solutions to all exercises in this chapter can be found on p. 345

Part II
Probability Calculus

Combinatorics

5

5.1 Introduction

Combinatorics is a special branch of mathematics. It has many applications not only in several interesting fields such as enumerative combinatorics (the classical application), but also in other fields, for example in graph theory and optimization.

First, we try to motivate and understand the role of combinatorics in statistics. Consider a simple example in which someone goes to a cafe. The person would like a hot beverage and a cake. Assume that one can choose among three different beverages, for example cappuccino, hot chocolate, and green tea, and three different cakes, let us say carrot cake, chocolate cake, and lemon tart. The person may consider different beverage and cake combinations when placing the order, for example carrot cake and cappuccino, carrot cake and tea, and hot chocolate and lemon tart. From a statistical perspective, the customer is evaluating the possible combinations before making a decision. Depending on their preferences, the order will be placed by choosing one of the combinations.

In this example, it is easy to calculate the number of possible combinations. There are three different beverages and three different cakes to choose from, leading to nine different (3×3) beverage and cake combinations. However, suppose there is a choice of 15 hot beverages and 8 different cakes. How many orders can be made? (Answer: 15×8) What if the person decides to order two cakes, how will it affect the number of possible combinations of choices? It will be a tedious task to count all the possibilities. So we need a systematic approach to count such possible combinations. Combinatorics deals with the counting of different possibilities in a systematic approach.

People often use the urn model to understand the system in the counting process. The urn model deals with the drawing of balls from an urn. The balls in the urn

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_5

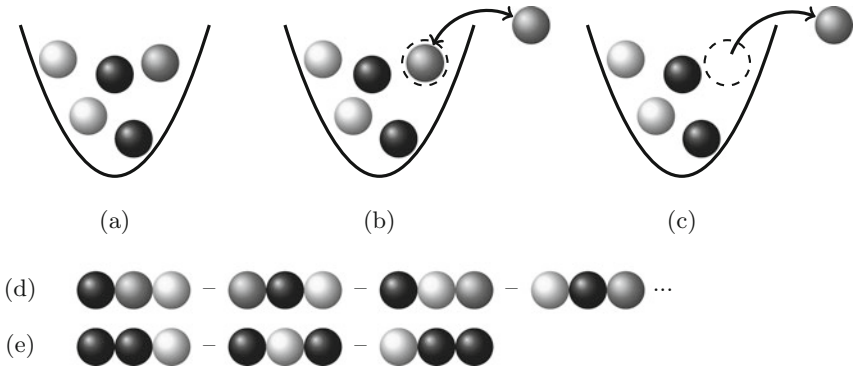


Fig. 5.1 **a** Representation of the urn model. Drawing from the urn model **b** with replacement and **c** without replacement. Compositions of three drawn balls: **d** all balls are distinguishable and **e** some balls are not distinguishable

represent the units of a population, or the features of a population. The balls may vary in colour or size to represent specific properties of a unit or feature. We illustrate this concept in more detail in Fig. 5.1.

Suppose there are 5 balls of three different colours—two black, one grey, and two white (see Fig. 5.1a). This can be generalized to a situation in which there are n balls in the urn and we want to draw m balls. Suppose we want to know

- how many different possibilities exist to draw m out of n balls (thus determining the number of distinguishable **combinations**).

To deal with such a question, we first need to decide whether a ball will be put back into the urn after it is drawn or not. Figure 5.1b illustrates that a grey ball is drawn from the urn and then placed back (illustrated by the two-headed arrow). We say the ball is drawn *with replacement*. Figure 5.1c illustrates a different situation in which the grey ball is drawn from the urn and is *not* placed back into the urn (illustrated by the one-headed arrow). We say the ball is drawn *without replacement*.

Further, we may be interested in knowing the

- total number of ways in which the chosen set of balls can be arranged in a distinguishable order (which we will define as **permutations** later in this chapter).

To answer the question how many permutations exist, we first need to decide whether all the chosen balls are distinguishable from each other or not. For example, in Fig. 5.1d, the three chosen balls have different colours; therefore, they are distinguishable. There are many options on how they can be arranged. In contrast, some

of the chosen balls in Fig. 5.1e are the same colour, they are therefore not distinguishable. Consequently, the number of combinations is much more limited. The concept of balls and urns just represents the features of observations from a sample. We illustrate this in more detail in the following example.

Example 5.1.1 Say a father promises his daughter three scoops of ice cream if she cleans up her room. For simplicity, let us assume the daughter has a choice of four flavours: chocolate, banana, cherry, and lemon. How many different choices does the daughter have? If each scoop has to be a different flavour she obviously has much less choice than if the scoops can have the same flavour. In the urn model, this is represented by the concept of "with/without replacement". The urn contains 4 balls of 4 different colours which represent the ice cream flavours. For each of the three scoops, a ball is drawn to determine the flavour. If we draw with replacement, each flavour can be potentially chosen multiple times; however, if we draw without replacement each flavour can be chosen only once. Then, the number of possible combinations is easy to calculate: it is 4, i.e. (chocolate, banana, and cherry); (chocolate, banana, and lemon); (chocolate, cherry, and lemon); and (banana, cherry, and lemon). But what if we have more choices? Or if we can draw flavours multiple times? We then need calculation rules which help us counting the number of options.

Now, let us assume that the daughter picked the flavours (chocolate [C], banana [B], and lemon [L]). Like many other children, she prefers to eat her most favourite flavour (chocolate) last, and her least favourite flavour (cherry) first. Therefore, the order in which the scoops are placed on top of the cone are important! In how many different ways can the scoops be placed on top of the cone? This relates to the question of the number of distinguishable *permutations*. The answer is 6: (C,B,L)–(C,L,B)–(B,L,C)–(B,C,L)–(L,B,C)–(L,C,B). But what if the daughter did pick a flavour multiple times, e.g. (chocolate, chocolate, lemon)? Since the two chocolate scoops are non-distinguishable, there are fewer permutations: (chocolate, chocolate, and lemon)–(chocolate, lemon, and chocolate)–(lemon, chocolate, and chocolate).

The bottom line of this example is that the number of combinations/options is determined by (i) whether we draw with or without replacement (i.e. allow flavours to be chosen more than once) and (ii) whether the arrangement in a particular order (=permutation) is of any specific interest.

Consider the urn example again. Suppose three balls of different colours, black, grey, and white, are drawn. Now there are two options: The first option is to take into account the order in which the balls are drawn. In such a situation, two possible sets of balls such as (black, grey, and white) and (white, black, and grey) constitute two different sets. Such a set is called an *ordered set*. In the second option, we do not take into account the order in which the balls are drawn. In such a situation, the two possible sets of balls such as (black, grey, and white) and (white, black, and grey) are the same sets and constitute an *unordered set* of balls.

Definition 5.1.1 A group of elements is said to be **ordered** if the order in which these elements are drawn is of relevance. Otherwise, it is called **unordered**.

Examples.

- Ordered samples:
  - The first three places in an Olympic 100 m race are determined by the order in which the athletes arrive at the finishing line. If 8 athletes are competing with each other, the number of possible results for the first three places is of interest. In the urn language, we are taking draws without replacement (since every athlete can only have one distinct place).
  - In a raffle with two prizes, the first drawn raffle ticket gets the first prize and the second raffle ticket gets the second prize.
  - There exist various esoteric tarot card games which claim to foretell someone's fortune with respect to several aspects of life. The order in which the cards are shown on the table is important for the interpretation.
- Unordered samples:
  - The selected members for a national football team. The order in which the selected names are announced is irrelevant.
  - Out of 10 economists, 10 medical doctors, and 10 statisticians, an advisory committee consisting of 4 economists, 3 medical doctors, and 2 statisticians is elected.
  - Fishing 20 fish from a lake.
  - A bunch of 10 flowers made from 21 flowers of 4 different colours.

Definition 5.1.2 The factorial function $n!$ is defined as

$$n! = \begin{cases} 1 & \text{for} \quad n = 0 \\ 1 \cdot 2 \cdot 3 \cdots n & \text{for} \quad n > 0. \end{cases} \tag{5.1}$$

Example 5.1.2 It follows from the definition of the factorial function that

$$0! = 1, \quad 1! = 1 \quad 2! = 1 \cdot 2 = 2, \quad 3! = 1 \cdot 2 \cdot 3 = 6\,.$$

This can be calculated in *R* as follows:

```
factorial(n)
```

5.2 Permutations

Definition 5.2.1 Consider a set of n elements. Each ordered composition of these n elements is called a **permutation**.

We distinguish between two cases: If all the elements are distinguishable, then we speak of *permutation without replacement*. However, if some or all of the elements are not distinguishable, then we speak of *permutation with replacement*. Please note that the meaning of "replacement" here is just a convention and does not directly refer to the drawings, e.g. from the urn model considered in Example 5.1.1.

5.2.1 Permutations without Replacement

If all the n elements are distinguishable, then there are

$$n! \tag{5.2}$$

different compositions of these elements.

Example 5.2.1 There were three candidate cities for hosting the 2020 Olympic Games: Tokyo (T), Istanbul (I), and Madrid (M). Before the election, there were $3! = 6$ possible outcomes, regarding the final rankings of the cities:

$$(\mathrm{M}, \mathrm{T}, \mathrm{I}),\ (\mathrm{M}, \mathrm{I}, \mathrm{T}),\ (\mathrm{T}, \mathrm{M}, \mathrm{I}),\ (\mathrm{T}, \mathrm{I}, \mathrm{M}),\ (\mathrm{I}, \mathrm{M}, \mathrm{T}),\ (\mathrm{I}, \mathrm{T}, \mathrm{M}).$$

5.2.2 Permutations with Replacement

Assume that not all n elements are distinguishable. The elements are divided into groups, and these groups are distinguishable. Suppose, there are s groups of sizes $n_1, n_2, \ldots, n_s$. The total number of different ways to arrange the n elements in s groups is:

$$\frac{n!}{n_1!\, n_2!\, n_3!\, \cdots n_s!}\,. \tag{5.3}$$

Example 5.2.2 Consider the data in Fig. 5.1e. There are two groups consisting of two black balls ($n_1 = 2$) and one white ball ($n_2 = 1$). So there are the following three possible combinations to arrange the balls: (black, black, and white), (black, white, and black), and (white, black, and black). This can be determined by calculating

$$\frac{3!}{2!\ 1!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3\,.$$

5.3 Combinations

Definition 5.3.1 The Binomial coefficient for any integers m and n with $n \geq m \geq 0$ is denoted and defined as

$$\binom{n}{m} = \frac{n!}{m!\,(n-m)!}\,. \tag{5.4}$$

It is read as "n choose m" and can be calculated in *R* using the following command:

```
choose(n,m)
```

There are several calculation rules for the binomial coefficient:

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \binom{n}{m} = \binom{n}{n-m}, \quad \binom{n}{m} = \prod_{i=1}^{m} \frac{n+1-i}{i}\,. \tag{5.5}$$

We now answer the question of how many different possibilities exist to draw m out of n elements, i.e. m out of n balls from an urn. It is necessary to distinguish between the following four cases:

(1) Combinations **without** replacement and **without** consideration of the order of the elements.
(2) Combinations **without** replacement and **with** consideration of the order of the elements.
(3) Combinations **with** replacement and **without** consideration of the order of the elements.
(4) Combinations **with** replacement and **with** consideration of the order of the elements.

5.3.1 Combinations without Replacement and without Consideration of the Order

When there is no replacement and the order of the elements is also not relevant, then the total number of distinguishable combinations in drawing m out of n elements is

$$\binom{n}{m}\,. \tag{5.6}$$

Example 5.3.1 Suppose a company elects a new board of directors. The board consists of 5 members and 15 people are eligible to be elected. How many combinations for the board of directors exist? Since a person cannot be elected twice, we have a

situation where there is no replacement. The order is also of no importance: either one is elected or not. We can thus apply (5.6) which yields

$$\binom{15}{5} = \frac{15!}{10!5!} = 3003$$

possible combinations. This result can be obtained in *R* by using the command `choose(15,5)`.

5.3.2 Combinations without Replacement and with Consideration of the Order

The total number of different combinations for the setting without replacement and with consideration of the order is

$$\frac{n!}{(n-m)!} = \binom{n}{m} m! \,. \tag{5.7}$$

Example 5.3.2 Consider a horse race with 12 horses. A possible bet is to forecast the winner of the race, the second horse of the race, and the third horse of the race. The total number of different combinations for the horses in the first three places is

$$\frac{12!}{(12-3)!} = 12 \cdot 11 \cdot 10 = 1320 \,.$$

This result can be explained intuitively: for the first place, there is a choice of 12 different horses. For the second place, there is a choice of 11 different horses (12 horses minus the winner). For the third place, there is a choice of 10 different horses (12 horses minus the first and second horses). The total number of combinations is the product $12 \cdot 11 \cdot 10$. This can be calculated in *R* as follows:

```
12 * 11 * 10
```

5.3.3 Combinations with Replacement and without Consideration of the Order

The total number of different combinations with replacement and without consideration of the order is

$$\binom{n+m-1}{m} = \frac{(n+m-1)!}{m!\,(n-1)!} = \binom{n+m-1}{n-1} \,. \tag{5.8}$$

Note that these are the two representations which follow from the definition of the binomial coefficient but typically only the first representation is used in textbooks. We will motivate the second representation after Example 5.3.3.

Example 5.3.3 A farmer has 2 fields and aspires to cultivate one out of 4 different organic products per field. Then, the total number of choices he has is

$$\binom{4+2-1}{2} = \binom{5}{2} = \frac{5!}{2!\,3!} = \frac{3!\cdot 4\cdot 5}{1\cdot 2\cdot 3!} = 10. \tag{5.9}$$

If 4 different organic products are denoted as a, b, c, and d, then the following combinations are possible:

(a, a)	(a, b)	(a, c)	(a, d)
	(b, b)	(b, c)	(b, d)
		(c, c)	(c, d)
			(d, d)

Please note that, for example, (a,b) is identical to (b,a) because the order in which the products a and b are cultivated on the first or second field is not important in this example.

We now try to give an intuitive explanation of formula (5.9) using Example 5.3.3. We have $n = 4$ products and $m = 2$ fields and apply the following technical "trick": we sort the combinations by the product symbols (a, b, c, or d). When we switch from one product to the next (e.g. from b to c), we make a note by adding a vertical line |. Whenever a product is skipped, we add a line too. For example, the combination (a, c) is denoted by $a||c|$, the combination (d, d) by $|||dd$, (c, c) by $||cc|$, and (a, a) by $aa|||$. Therefore, the number of characters equates to the 2 chosen symbols of the set (a, b, c, d) plus the 3 vertical lines, in summary $(4 + 2) - 1 = 5$ places where $3 = n - 1$ places are selected for the vertical line |. How many different line/letter combinations exist? There are 3 out of 5 possible positions for |, i.e. $\binom{5}{3} = 10$ possible combinations, and this is nothing but the right-hand side of (5.9).

5.3.4 Combinations with Replacement and with Consideration of the Order

The total number of different combinations for the integers m and n with replacement and when the order is of relevance is

$$n^m\,. \tag{5.10}$$

Example 5.3.4 Consider a credit card with a four-digit personal identification number (PIN) code. The total number of possible combinations for the PIN is

$$n^m = 10^4 = 10,000.$$

Note that every digit in the first, second, third, and fourth places ($m = 4$) can be chosen out of ten digits from 0 to 9 ($n = 10$).

5.4 Key Points and Further Issues

Note:

✓ The rules of combinatorics are as follows:

Combinations	without replacement	with replacement
without order	$\binom{n}{m}$	$\binom{n+m-1}{m}$
with order	$\binom{n}{m} m!$	n^m

✓ Combinations with and without *replacement* are also often called combinations with and without *repetition*.

✓ The permutation rules are as follows:

	without replacement	with replacement
Permutations	$n!$	$\frac{n!}{n_1! \cdots n_s!}$

5.5 Exercises

Exercise 5.1 At a party with 10 guests, every guest shakes hands with each other guest. How many handshakes can be counted in total?

Exercise 5.2 A language teacher is concerned about the vocabularies of his students. He thus tests 5 students in each lecture. What are the total number of possible combinations

(a) if a student is tested only once per lecture and
(b) if a student is tested more than once per lecture?

Use *R* to quantify numbers which you cannot calculate manually.

Exercise 5.3 "Gobang" is a popular game in which two players set counters on a board with 381 knots. One needs to place 5 consecutive counters in a row to win the game. There are also rules on how to remove counters from the other player. Consider a match where 64 counters have already been placed on the board. How many possible combinations exist to place 64 counters on the board?

Exercise 5.4 A shop offers a special tray of beer: "Munich's favourites". Customers are allowed to fill the tray, which holds 20 bottles, with any combination of Munich's 6 most popular beers (from 6 different breweries).

(a) What are the number of possible combinations to fill the tray?
(b) A customer insists of having at least one beer from each brewery in his tray. How many options does he have to fill the tray?

Exercise 5.5 The FIFA World Cup 2018 in Russia consists of 32 teams. How many combinations for the top 3 teams exist when

(a) taking into account the order of these top 3 teams and
(b) *without* taking into account the order of these top 3 teams?

Exercise 5.6 An online book store assigns membership codes to each member. For administrative reasons, these codes consist of four letters between "A" and "L". A special discount period increased the total number of members from 18, 200 to 20, 500. Are there enough combinations of codes left to be assigned for the new membership codes?

Exercise 5.7 In the old scoring system of ice skating (valid until 2004), each member of a jury of 9 people judged the performance of the skaters on a scale between 0 and 6. It was a decimal scale and thus scores such as 5.1 and 5.2 were possible. Calculate the number of possible score combinations from the jury.

Exercise 5.8 It is possible in Pascal's triangle (Fig. 5.2, left) to view each entry as the sum of the two entries directly above it. For example, the 3 on the fourth line

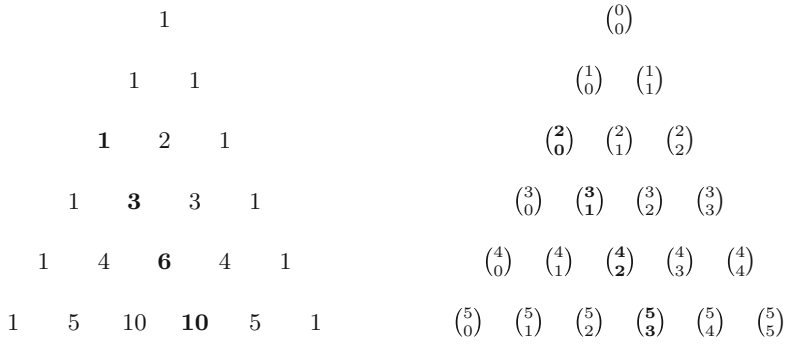


Fig. 5.2 Excerpt from Pascal's triangle (*left*) and its representation by means of binomial coefficients (*right*)

from the top is the sum of the 1 and 2 above the 3. Another interpretation refers to a geometric representation of the binomial coefficient, $\binom{n}{k}$ (Fig. 5.2, right) with $k = 0, 1, 2, \ldots$ being the column index and $n = 0, 1, 2, \ldots$ being the row index.

(a) Show that each entry in the bold third diagonal line can be represented via $\binom{n}{2}$.
(b) Now show that the sum of two consecutive entries in the bold third diagonal line always corresponds to quadratic numbers.

→ Solutions to all exercises in this chapter can be found on p. 358

Elements of Probability Theory

6

Let us first consider some simple examples to understand the need for probability theory. Often one needs to make a decision whether to carry an umbrella or not when leaving the house; a company might wonder whether to introduce a new advertisement to possibly increase sales or to continue with their current advertisement; or someone may want to choose a restaurant based on where he can get his favourite dish. In all these situations, randomness is involved. For example, the decision of whether to carry an umbrella or not is based on the possibility or chance of rain. The sales of the company may increase, decrease, or remain unchanged with a new advertisement. The investment in a new advertising campaign may therefore only be useful if the probability of its success is higher than that of the current advertisement. Similarly, one may choose the restaurant where one is most confident of getting the food of one's choice. In all such cases, an event may be happening or not and depending on its likelihood, actions are taken. The purpose of this chapter is to learn how to calculate such likelihoods of events happening and not happening.

6.1 Basic Concepts and Set Theory

A simple (not rigorous) definition of a **random experiment** requires that the experiment can be repeated any number of times under the same set of conditions, and its outcome is known only after the completion of the experiment. A simple and classical example of a random experiment is the tossing of a coin or the rolling of a die. When tossing a coin, it is unknown what the outcome will be, head or tail, until the coin is tossed. The experiment can be repeated and different outcomes may be observed in each repetition. Similarly, when rolling a die, it is unknown how many dots will appear on the upper surface until the die is rolled. Again, the die can be rolled repeatedly and different numbers of dots are obtained in each trial. A possible

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_6

outcome of a random experiment is called a **simple event** (or **elementary event**) and denoted by ω_i. The set of all possible outcomes, $\{\omega_1, \omega_2, \ldots, \omega_k\}$, is called the **sample space** and is denoted as Ω, i.e. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$. Subsets of Ω are called **events** and are denoted by capital letters such as A, B, C. The set of all simple events that are contained in the event A is denoted by Ω_A. The event $\bar{A}$ refers to the non-occurring of A and is called **a composite or complementary event**. Also Ω is an event. Since it contains all possible outcomes, we say that Ω will always occur and we call it a **sure event** or **certain event**. On the other hand, if we consider the null set $\emptyset = \{\}$ as an event, then this event can never occur and we call it an **impossible event**. The sure event therefore is the set of all elementary events, and the impossible event is the set with no elementary events.

The above concepts of "events" form the basis of a definition of "probability". Once we understand the concept of probability, we can develop a framework to make conclusions about the population of interest, using a sample of data.

Example 6.1.1 (Rolling a die) If a die is rolled once, then the possible outcomes are the number of dots on the upper surface: $1, 2, \ldots, 6$. Therefore, the sample space is the set of simple events $\omega_1 =$ "1", $\omega_2 =$ "2", $\ldots$, $\omega_6 =$ "6" and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$. Any subset of Ω can be used to define an event. For example, an event A may be "an even number of dots on the upper surface of the die". There are three possibilities that this event occurs: ω_2, ω_4, or ω_6. If an odd number shows up, then the composite event $\bar{A}$ occurs instead of A. If an event is defined to observe only one particular number, say $\omega_1 =$ "1", then it is an elementary event. An example of a sure event is "a number which is greater than or equal to 1" because any number between 1 and 6 is greater than or equal to 1. An impossible event is "the number is 7".

Example 6.1.2 (Rolling two dice) Suppose we throw two dice simultaneously and an event is defined as the "number of dots observed on the upper surface of both the dice"; then, there are 36 simple events defined as (number of dots on first die, number of dots on second die), i.e. $\omega_1 = (1, 1), \omega_2 = (1, 2), \ldots, \omega_{36} = (6, 6)$. Therefore Ω is

$$\Omega = \begin{matrix} \{(1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6)\}. \end{matrix}$$

One can define different events and their corresponding sample spaces. For example, if an event A is defined as "upper faces of both the dice contain the same number of dots", then the sample space is $\Omega_A = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$. If another event B is defined as "the sum of numbers on the upper faces is 6", then

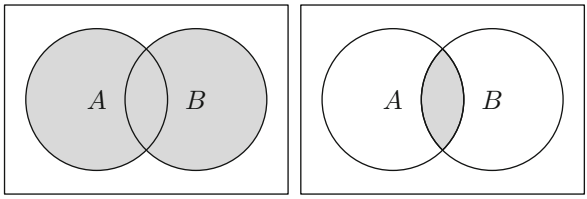


Fig. 6.1 $A \cup B$ and $A \cap B^*$

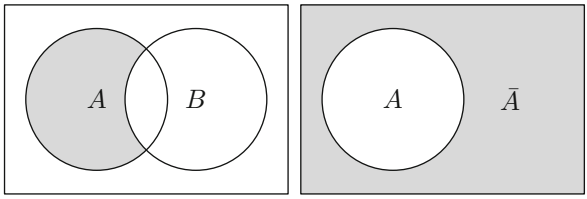


Fig. 6.2 $A \backslash B$ and $\bar{A} = \Omega \backslash A^*$

the sample space is $\Omega_B = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$. A sure event is "get either an even number or an odd number"; an impossible event would be "the sum of the two dice is greater than 13".

It is possible to view events as sets of simple events. This helps to determine how different events relate to each other. A popular technique to visualize this approach is to use **Venn diagrams**. In Venn diagrams, two or more sets are visualized by circles. Overlapping circles imply that both events have one or more identical simple events. Separated circles mean that none of the simple events of event A are contained in the sample space of B. We use the following notations:

$A \cup B$ The union of events $A \cup B$ is the set of all simple events of A and B which occurs if at least one of the simple events of A *or* B occurs (Fig. 6.1, left side, grey shaded area). Please note that we use the word "or" from a statistical perspective: "A or B" means that either a simple event from A occurs, or a simple event from B occurs, or a simple event which is part of both A and B occurs.

$A \cap B$ The intersection of events $A \cap B$ is the set of all simple events A and B which occur when a simple event occurs that belongs to A *and* B (Fig. 6.1, right side, grey shaded area).

$A \backslash B$ The event $A \backslash B$ contains all simple events of A, which are not contained in B. The event "A but not B" or "A minus B" occurs, if A occurs but B does not occur. Also $A \backslash B = A \cap \bar{B}$ (Fig. 6.2, left side, grey shaded area).

$\bar{A}$ The event $\bar{A}$ contains all simple events of Ω, which are not contained in A. The complementary event of A (which is "Not-A" or "$\bar{A}$" occurs whenever A does not occur (Fig. 6.2, right side, grey shaded area).

$A \subseteq B$ A is a subset of B. This means that all simple events of A are also part of the sample space of B.

Example 6.1.3 Consider Example 6.1.1 where the sample space of rolling a die was determined as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$ with $\omega_1 =$ "1", $\omega_2 =$ "2", ..., $\omega_6 =$ "6".

- If $A = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ and B is the set of all odd numbers, then $B = \{\omega_1, \omega_3, \omega_5\}$ and thus $B \subseteq A$.
- If $A = \{\omega_2, \omega_4, \omega_6\}$ is the set of even numbers and $B = \{\omega_3, \omega_6\}$ is the set of all numbers which are divisible by 3, then $A \cup B = \{\omega_2, \omega_3, \omega_4, \omega_6\}$ is the collection of simple events for which the number is either even or divisible by 3 or both.
- If $A = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers and $B = \{\omega_3, \omega_6\}$ is the set of the numbers which are divisible by 3, then $A \cap B = \{\omega_3\}$ is the set of simple events in which the numbers are odd and divisible by 3.
- If $A = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers and $B = \{\omega_3, \omega_6\}$ is the set of the numbers which are divisible by 3, then $A \backslash B = \{\omega_1, \omega_5\}$ is the set of simple events in which the numbers are odd but not divisible by 3.
- If $A = \{\omega_2, \omega_4, \omega_6\}$ is the set of even numbers, then $\bar{A} = \{\omega_1, \omega_3, \omega_5\}$ is the set of odd numbers.

Remark 6.1.1 Some textbooks also use the following notations:

$$\begin{array}{lll} A + B & \text{for} & A \cup B \\ AB & \text{for} & A \cap B \\ A - B & \text{for} & A \backslash B. \end{array}$$

We can use these definitions and notations to derive the following properties of a particular event A:

$$\begin{array}{ll} A \cup A = A & A \cap A = A \\ A \cup \Omega = \Omega & A \cap \Omega = A \\ A \cup \emptyset = A & A \cap \emptyset = \emptyset \\ A \cup \bar{A} = \Omega & A \cap \bar{A} = \emptyset. \end{array}$$

Definition 6.1.1 Two events A and B are ***disjoint*** if $A \cap B = \emptyset$ holds, i.e. if both events cannot occur simultaneously.

Example 6.1.4 The events A and $\bar{A}$ are disjoint events.

Definition 6.1.2 The events $A_1, A_2, \ldots, A_m$ are said to be mutually or pairwise disjoint, if $A_i \cap A_j = \emptyset$ whenever $i \neq j = 1, 2, ..., m$.

Example 6.1.5 Recall Example 6.1.1. If $A = \{\omega_1, \omega_3, \omega_5\}$ and $B = \{\omega_2, \omega_4, \omega_6\}$ are the sets of odd and even numbers, respectively, then the events A and B are disjoint.

Definition 6.1.3 The events $A_1, A_2, \ldots, A_m$ form a **complete decomposition** of Ω if and only if

$$A_1 \cup A_2 \cup \cdots \cup A_m = \Omega$$

and

$$A_i \cap A_j = \emptyset \quad (\text{for all } i \neq j).$$

Example 6.1.6 Consider Example 6.1.1. The elementary events $A_1 = \{\omega_1\}$, $A_2 = \{\omega_2\}, \ldots, A_6 = \{\omega_6\}$ form a complete decomposition. Other complete decompositions are, e.g.

- $A_1 = \{\omega_1, \omega_3, \omega_5\}, \quad A_2 = \{\omega_2, \omega_4, \omega_6\}$
- $A_1 = \{\omega_1\}, \quad A_2 = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$
- $A_1 = \{\omega_1, \omega_2, \omega_3\}, \quad A_2 = \{\omega_4, \omega_5, \omega_6\}.$

6.2 Relative Frequency and Laplace Probability

There is a close connection between the relative frequency and the probability of an event. A random experiment is described by its possible outcomes, for example getting a number between 1 and 6 when rolling a die. Suppose an experiment has m possible outcomes (events) $A_1, A_2, \ldots, A_m$ and the experiment is repeated n times. Now we can count how many times each of the possible outcome has occurred. In other words, we can calculate the absolute frequency $n_i = n(A_i)$ which is equal to the number of times an event A_i, $i = 1, 2, \ldots, m$, occurs. The relative frequency $f_i = f(A_i)$ of a random event A_i, with n repetitions of the experiment, is calculated as

$$f_i = f(A_i) = \frac{n_i}{n}. \tag{6.1}$$

Example 6.2.1 Consider roulette, a game frequently played in casinos. The roulette table consists of 37 numbers from 0 to 36. Out of these 37 numbers, 18 numbers are red, 18 are black and one (zero) is green. Players can place their bets on either a single number or a range of numbers, the colours red or black, whether the number is odd or even, among many other choices. A casino employee spins a wheel (containing pockets representing the 37 numbers) in one direction and then spins a ball over the wheel in the opposite direction. The wheel and ball gradually slow down and the ball finally settles in a pocket. The pocket number in which the ball sits down when the wheel stops is the winning number. Consider three possible outcomes A_1: "red", A_2:"black", and A_3: "green (zero)". Suppose the roulette ball is spun $n = 500$ times. All the outcomes are counted and recorded as follows: A_1 occurs 240 times, A_2 occurs 250 times and A_3 occurs 10 times. Then, the absolute frequencies are given

by $n_1 = n(A_1) = 240$, $n_2 = n(A_2) = 250$, and $n_3 = n(A_3) = 10$. We therefore get the relative frequencies as

$$f_1 = f(A_1) = \frac{240}{500} = 0.48\,, \quad f_2 = f(A_2) = \frac{250}{500} = 0.5,$$
$$f_3 = f(A_3) = \frac{10}{500} = 0.02.$$

If we assume that the experiment is repeated a large number of times (mathematically, this would mean that n tends to infinity) and the experimental conditions remain the same (at least approximately) over all the repetitions, then the relative frequency $f(A)$ converges to a limiting value for A. This limiting value is interpreted as the probability of A and denoted by $P(A)$, i.e.

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

where $n(A)$ denotes the number of times an event A occurs out of n times.

Example 6.2.2 Suppose a fair coin is tossed $n = 20$ times and we observe the number of heads $n(A_1) = 8$ times and number of tails $n(A_2) = 12$ times. The meaning of a fair coin in this case is that the probabilities of head and tail are equal (i.e. 0.5). Then, the relative frequencies in the experiment are $f(A_1) = 8/20 = 0.4$ and $f(A_2) = 12/20 = 0.6$. When the coin is tossed a large number of times and n tends to infinity, then both $f(A_1)$ and $f(A_2)$ will have a limiting value 0.5 which is simply the probability of getting a head or tail in tossing a fair coin.

Example 6.2.3 In Example 6.2.1, the relative frequency of $f(\text{red}) = f(A_1)$ tends to $18/37$ as n tends to infinity because 18 out of 37 numbers are red.

The reader will gain a more theoretical understanding of how repeated experiments relate to expected quantities in the following chapters after learning the Theorem of Large Numbers described in Appendix A.3.

A different definition of probability was given by Pierre-Simon Laplace (1749–1827). We call an experiment a **Laplace experiment** if the number of possible simple events is finite and all the outcomes are equally probable. The probability of an arbitrary event A is then defined as follows:

Definition 6.2.1 The proportion

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of "favourable simple events" for } A}{\text{Total number of possible simple events}} \tag{6.2}$$

is called the **Laplace probability**, where $|A|$ is the cardinal number of A, i.e. the number of simple events contained in the set A, and $|\Omega|$ is the cardinal number of Ω, i.e. the number of simple events contained in the set Ω.

The cardinal numbers $|A|$ and $|\Omega|$ are often calculated using the combinatoric rules introduced in Chap. 5.

Example 6.2.4 (Example 6.1.2 continued) The sample space contains 36 simple events. All of these simple events have equal probability 1/36. To calculate the probability of the event A that the sum of the dots on the two dice is at least 4 and at most 6, we count the favourable simple events which fulfil this condition. The simple events are (1, 3), (2, 2), (3, 1) (sum is 4), (1, 4), (2, 3), (4, 1), (3, 2) (sum is 5) and (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) (sum is 6). In total, there are $(3 + 4 + 5) = 12$ favourable simple events, i.e.

$$A = \{(1, 3), (2, 2), (3, 1), (1, 4), (2, 3), (4, 1), \\ (3, 2), (1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} .$$

The probability of the event A is therefore $12/36 = 1/3$.

6.3 The Axiomatic Definition of Probability

An important foundation for modern probability theory was established by A.N. Kolmogorov in 1933 when he proposed the following **axioms of probability**.

Axiom 1 Every random event A has a probability in the (closed) interval [0, 1], i.e.

$$0 \leq P(A) \leq 1.$$

Axiom 2 The sure event has probability 1, i.e.

$$P(\Omega) = 1.$$

Axiom 3 If A_1 and A_2 are disjoint events, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

holds.

Remark Axiom 3 also holds for three or more disjoint events and is called the **theorem of additivity of disjoint events**. For example, if A_1, A_2, and A_3 are disjoint events, then $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$.

Example 6.3.1 Suppose the two events in tossing a coin are A_1: "appearance of head" and A_2: "appearance of tail" which are disjoint. The event $A_1 \cup A_2$: "appearance of head or tail" has the probability

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) = 1/2 + 1/2 = 1.$$

Example 6.3.2 Suppose an event is defined as the number of points observed on the upper surface of a die when rolling it. There are six events, i.e. the natural numbers 1, 2, 3, 4, 5, 6. These events are disjoint and they have equal probability of occurring: $P(1) = P(2) = \cdots = P(6) = 1/6$. The probability of getting an even number is then

$$P(\text{"even number"}) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 1/2.$$

6.3.1 Corollaries Following from Kolomogorov's Axioms

We already know that $A \cup \bar{A} = \Omega$ (sure event). Since A and $\bar{A}$ are disjoint, using Axiom 3 we have

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1\,.$$

Based on this, we have the following corollaries.

Corollary 1 *The probability of the complementary event of A, (i.e. $\bar{A}$) is*

$$P(\bar{A}) = 1 - P(A). \tag{6.3}$$

Example 6.3.3 Suppose a box of 30 chocolates contains chocolates of 6 different flavours with 5 chocolates of each flavour. Suppose an event A is defined as $A = \{$"marzipan flavour"$\}$. The probability of finding a marzipan chocolate (without looking into the box) is $P(\text{"marzipan"}) = 5/30$. Then, the probability of the complementary event $\bar{A}$, i.e. the probability of not finding a marzipan chocolate is therefore

$$P(\text{"no marzipan flavour"}) = 1 - P(\text{"marzipan flavour"}) = 25/30.$$

Corollary 2 *The probability of occurrence of an impossible event $\emptyset$ is zero:*

$$P(\emptyset) = P(\bar{\Omega}) = 1 - P(\Omega) = 0.$$

Corollary 3 *Let A_1 and A_2 be not necessarily disjoint events. The probability of occurrence of A_1 or A_2 is*

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \tag{6.4}$$

The rule in (6.4) is known as **the additive theorem of probability**. Again we use the word "or" in the statistical sense: either A_1 is occurring, A_2 is occurring, or both of them. This means we have to add the probabilities $P(A_1)$ and $P(A_2)$ but need to make sure that the simple events which are contained in both sets are not counted twice, thus we subtract $P(A_1 \cap A_2)$.

Example 6.3.4 There are 10 actors acting in a play. Two actors, one of whom is male, are portraying evil characters. In total, there are 6 female actors. Let an event A describe whether the actor is male and another event B describe whether the character is evil. Suppose we want to know the probability of a randomly chosen actor being male or evil. We can then calculate

$$\begin{aligned} &P(\text{actor is male or evil}) = \\ &= P(\text{actor is male}) + P(\text{actor is evil}) - P(\text{actor is male and evil}) \\ &= \frac{4}{10} + \frac{2}{10} - \frac{1}{10} = \frac{1}{2}\,. \end{aligned}$$

Corollary 4 *If $A \subseteq B$ then $P(A) \leq P(B)$.*

Proof We use the representation $B = A \cup (\bar{A} \cap B)$ where A and $\bar{A} \cap B$ are the disjoint events. Then using Axiom 3 and Axiom 1, we get

$$P(B) = P(A) + P(\bar{A} \cap B) \geq P(A)\,.$$

6.3.2 Calculation Rules for Probabilities

The introduced axioms and corollaries can be summarized as follows:

(1) $0 \leq P(A) \leq 1$
(2) $P(\Omega) = 1$
(3) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, if A_1 and A_2 are disjoint
(4) $P(\emptyset) = 0$
(5) $P(\bar{A}) = 1 - P(A)$
(6) $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
(7) $P(A) \leq P(B)$, if $A \subseteq B$

6.4 Conditional Probability

Consider the following example to understand the concept of conditional probability: Suppose a new medical test is developed to diagnose a particular infection of the blood. The test is conducted on blood samples from 100 randomly selected patients and the outcomes of the tests are presented in Table 6.1.

There are the following four possible outcomes:

- The blood sample has an infection and the test diagnoses it, i.e. the test is correctly diagnosing the infection.
- The blood sample does not have an infection and the test does not diagnose it, i.e. the test is correctly diagnosing that there is no infection.
- The blood sample has an infection and the test does not diagnose it, i.e. the test is incorrect in stating that there is no infection.
- The blood sample does not have an infection but the test diagnoses it, i.e. the test is incorrect in stating that there is an infection.

Table 6.2 contains the relative frequencies of Table 6.1. In the following, we interpret the relative frequencies as probabilities, i.e. we assume that the values in Table 6.2 would be observed if the number n of patients was much larger than 100.

It can be seen that the probability that a test is positive is $P(T+) = 0.30 + 0.10 = 0.40$ and the probability that an infection is present is $P(IP) = 0.30 + 0.15 = 0.45$.

Table 6.1 Absolute frequencies of test results and infection status

		Infection		Total (row)
		Present	Absent	
Test	Positive (+)	30	10	40
	Negative (−)	15	45	60
	Total (column)	45	55	Total = 100

Table 6.2 Relative frequencies of patients and test

		Infection		Total (row)
		Present (IP)	Absent (IA)	
Test	Positive (+)	0.30	0.10	0.40
	Negative (−)	0.15	0.45	0.60
	Total (column)	0.45	0.55	Total = 1

If one already knows that the test is positive and wants to determine the probability that the infection is indeed present, then this can be achieved by the respective **conditional probability** $P(IP|T+)$ which is

$$P(IP|T+) = \frac{P(IP \cap T+)}{P(T+)} = \frac{0.3}{0.4} = 0.75.$$

Note that $IP \cap T+$ denotes the "relative frequency of blood samples in which the disease is present *and* the test is positive" which is 0.3.

More generally, recall Definition 4.1.1 from Chap. 4 where we defined conditional, joint, and marginal frequency distributions in contingency tables. The present example simply applies these rules to the contingency tables of relative frequencies and interprets the relative frequencies as an approximation to the probabilities of interest, as already explained.

We use the intersection operator $\cap$ to describe events which occur for $A = a$ *and* $B = b$. This relates to the joint relative frequencies. The marginal relative frequencies (i.e. probabilities $P(A = a)$) can be observed from the column and row sums, respectively; and the conditional probabilities can be observed as the joint frequencies in relation to the marginal frequencies.

For simplicity, assume that all simple events in $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ are equally probable, i.e. $P(\omega_j) = \frac{1}{k}$, $j = 1, 2, \ldots, k$. Let A and B be two events containing n_A and n_B numbers of simple events. Let further $A \cap B$ contain n_{AB} numbers of simple events. The Laplace probability using (6.2) is

$$P(A) = \frac{n_A}{k}, \quad P(B) = \frac{n_B}{k}, \quad P(A \cap B) = \frac{n_{AB}}{k}.$$

Assume that we have prior information that A has already occurred. Now we want to find out how the probability of B is to be calculated. Since A has already occurred, we know that the sample space is reduced by the number of simple events which

are contained in A. There are n_A such simple events. Thus, the total sample space Ω is reduced by the sample space of A. Therefore, the simple events in $A \cap B$ are those simple events which are realized when B is realized. The Laplace probability for B under the prior information on A, or under the condition that A is known, is therefore

$$P(B|A) = \frac{n_{AB}/k}{n_A/k} = \frac{P(A \cap B)}{P(A)}. \tag{6.5}$$

This can be generalized to the case when the probabilities for simple events are unequal.

Definition 6.4.1 Let $P(A) > 0$. Then the **conditional probability** of event B occurring, given that event A has already occurred, is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{6.6}$$

The roles of A and B can be interchanged to define $P(A|B)$ as follows. Let $P(B) > 0$. The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{6.7}$$

We now introduce a few important theorems which are relevant to calculating conditional and other probabilities.

Theorem 6.4.1 (Multiplication Theorem of Probability) *For two arbitrary events A and B, the following holds:*

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \tag{6.8}$$

This theorem follows directly from the two definitions (6.6) and (6.7) (but does not require that $P(A) > 0$ and $P(B) > 0$).

Theorem 6.4.2 (Law of Total Probability) *Assume that $A_1, A_2, \ldots, A_m$ are events such that $\cup_{i=1}^m A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$, $P(A_i) > 0$ for all i, i.e. $A_1, A_2, \ldots, A_m$ form a complete decomposition of $\Omega = \cup_{i=1}^m A_i$ in pairwise disjoint events, then the probability of an event B can be calculated as*

$$P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i)\,. \tag{6.9}$$

6.4.1 Bayes' Theorem

Bayes' Theorem gives a connection between $P(A|B)$ and $P(B|A)$. For events A and B with $P(A) > 0$ and $P(B) > 0$, using (6.6) and (6.7) or (6.8), we get

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A)} \frac{P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B)} . \end{aligned} \tag{6.10}$$

Let $A_1, A_2, \ldots, A_m$ be events such that $\cup_{i=1}^m A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$, $P(A_i) > 0$ for all i, and B is another event than A, then using (6.9) and (6.10), we get

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)} . \tag{6.11}$$

The probabilities $P(A_i)$ are called **prior probabilities**, $P(B|A_i)$ are sometimes called **model probabilities** and $P(A_j|B)$ are called **posterior probabilities**.

Example 6.4.1 Suppose someone rents movies from two different DVD stores. Sometimes it happens that the DVD does not work because of scratches. We consider the following events: A_i $(i = 1, 2)$: "the DVD is rented from store i". Further let B denote the event that the DVD is working without any problems. Assume we know that $P(A_1) = 0.6$ and $P(A_2) = 0.4$ (note that $A_2 = \bar{A}_1$) and $P(B|A_1) = 0.95$, $P(B|A_2) = 0.75$ and we are interested in the probability that a rented DVD works fine. We can then apply the Law of Total Probability and get

$$\begin{aligned} P(B) &\overset{(6.9)}{=} P(B|A_1)P(A_1) + P(B|A_2)P(A_2) \\ &= 0.6 \cdot 0.95 + 0.4 \cdot 0.75 = 0.87. \end{aligned}$$

We may also be interested in the probability that the movie was rented from store 1 *and* is working which is

$$P(B \cap A_1) \overset{(6.8)}{=} P(B|A_1)P(A_1) = 0.95 \cdot 0.6 = 0.57.$$

Now suppose we have a properly working DVD. What is the probability that it is rented from store 1? This is obtained as follows:

$$P(A_1|B) \overset{(6.7)}{=} \frac{P(A_1 \cap B)}{P(B)} = \frac{0.57}{0.87} = 0.6552.$$

Now assume we have a DVD which does not work, i.e. $\bar{B}$ occurs. The probability that a DVD is not working given that it is from store 1 is $P(\bar{B}|A_1) = 0.05$. Similarly,

$P(\bar{B}|A_2) = 0.25$ for store 2. We can now calculate the conditional probability that a DVD is from store 1 given that it is not working:

$$P(A_1|\bar{B}) \overset{(6.11)}{=} \frac{P(\bar{B}|A_1)P(A_1)}{P(\bar{B}|A_1)P(A_1) + P(\bar{B}|A_2)P(A_2)}$$
$$= \frac{0.05 \cdot 0.6}{0.05 \cdot 0.6 + 0.25 \cdot 0.4} = 0.2308.$$

The result about $P(\bar{B})$ used in the denominator can also be directly obtained by using $P(\bar{B}) = 1 - 0.87 = 0.13$.

6.5 Independence

Intuitively, two events are independent if the occurrence or non-occurrence of one event does not affect the occurrence or non-occurrence of the other event. In other words, two events A and B are independent if the probability of occurrence of B has no effect on the probability of occurrence of A. In such a situation, one expects that

$$P(A|B) = P(A) \quad \text{and} \quad P(A|\bar{B}) = P(A)\,.$$

Using this and (6.7), we can write

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$= \frac{P(A \cap \bar{B})}{P(\bar{B})} = P(A|\bar{B}). \tag{6.12}$$

This yields:

$$\begin{aligned} P(A \cap B)P(\bar{B}) &= P(A \cap \bar{B})P(B) \\ P(A \cap B)(1 - P(B)) &= P(A \cap \bar{B})P(B) \\ P(A \cap B) &= (P(A \cap \bar{B}) + P(A \cap B))P(B) \\ P(A \cap B) &= P(A)P(B)\,. \end{aligned} \tag{6.13}$$

This leads to the following definition of stochastic independence.

Definition 6.5.1 Two random events A and B are called **(stochastically) independent** if

$$P(A \cap B) = P(A)P(B)\,, \tag{6.14}$$

i.e. if the probability of simultaneous occurrence of both events A and B is the product of the individual probabilities of occurrence of A and B.

This definition of independence can be extended to the case of more than two events as follows:

Definition 6.5.2 The n events $A_1, A_2, \ldots, A_n$ are stochastically mutually independent, if for any subset of m events $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$ $(m \leq n)$

$$P(A_{i_1} \cap A_{i_2} \cdots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdot \ldots \cdot P(A_{i_m}) \tag{6.15}$$

holds.

A weaker form of independence is pairwise independence. If condition (6.15) is fulfilled only for two arbitrary events, i.e. $m = 2$, then the events are called **pairwise independent**. The difference between pairwise independence and general stochastic independence is explained in the following example.

Example 6.5.1 Consider an urn with four balls. The following combinations of zeroes and ones are printed on the balls: 110, 101, 011, 000. One ball is drawn from the urn. Define the following events:

A_1 : The first digit on the ball is 1.
A_2 : The second digit on the ball is 1.
A_3 : The third digit on the ball is 1.

Since there are two favourable simple events for each of the events A_1, A_2 and A_3, we get

$$P(A_1) = P(A_2) = P(A_3) = \frac{2}{4} = \frac{1}{2}.$$

The probability that all the three events simultaneously occur is zero because there is no ball with 111 printed on it. Therefore, A_1, A_2, and A_3 are not stochastically independent because

$$P(A_1)P(A_2)P(A_3) = \frac{1}{8} \neq 0 = P(A_1 \cap A_2 \cap A_3).$$

However,

$$P(A_1 \cap A_2) = \frac{1}{4} = P(A_1)P(A_2)\,,$$
$$P(A_1 \cap A_3) = \frac{1}{4} = P(A_1)P(A_3)\,,$$
$$P(A_2 \cap A_3) = \frac{1}{4} = P(A_2)P(A_3)\,,$$

which means that the three events are pairwise independent.

6.6 Key Points and Further Issues

Note:

✓ We summarize some important theorems and laws:

- The Laplace probability is the ratio

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Number of "favourable simple events" for } A}{\text{Total number of possible simple events}}.$$

- The Law of Total Probability is

$$P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i).$$

- Bayes' Theorem is

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}.$$

- n events $A_1, A_2, \ldots, A_n$ are (stochastically) independent, if

$$P(A_1 \cap A_2 \cdots \cap A_n) = P(A_1)P(A_2) \cdot \ldots \cdot P(A_n).$$

✓ In Sect. 10.8, we present the χ^2-independence test, which can test whether discrete random variables (see Chap. 7) are independent or not.

6.7 Exercises

Exercise 6.1

(a) Suppose $\Omega = \{0, 1, \ldots, 15\}$, $A = \{0, 8\}$, $B = \{1, 2, 3, 5, 8, 10, 12\}$, $C = \{0, 4, 9, 15\}$. Determine $A \cap B$, $B \cap C$, $A \cup C$, $C \setminus A$, $\Omega \setminus (B \cup A \cup C)$.
(b) Now consider the three pairwise disjoint events E, F, G with $\Omega = E \cup F \cup G$ and $P(E) = 0.2$ and $P(F) = 0.5$. Calculate $P(\bar{F})$, $P(G)$, $P(E \cap G)$, $P(E \setminus E)$, and $P(E \cup F)$.

Exercise 6.2 A driving licence examination consists of two parts which are based on a theoretical and a practical examination. Suppose 25 % of people fail the practical examination, 15 % of people fail the theoretical examination, and 10 % of people fail both the examinations. If a person is randomly chosen, then what is the probability that this person

(a) fails at least one of the examinations?
(b) only fails the practical examination, but not the theoretical examination?
(c) successfully passes both the tests?
(d) fails any of the two examinations?

Exercise 6.3 A new board game uses a twelve-sided die. Suppose the die is rolled once, what is the probability of getting

(a) an even number?
(b) a number greater than 9?
(c) an even number greater than 9?
(d) an even number or a number greater than 9?

Exercise 6.4 The Smiths are a family of six. They are celebrating Christmas and there are 12 gifts, two for each family member. The name tags for each family member have been attached to the gifts. Unfortunately the name tags on the gifts are damaged by water. Suppose each family member draws two gifts at random. What is the probability that someone

(a) gets his/her two gifts, rather than getting the gifts for another family member?
(b) gets none of his/her gifts, but rather gets the gifts for other family members?

Exercise 6.5 A chef from a popular TV cookery show sometimes puts too much salt in his pumpkin soup and the probability of this happening is 0.2. If he is in love (which he is with probability 0.3), then the probability of using too much salt is 0.6.

(a) Create a contingency table for the probabilities of the two variables "in love" and "too much salt".
(b) Determine whether the two variables are stochastically independent or not.

Exercise 6.6 Dr. Obermeier asks his neighbour to take care of his basil plant while he is away on leave. He assumes that his neighbour does not take care of the basil with a probability of $\frac{1}{3}$. The basil dies with probability $\frac{1}{2}$ when someone takes care of it and with probability $\frac{3}{4}$ if no one takes care of it.

(a) Calculate the probability of the basil plant surviving after its owner's leave.
(b) It turns out that the basil eventually dies. What is the probability that Dr. Obermeier's neighbour did not take care of the plant?

Exercise 6.7 A bank considers changing its credit card policy. Currently 5 % of credit card owners are not able to pay their bills in any month, i.e. they never pay their bills. Among those who are generally able to pay their bills, there is still a 20 % probability that the bill is paid too late in a particular month.

(a) What is the probability that someone is not paying his bill in a particular month?
(b) A credit card owner did not pay his bill in a particular month. What is the probability that he never pays back the money?
(c) Should the bank consider blocking the credit card if a customer does not pay his bill on time?

Exercise 6.8 There are epidemics which affect animals such as cows, pigs, and others. Suppose 200 cows are tested to see whether they are infected with a virus or not. Let event A describe whether a cow has been transported by a truck recently or not and let B denote the event that a cow has been tested positive with a virus. The data are summarized in the following table:

	B	$\bar{B}$	Total
A	40	60	100
$\bar{A}$	20	80	100
Total	60	140	200

(a) What is the probability that a cow is infected and has been transported by a truck recently?
(b) What is the probability of having an infected cow given that it has been transported by the truck?
(c) Determine and interpret $P(B)$.

Exercise 6.9 A football practice target is a portable wall with two holes (which are the target) in it for training shots. Suppose there are two players A and B. The probabilities of hitting the target by A and B are 0.4 and 0.5, respectively.

(a) What is the probability that at least one of the players succeeds with his shot?
(b) What is the probability that exactly one of the players hits the target?
(c) What is the probability that only B scores?

→ Solutions to all exercises in this chapter can be found on p. 361

*Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

Random Variables

7

In the first part of the book we highlighted how to *describe* data. Now, we discuss the concepts required to draw statistical conclusions from a sample of data about a population of interest. For example, suppose we know the starting salary of a sample of 100 students graduating in law. We can use this knowledge to draw conclusions about the expected salary for the population of all students graduating in law. Similarly, if a newly developed drug is given to a sample of selected tuberculosis patients, then some patients may show improvement and some patients may not, but we are interested in the consequences for the entire population of patients. In the remainder of this chapter, we describe the theoretical concepts required for making such conclusions. They form the basis for statistical tests and inference which are introduced in Chaps. 9–11.

7.1 Random Variables

Random variables help us to view the collected data as an outcome of a random experiment. Consider the simple experiment of tossing a coin. If a coin is tossed, then one can observe either "head" (H) or "tail" (T). The occurrence of "head" or "tail" is random, and the exact outcome will only be known after the coin is tossed. We can toss the coin many times and obtain a sequence of outputs. For example, if a coin is tossed seven times, then one of the outcomes may be H, H, T, H, T, T, T. This outcome is the consequence of a random experiment, and it may be helpful if we can distill the sequence of outcomes in meaningful numbers. One option is to summarize them by a variable X, which takes the values $x_1 = 1$ (denoting head) and $x_2 = 0$ (denoting tail). We have learnt from Chap. 6 that this can be described in the framework of a random experiment where $\Omega = \{\omega_1, \omega_2\}$ with the events $A_1 = \{\omega_1\} = 1 =$ head and $A_2 = \{\omega_2\} = 0 =$ tail. The random variable X is

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_7

Table 7.1 Examples of random variables

X	Event	Realizations of X
Roll of a die	A_i: number i $(i = 1, 2, \ldots, 6)$	$x = i$
Lifetime of TV	A_i: survival time is i months $(i = 1, 2, \ldots)$	$x = i$
Roulette	A_1: red	$x_1 = 1$
	A_2: black	$x_2 = 2$
	A_3: green (zero)	$x_3 = 0$

now mapped to real numbers, and therefore, it describes the possible outcome of *any* coin toss experiment. The observed outcomes H, H, T, H, T, T, T relate to a specific sample, a unique *realization* of this experiment. We can write $X(\omega_1) = 1$ and $X(\omega_2) = 0$ with $\omega_1, \omega_2 \in \Omega$ and $1, 0 \in \mathcal{R}$ where $\mathcal{R}$ is the set of real numbers. We know that in any coin tossing experiment, the probability of head being observed is $P(X(\omega_1) = 1) = 0.5$ and of tail being observed is $P(X(\omega_2) = 0) = 0.5$. We may therefore view X as a random variable which collects the possible outcomes of a random experiment and captures the uncertainty associated with them.

Definition 7.1.1 Let Ω represent the sample space of a random experiment, and let $\mathcal{R}$ be the set of real numbers. A random variable is a function X which assigns to each element $\omega \in \Omega$ one and only one number $X(\omega) = x, x \in \mathcal{R}$, i.e.

$$X : \Omega \to \mathcal{R}. \tag{7.1}$$

Example 7.1.1 The features of a die roll experiment, a roulette game, or the lifetime of a TV can all be described by a random variable, see Table 7.1. The events involve randomness, and if we have knowledge about the random process, we can assign probabilities $P(X = x_i)$ to each event, e.g. when rolling a die, the probability of getting a "1" is $P(X = 1) = 1/6$ and the probability of getting a "2" is $P(X = 2) = 1/6$.

Note that it is a convention to denote random variables by capital letters (e.g. X) and their values by small letters (e.g. x). It is evident from the coin tossing experiment that we need to know $P(X = x)$ to describe the respective random variable. We assume in this chapter that we have this knowledge. However, Chaps. 9–11 show how a sample of data can be used to estimate unknown probabilities and other quantities given a prespecified uncertainty level. More generally, we can say that it is mandatory to know $P(X \in A)$ for all possible A which are subsets of $\mathcal{R}$. If we choose $A = (-\infty, x], x \in \mathcal{R}$, we have

$$P(X \in A) = P(X \in (-\infty, x]) = P(-\infty < X \leq x) = P(X \leq x).$$

This consideration gives rise to the definition of the cumulative distribution function. Recall that we developed the concept of the empirical cumulative distribution function (ECDF) in Chap. 2, Sect. 2.2, but the definition there was empirical. Now, we develop it theoretically.

7.2 Cumulative Distribution Function (CDF)

Definition 7.2.1 The **cumulative distribution function (CDF)** of a random variable X is defined as

$$F(x) = P(X \leq x). \tag{7.2}$$

As in Chap. 2, we can see that the CDF is useful in obtaining the probabilities related to the occurrence of random events. Note that the empirical cumulative distribution function (ECDF, Sect. 2.2) and the cumulative distribution function are closely related and therefore have a similar definition and similar calculation rules. However, in Chap. 2, we work with the cumulative distribution of *observed* values in a particular sample whereas in this chapter, we deal with random variables modelling the distribution of a general population.

The definition 7.2 implies the following properties of the cumulative distribution function:

- $F(x)$ is a monotonically non-decreasing function (if $x_1 \leq x_2$, it follows that $F(x_1) \leq F(x_2)$),
- $\lim_{x \to -\infty} F(x) = 0$ (the lower limit of F is 0),
- $\lim_{x \to +\infty} F(x) = 1$ (the upper limit of F is 1),
- $F(x)$ is continuous from the right, and
- $0 \leq F(x) \leq 1$ for all $x \in \mathcal{R}$.

Another notation for $F(x) = P(X \leq x)$ is $F_X(x)$, but we use $F(x)$.

7.2.1 CDF of Continuous Random Variables

Before giving some examples about the meaning and interpretation of the CDF, we first need to consider some definitions and theorems.

Definition 7.2.2 A random variable X is said to be **continuous** if there is a function $f(x)$ such that for all $x \in \mathcal{R}$

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \tag{7.3}$$

holds. $F(x)$ is the cumulative distribution function (CDF) of X, and $f(x)$ is the probability density function (PDF) of x and $\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$ for all x that are continuity points of f.

Theorem 7.2.1 *For a function $f(x)$ to be a **probability density function (PDF)** of X, it needs to satisfy the following conditions:*

(1) $f(x) \geq 0$ *for all* $x \in \mathcal{R}$,
(2) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

Theorem 7.2.2 *Let X be a random variable with CDF $F(x)$. If $x_1 < x_2$, where x_1 and x_2 are known constants, $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$.*

Theorem 7.2.3 *The probability of a continuous random variable taking a particular value x_0 is zero:*

$$P(X = x_0) = 0. \tag{7.4}$$

The proof is provided in Appendix C.2.

Example 7.2.1 Consider the continuous random variable "waiting time for the train". Suppose that a train arrives every 20 min. Therefore, the waiting time of a particular person is random and can be any time contained in the interval [0, 20]. We can start describing the required probability density function as

$$f(x) = \begin{cases} k & \text{for} \quad 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

where k is an unknown constant. Now, using condition (2) of Theorem 7.2.1, we have

$$1 = \int_0^{20} f(x)\mathrm{d}x = [kx]_0^{20} = 20k$$

which needs to be fulfilled. This yields $k = 1/20$ which is always greater than 0, and therefore, condition (1) of Theorem 7.2.1 is also fulfilled. It follows that

$$f(x) = \begin{cases} \frac{1}{20} & \text{for} \quad 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

is the probability density function describing the waiting time for the train. We can now use Definition 7.2.2 to determine the cumulative distribution function:

$$F(x) = \int_0^x f(t)\mathrm{d}t = \int_0^x \frac{1}{20}\mathrm{d}t = \frac{1}{20}[t]_0^x = \frac{1}{20}x.$$

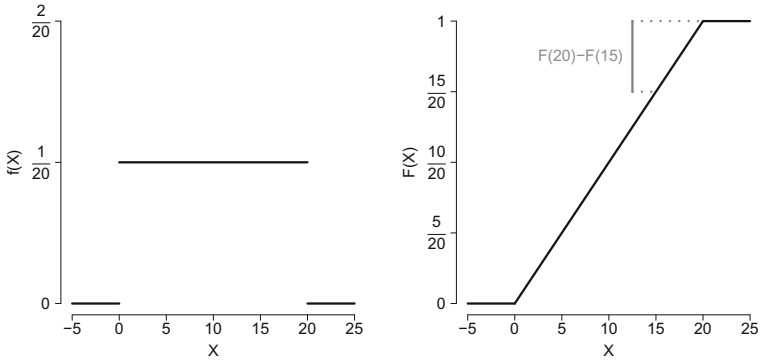
Suppose we are interested in calculating the probability of a waiting time between 15 and 20 min. This can be calculated using Theorem 7.2.2:

$$P(15 \leq X \leq 20) = F(20) - F(15) = \frac{20}{20} - \frac{15}{20} = 0.25.$$

We can obtain this probability from the graph of the CDF as well, see Fig. 7.1 where both the PDF and CDF of this example are illustrated.

Defining a function, for example the CDF, is simple in *R*: One can use the `function` command followed by specifying the variables the function evaluates in round brackets (e.g. x) and the function itself in braces (e.g. $x/20$). Functions can be plotted using the `curve` command:

```
cdf <- function(x){1/20*x}
curve(cdf,from=0,to=20)
```



(a) PDF (b) CDF

Fig. 7.1 Probability density function (PDF) and cumulative distribution function (CDF) for waiting time in Example 7.2.1

Alternatively, the `plot` command can be used to plot vectors against each other; for example, after defining a function, we can define a sequence (`x<-seq(0,20,0.01)`), evaluate this sequence via the specified function (`cdf(x)`), and plot them against each other and connect the points from the sequence with a line (`plot(x,cdf(x),type='l')`).

This example illustrates how the cumulative distribution function can be used to obtain probabilities of interest. Most importantly, if we want to calculate the probability that the random variable X takes values in the interval $[x_1, x_2]$, we simply have to look at the difference of the respective CDF values at x_1 and x_2. Figure 7.2a highlights that the interval probability corresponds to the difference of the CDF values on the y-axis.

We can also use the probability density function to visualize $P(x_1 \leq X \leq x_2)$. We know from Theorem 7.2.1 that $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$, and therefore, the area under the PDF equals 1. Thus, we can interpret interval probabilities as the area under the PDF between x_1 and x_2. This is presented in Fig. 7.2b.

7.2.2 CDF of Discrete Random Variables

Definition 7.2.3 A random variable X is defined to be **discrete** if its probability space is either finite or countable, i.e. if it takes only a finite or countable number of values. Note that a set V is said to be **countable**, if its elements can be listed, i.e. there is a one-to-one correspondence between V and the positive integers.

Example 7.2.2 Consider the example of tossing of a coin where each trial results in either a head (H) or a tail (T), each occurring with the same probability

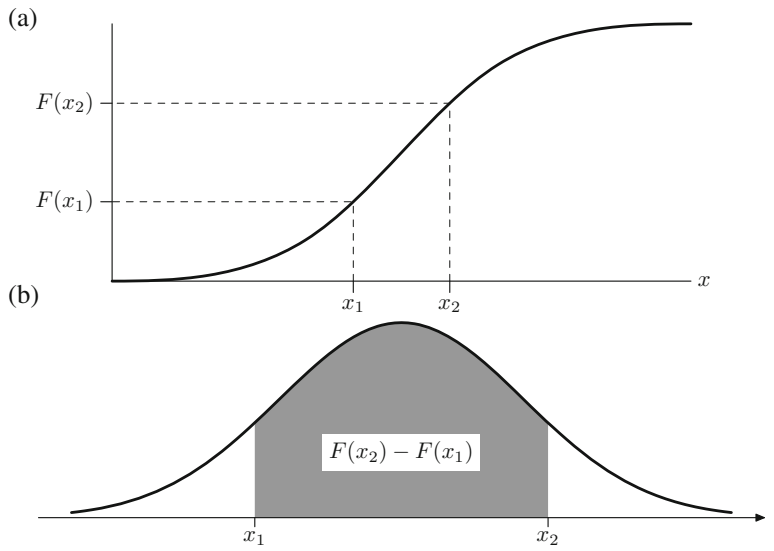


Fig. 7.2 Graphical representation of the probability $P(x_1 \leq X \leq x_2)$ **a** via the CDF and **b** via the PDF*

0.5. When the coin is tossed multiple times, we may observe sequences such as H, T, H, H, T, H, H, T, and $T, \ldots$. The sample space is $\Omega = \{H, T\}$. Let the random variable X denote the number of trials required to get the third head, then $X = 4$ for the given sequence above. Clearly, the space of X is the set $(3, 4, 5, \ldots)$. We can see that X is a discrete random variable because its space is finite and can be counted. We can also assign certain probabilities to each of these values, e.g. $P(X = 3) = p_1$ and $P(X = 4) = p_2$.

Definition 7.2.4 Let X be a discrete random variable which takes k different values. The **probability mass function** (PMF) of X is given by

$$f(X) = P(X = x_i) = p_i \quad \text{for each} \quad i = 1, 2, \ldots, k. \tag{7.5}$$

It is required that the probabilities p_i satisfy the following conditions:

(1) $0 \leq p_i \leq 1$,
(2) $\sum_{i=1}^{k} p_i = 1$.

Definition 7.2.5 Given (7.5), we can write the CDF of a discrete random variable as

$$F(x) = \sum_{i=1}^{k} I_{\{x_i \le x\}} p_i, \tag{7.6}$$

where I is an indicator function defined as

$$I_{\{x_i \le x\}} = \begin{cases} 1 & \text{if} \quad x_i \le x \\ 0 & \text{otherwise.} \end{cases}$$

The CDF of a discrete variable is always a **step function**.

Working with the CDF for Discrete Random variables

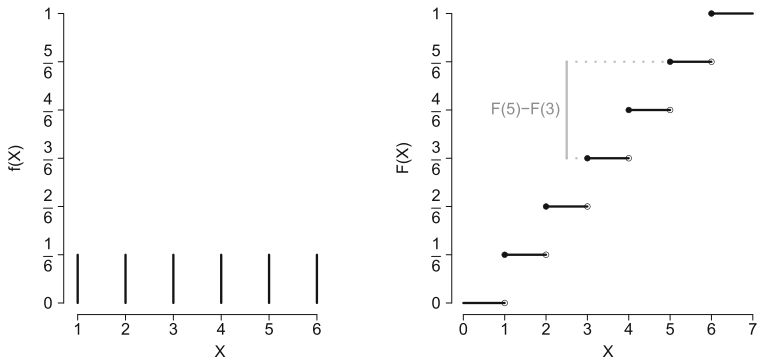
We can easily calculate various types of probabilities for discrete random variables using the CDF. Let a and b be some known constants, then

$$P(X \le a) = F(a), \tag{7.7}$$
$$P(X < a) = P(X \le a) - P(X = a) = F(a) - P(X = a), \tag{7.8}$$
$$P(X > a) = 1 - P(X \le a) = 1 - F(a), \tag{7.9}$$
$$P(X \ge a) = 1 - P(X < a) = 1 - F(a) + P(X = a), \tag{7.10}$$
$$\begin{aligned} P(a \le X \le b) &= P(X \le b) - P(X < a) \\ &= F(b) - F(a) + P(X = a), \end{aligned} \tag{7.11}$$
$$P(a < X \le b) = F(b) - F(a), \tag{7.12}$$
$$P(a < X < b) = F(b) - F(a) - P(X = b), \tag{7.13}$$
$$P(a \le X < b) = F(b) - F(a) - P(X = b) + P(X = a). \tag{7.14}$$

Remark 7.2.1 The Eqs. (7.7)–(7.14) can also be used for continuous variables, but in this case, $P(X = a) = P(X = b) = 0$ (see Theorem 7.2.3), and therefore, Eqs. (7.7)–(7.14) can be modified accordingly.

Example 7.2.3 Consider the experiment of rolling a die. There are six possible outcomes. If we define the random variable X as the number of dots observed on the upper surface of the die, then the six possible outcomes can be described as $x_1 = 1, x_2 = 2, \ldots, x_6 = 6$. The respective probabilities are $P(X = x_i) = 1/6$; $i = 1, 2, \ldots, 6$. The PMF and CDF are therefore defined as follows:

$$f(x) = \begin{cases} 1/6 & \text{if } x = 1 \\ 1/6 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 1/6 & \text{if } x = 4 \\ 1/6 & \text{if } x = 5 \\ 1/6 & \text{if } x = 6 \\ 0 & \text{elsewhere.} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{if } -\infty < x < 1 \\ 1/6 & \text{if } 1 \le x < 2 \\ 2/6 & \text{if } 2 \le x < 3 \\ 3/6 & \text{if } 3 \le x < 4 \\ 4/6 & \text{if } 4 \le x < 5 \\ 5/6 & \text{if } 5 \le x < 6 \\ 1 & \text{if } 6 \le x < \infty. \end{cases}$$



(a) PDF (b) CDF

Fig. 7.3 Probability density function and cumulative distribution function for rolling a die in Example 7.2.3. "•" relates to an included value and "∘" to an excluded value

Both the CDF and the PDF are displayed in Fig. 7.3.

We can use the CDF to calculate any desired probability, e.g. $P(X \leq 5) = F(5) = 5/6$. This is shown in Fig. 7.3b where for $X = 5$, we obtain $F(5) = 5/6$ when evaluating on the y-axis. Similarly, $P(3 < X \leq 5) = F(5) - F(3) = (5/6) - (3/6) = 2/6$ can be interpreted as the difference of $F(5)$ and $F(3)$ on the y-axis.

7.3 Expectation and Variance of a Random Variable

We have seen that both the probability density function (or probability mass function) and the cumulative distribution function are helpful in characterizing the features of a random variable. Some other features of random variables are characterized by the concepts of *expectation* and *variance*.

7.3.1 Expectation

Definition 7.3.1 The expectation of a continuous random variable X, having the probability density function $f(x)$ with $\int |x| f(x) \mathrm{d}x < \infty$, is defined as

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x. \tag{7.15}$$

For a discrete random variable X, which takes the values $x_1, x_2, \ldots$ with respective probabilities $p_2, p_2, \ldots$, the **expectation** of X is defined as

$$\mathrm{E}(X) = \sum_{i=1}^{k} x_i p_i = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_k P(X = x_k). \quad (7.16)$$

The *expectation* of X, i.e. $\mathrm{E}(X)$, is usually denoted by $\mu = \mathrm{E}(X)$ and relates to the arithmetic mean of the distribution of the population. It reflects the central tendency of the population.

Example 7.3.1 Consider again Example 7.2.1 where the waiting time for a train was described by the following probability density function:

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \le x \le 20 \\ 0 & \text{otherwise.} \end{cases}$$

We can calculate the expectation as follows:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_{-\infty}^{0} x f(x)\,\mathrm{d}x + \int_{0}^{20} x f(x)\,\mathrm{d}x + \int_{20}^{\infty} x f(x)\,\mathrm{d}x \\ &= 0 + \int_{0}^{20} \frac{1}{20} x\,\mathrm{d}x + 0 = \left[\frac{1}{40} x^2\right]_0^{20} = \frac{400}{40} - 0 = 10. \end{aligned}$$

The "average" waiting time for the train is therefore 10 min. This means that if a person has to wait for the train every day, then the person will experience waiting times varying randomly between 0 and 20 min and, on average, has to wait for 10 min.

Example 7.3.2 Consider again the die roll experiment from Example 7.2.3. The probabilities for the occurrence of any $x_i, i = 1, 2, \ldots, 6$, are $P(X = x_i) = 1/6$. The expectation can thus be calculated as

$$\begin{aligned} \mathrm{E}(X) &= \sum_{i=1}^{6} x_i p_i \\ &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + 4 \cdot P(X = 4) \\ &\quad + 5 \cdot P(X = 5) + 6 \cdot P(X = 6) \\ &= (1 + 2 + 3 + 4 + 5 + 6)\frac{1}{6} = \frac{21}{6} = 3.5. \end{aligned}$$

7.3.2 Variance

The *variance* describes the variability of a random variable. It gives an idea about the concentration or dispersion of values around the arithmetic mean of the distribution.

Definition 7.3.2 The **variance** of a random variable X is defined as

$$\mathrm{Var}(X) = \mathrm{E}[X - \mathrm{E}(X)]^2. \tag{7.17}$$

The variance of a continuous random variable X is

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathrm{E}(X))^2 f(x)\mathrm{d}x \tag{7.18}$$

where $\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x)\mathrm{d}x$. Similarly, the variance of a discrete random variable X is

$$\mathrm{Var}(X) = \sum_{i=1} (x_i - \mathrm{E}(X))^2 p_i \tag{7.19}$$

where $\mathrm{E}(X) = \sum_i x_i p_i$. The variance is usually denoted by $\sigma^2 = \mathrm{Var}(X)$.

Definition 7.3.3 The positive square root of the variance is called the **standard deviation**.

Example 7.3.3 Recall Examples 7.2.1, and 7.3.1. We can calculate the variance of the waiting time for a train using the probability density function

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } \ 0 \le x \le 20 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathrm{E}(X) = 10$ (already calculated in Example 7.3.1). Using (7.18), we obtain:

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathrm{E}(x))^2 f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} (x-10)^2 f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{0} (x-10)^2 f(x)\,\mathrm{d}x + \int_{0}^{20} (x-10)^2 f(x)\,\mathrm{d}x + \int_{20}^{\infty} (x-10)^2 f(x)\,\mathrm{d}x \\ &= 0 + \int_0^{20} (x-10)^2 \cdot \frac{1}{20}\,\mathrm{d}x + 0 = \int_0^{20} \frac{1}{20}(x^2 - 20x + 100)\,\mathrm{d}x \\ &= \left[\frac{1}{20}\left(\frac{1}{3}x^3 - 10x^2 + 100x\right)\right]_0^{20} = 33\frac{1}{3}. \end{aligned}$$

The standard deviation is $\sqrt{33\frac{1}{3}\,\mathrm{min}^2} \approx 5.77\,\mathrm{min}$.

Recall that in Chap. 3, we introduced the sample variance and the sample standard deviation. We already know that the standard deviation has the same unit of measurement as the variable, whereas the unit of the variance is the square of the measurement unit. The standard deviation measures how the values of a random variable are dispersed around the population 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 data values around the mean, and the observed values may be far away from the mean. These considerations are helpful in making connections between random variables and samples of data, see Chap. 9 for the construction of confidence intervals.

Example 7.3.4 Recall Example 7.3.2 where we calculated the expectation of a die roll experiment as $\mathrm{E}(X) = 3.5$. With $x_i \in \{1, 2, 3, 4, 5, 6\}$ and $p_i = 1/6$ for all $i = 1, 2, 3, 4, 5, 6$, the variance for this example corresponds to

$$\mathrm{Var}(X) = \sum_{i=1}(x_i - \mathrm{E}(X))^2 p_i = (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + (3 - 3.5)^2 \cdot \frac{1}{6}$$
$$+(4 - 3.5)^2 \cdot \frac{1}{6} + (5 - 3.5)^2 \cdot \frac{1}{6} + (6 - 3.5)^2 \cdot \frac{1}{6} \approx 2.92.$$

Theorem 7.3.1 *The variance of a random variable X can be expressed as*

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2. \tag{7.20}$$

The proof is given in Appendix C.2.

Example 7.3.5 In Examples 7.2.1, 7.3.1, and 7.3.3, we evaluated the waiting time for a train using the PDF

$$f(X) = \begin{cases} \frac{1}{20} & \text{for } 0 < X \leq 20 \\ 0 & \text{otherwise.} \end{cases}$$

We calculated the expectation and variance in Eqs. (7.15) and (7.17) as 10 min and $33\frac{1}{3}$ min^2, respectively. Theorem 7.3.1 tells us that we can calculate the variance in a different way as follows:

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \int_0^{20} \frac{1}{20} x^2\,\mathrm{d}x$$
$$= \left[\frac{1}{60}x^3\right]_0^{20} = 133\frac{1}{3}$$
$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 133\frac{1}{3} - 10^2 = 33\frac{1}{3}.$$

This yields the same result as Eq. (7.18) but is much quicker.

7.3.3 Quantiles of a Distribution

We introduced the concept of quantiles in Chap. 3, Sect. 3.1.2. Now, we define quantiles in terms of the distribution function.

Definition 7.3.4 The value x_p for which the cumulative distribution function is

$$F(x_p) = p \quad (0 < p < 1) \tag{7.21}$$

is called the **p-quantile**.

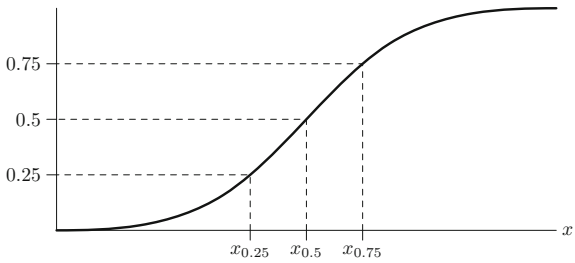


Fig. 7.4 First quartile, median, and third quartile*

It follows from Definition 7.3.4 that x_p is the value which divides the cumulative distribution function into two parts: the probability of observing a value left of x_p is p, whereas the probability of observing a value right of x_p is $1 - p$. For example, the 0.25-quantile $x_{0.25}$ describes the x-value for which the probability of observing $x_{0.25}$ or any smaller value is 0.25. Figure 7.4 shows the 0.25-quantile (first quartile), the 0.5-quantile (median), and the 0.75-quantile (third quartile) in a cumulative distribution function.

Example 7.3.6 Recall Examples 7.2.1, 7.3.1, 7.3.5 and Fig. 7.1b where we described the waiting time for a train by using the following CDF:

$$F(x) = \frac{1}{20}x.$$

The first quartile $x_{0.25}$ is 5 because $F(5) = 5/20 = 0.25$. This means that the probability of waiting for the train for 5 min or less is 25 % and of waiting for longer than 5 min is 75 %.

For continuous variables, there is a unique value which describes the p-quantile. However, for discrete variables, this may not necessarily be true. In this case, the p-quantile is chosen such that

$$\begin{aligned} F(x_p) &\geq p, \\ F(x) &< p \quad \text{for} \quad x < x_p \end{aligned}$$

holds.

Example 7.3.7 The cumulative distribution function for rolling a die is described in Example 7.2.3 and Fig. 7.3b. The first quartile $x_{0.25}$ is 2 because $F(2) = 2/6 > 0.25$ and $F(x) < 0.25$ for $x < 2$.

7.3.4 Standardization

Standardization transforms a random variable in such a way that it has an expectation of zero and a variance of one. More details on the need for standardization are discussed in Chap. 10.

Definition 7.3.5 A random variable Y is called **standardized** when

$$\mathrm{E}(Y) = 0 \quad \text{and} \quad \mathrm{Var}(Y) = 1.$$

Theorem 7.3.2 *Suppose a random variable X has mean $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Then, it can be standardized as follows:*

$$Y = \frac{X - \mu}{\sigma} = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}. \tag{7.22}$$

Example 7.3.8 In Examples 7.2.1, 7.3.1, and 7.3.5, we considered the waiting time X for a train. The random variable X can take values between 0 and 20 min, and we calculated $\mathrm{E}(X) = 10$ and $\mathrm{Var}(X) = 33\frac{1}{3}$. The standardized variable of X is

$$Y = \frac{X - \mu}{\sigma} = \frac{X - 10}{\sqrt{33\frac{1}{3}}}.$$

One can show that $\mathrm{E}(Y) = 0$ and $\mathrm{Var}(Y) = 1$, see also Exercise 7.10 for more details.

7.4 Tschebyschev's Inequality

If we do not know the distribution of a random variable X, we can still make statements about the probability that X takes values in a certain interval (which has to be symmetric around the expectation μ) if the mean μ and the variance σ^2 of X are known.

Theorem 7.4.1 (Tschebyschev's inequality) *Let X be a random variable with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. It holds that*

$$P(|X - \mu| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}. \tag{7.23}$$

This is equivalent to

$$P(|X - \mu| < c) \geq 1 - \frac{\mathrm{Var}(X)}{c^2}. \tag{7.24}$$

The proof is given in Appendix C.2.

Example 7.4.1 In Examples 7.2.1, 7.3.1, and 7.3.5, we have worked with a random variable which describes the waiting time for a train. We determined $\mathrm{E}(X) = 10$ and $\mathrm{Var}(X) = 33\frac{1}{3}$. We can calculate the probability of waiting between $10 - 7 = 3$ and $10 + 7 = 17$ min:

$$P(|X - \mu| < c) \geq 1 - \frac{\mathrm{Var}(X)}{c^2}$$

$$P(|X - 10| < 7) \geq 1 - \frac{33\frac{1}{3}}{7^2} \approx 0.32.$$

The probability is therefore at least 0.32. However, if we apply our distributional knowledge that $F(x) = \frac{1}{20}x$ (for $0 \leq X \leq 20$), then we obtain a much more precise result which is

$$P(3 < X < 17) = F(17) - F(3) = \frac{17}{20} - \frac{3}{20} = 0.7.$$

We can clearly see that Tschebyschev's inequality gives us the correct answer, that is $P(3 < X < 17)$ is greater 0.32. Nevertheless, the approximation to the exact probability, 0.7, is rather poor. One needs to keep in mind that only the lack of distributional knowledge makes the inequality useful.

7.5 Bivariate Random Variables

There are many situations in which we are interested in analysing more than one variable, say two variables. When we have more than one variable, then not only their individual distributions but also their joint distribution can be of interest. For example, we know that driving a car after drinking alcohol is not necessarily safe. If we consider two variables, the blood alcohol content X and number of car accidents Y, then we may be interested in the probability of having a high blood alcohol content *and* a car accident at the same time. If we analyse (X, Y) jointly, then we are interested in their joint **bivariate** distribution $f_{XY}(x, y)$. This distribution can either be discrete or continuous.

Discrete Bivariate Random Variables. Suppose we have two categorical variables X and Y which can take the values $x_1, x_2, \ldots, x_I$ and $y_1, y_2, \ldots, y_J$, respectively. Their **joint probability distribution function** is characterized by

$$P(X = x_i, Y = y_j) = p_{ij} \quad (i = 1, 2, \ldots, I\,;\, j = 1, 2, \ldots, J)$$

with $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$. This means that the probability of observing x_i *and* y_j together is p_{ij}. We can summarize this information in a contingency table as follows:

		Y				
		1	2	...	J	Total
X	1	p_{11}	p_{12}	...	p_{1J}	p_{1+}
	2	p_{21}	p_{22}	...	p_{2J}	p_{2+}
	⋮	⋮				⋮
	I	p_{I1}	p_{I2}	...	p_{IJ}	p_{I+}
	Total	p_{+1}	p_{+2}	...	p_{+J}	1

Each cell contains a "piece" of the joint distribution. The entries $p_{+1}, p_{+2}, \ldots, p_{+J}$ in the bottom row of the table summarize the **marginal distribution** of Y, which is the distribution of Y without giving reference to X. The entries $p_{1+}, p_{+2}, \ldots, p_{I+}$

in the last column summarize the marginal distribution of X. The marginal distributions can therefore be expressed as

$$P(X = x_i) = \sum_{j=1}^{J} p_{ij} = p_{i+} \quad i = 1, 2, \ldots, I,$$

$$P(Y = y_j) = \sum_{i=1}^{I} p_{ij} = p_{+j} \quad j = 1, 2, \ldots, J\,.$$

The **conditional distributions** of X given $Y = y_j$ and Y given $X = x_j$ are given as follows:

$$P(X = x_i|Y = y_j) = p_{i|j} = \frac{p_{ij}}{p_{+j}} \quad i = 1, 2, \ldots, I,$$

$$P(Y = y_j|X = x_i) = p_{j|i} = \frac{p_{ij}}{p_{i+}} \quad j = 1, 2, \ldots, J.$$

They summarize the distribution of X for a given value of y_j (or the distribution of Y for a given value of x_i) and play a crucial role in the construction of regression models such as the linear regression model introduced in Chap. 11. Please also recall the definitions of Sect. 4.1 where we introduced conditional and marginal distributions for data samples rather than random variables.

Example 7.5.1 Suppose we have a contingency table on smoking behaviour X ($1 =$ never smoking, $2 =$ smoking sometimes, and $3 =$ smoking regularly) and education level Y ($1 =$ primary education, $2 =$ Secondary education, and $3 =$ tertiary education):

		Y			
		1	2	3	Total
X	1	0.10	0.20	0.30	0.60
	2	0.10	0.10	0.10	0.30
	3	0.08	0.01	0.01	0.10
	Total	0.28	0.31	0.41	1

The cell entries represent the joint distribution of smoking behaviour and education level. We can interpret each entry as the probability of observing $X = x_i$ *and* $Y = y_j$ simultaneously. For example, $p_{23} = P$ ("smoking sometimes and tertiary education") $= 0.10$. The marginal distribution of X is contained in the last column of the table and lists the probabilities of smoking (unconditional on education level), e.g. the probability of being a non-smoker in this population is 60 %. We can also interpret the conditional distributions: $P(X|Y = 3)$ represents the distribution of smoking behaviour among those who have tertiary education. If we are interested in the probability of smoking sometimes given tertiary education is completed, then we calculate $P(X = 2|Y = 3) = p_{2|3} = \frac{0.10}{0.41} = 0.24$.

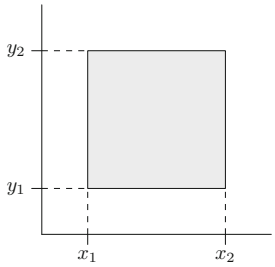


Fig. 7.5 Area covering all points of (X, Y) with $(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)^*$

Continuous Bivariate Random Variables.

Definition 7.5.1 A bivariate random variable (X, Y) is continuous if there is a function $f_{XY}(x, y)$ such that

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x, y)\, \mathrm{d}x\, \mathrm{d}y \tag{7.25}$$

holds.

The function $F_{XY}(x, y)$ is the **joint cumulative distribution function** of X and Y; the joint distribution function is denoted by $f_{XY}(x, y)$, and $f_{XY}(x, y)$ has to fulfil the usual conditions of a density function. Necessary and sufficient conditions that a function $F_{XY}(x, y)$ is a bivariate cumulative distribution function are as follows:

$$\lim_{x \to -\infty} F_{XY}(x, y) = 0 \qquad \lim_{y \to -\infty} F_{XY}(x, y) = 0$$
$$\lim_{x \to \infty} F_{XY}(x, y) = 1 \qquad \lim_{y \to \infty} F_{XY}(x, y) = 1$$

and $F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0$ for all $x_1 < x_2, y_1 < y_2$.

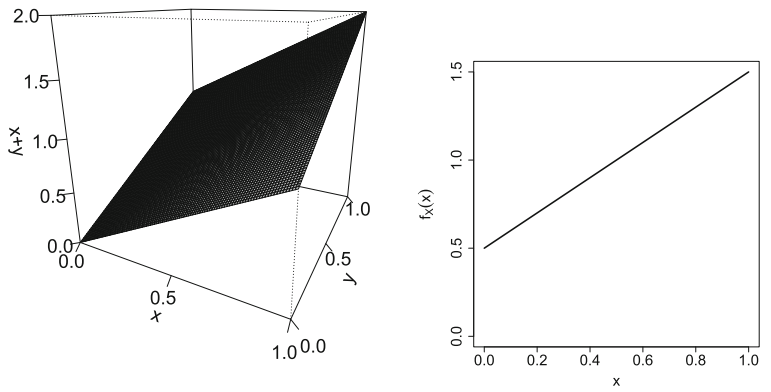
The last condition is sometimes referred to as the *rectangle inequality*. As in the univariate case, we can use the cumulative distribution function to calculate interval probabilities; similarly, we look at the rectangular area defined by (x_1, y_1), (x_1, y_2), (x_2, y_1), and (x_2, y_2) in the bivariate case (instead of an interval $[a, b]$), see Fig. 7.5.

We can calculate the desired probabilities as follows:

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

The **marginal distributions** of X and Y are

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}x,$$



(a) Joint distribution $f_{XY}(x, y)$ (b) Marginal distribution $f_X(x)$

Fig. 7.6 Joint and marginal distribution for Example 7.5.2

respectively. Similar to the discrete case, $f_X(x)$ and $f_Y(y)$ also describe the distribution of X unconditional on Y and the distribution of Y unconditional on X. The **cumulative marginal distributions** are

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t, \quad F_Y(y) = \int_{-\infty}^{y} f_Y(t)\mathrm{d}t.$$

The **conditional distributions** can be obtained by the ratio of the joint and marginal distributions:

$$f_{X|Y}(x, y) = \frac{f(x, y)}{f(y)}, \quad f_{Y|X}(x, y) = \frac{f(x, y)}{f(x)}.$$

Example 7.5.2 Consider the function

$$f_{XY}(x, y) = \begin{cases} x + y & \text{for } 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Suppose X and Y represent the concentrations of two drugs in the human body. Then, $f_{XY}(x, y)$ may represent the sum of two drug concentrations in the human body. Since there are infinite possible realizations of both X and Y, we represent their joint distribution in a figure rather than a table, see Fig. 7.6a.

The marginal distributions for X and Y can be calculated as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}y = \int_0^1 (x + y)\, dy = \left[xy + \frac{1}{2}y^2\right]_0^1 = x + \frac{1}{2},$$

$$f_Y(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}x = \int_0^1 (x + y)\, \mathrm{d}x = \left[\frac{1}{2}x^2 + xy\right]_0^1 = y + \frac{1}{2}.$$

Figure 7.6b depicts the marginal distribution for X. The slope of the marginal distribution is essentially the slope of the surface of the joint distribution shown in Fig. 7.6a. It is easy to see in this simple example that the marginal distribution of

X is nothing but a cut in the surface of the joint distribution. Note that the conditional distributions $f_{X|Y}(x, y)$ and $f_{Y|X}(x, y)$ can be easily calculated; for example, $f_{X|Y}(x, y) = f(x, y)/f(y) = (x + y)/(y + 0.5)$.

Stochastic Independence.

Definition 7.5.2 Two continuous random variables X and Y are said to be **stochastically independent** if

$$f_{XY}(x, y) = f_X(x) f_Y(y). \tag{7.26}$$

For discrete variables, this is equivalent to

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j) \tag{7.27}$$

being valid for all (i, j).

Example 7.5.3 In Example 7.5.2, we considered the function

$$f_{XY}(x, y) = \begin{cases} x + y & \text{for } 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

with the marginal distributions of X and Y as $f_X = x + 0.5$ and $f_Y = y + 0.5$, respectively. Since $f_X \cdot f_Y = (x + \frac{1}{2})(y + \frac{1}{2}) \neq f_{XY}$, it follows that X and Y are not independent. The interpretation is that the concentrations of the two drugs are not independent.

7.6 Calculation Rules for Expectation and Variance

Calculation Rules for the Expectation. For any constant values a and b, and any random variables X and Y, the following rules hold:

$$E(a) = a, \tag{7.28}$$

$$E(bX) = b\mathrm{E}(X), \tag{7.29}$$

$$E(a + bX) = a + b\mathrm{E}(X), \tag{7.30}$$

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) \ (\textit{additivity}). \tag{7.31}$$

The proof of rule (7.30) is given in Appendix C.2.

Example 7.6.1 Consider again Example 7.2.3 where we illustrated how the outcome of a die roll experiment can be captured by a random variable. There were 6 events, and X could take the values $x_1 = 1, x_2 = 2, \ldots, x_6 = 6$. The probability of the occurrence of any number was $P(X = x_i) = 1/6$, and the expectation was calculated as 3.5. Consider two different situations:

(i) Suppose the die takes the value 10, 20, 30, 40, 50, and 60 instead of the values 1, 2, 3, 4, 5, and 6. The random variable $Y = 10X$ describes this suitably, and its expectation is

$$\mathrm{E}(Y) = \mathrm{E}(10X) = 10\mathrm{E}(X) = 10 \cdot 3.5 = 35$$

which follows from (7.29).

(ii) If we are rolling two dices X_1 and X_2, then the expectation for the sum of the two outcomes is

$$\mathrm{E}(X) = \mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 3.5 + 3.5 = 7$$

due to (7.31).

Calculation Rules for the Variance. Let a and b be any known constants and X be a random variable (discrete or continuous). Then, we have the following rules:

$$\mathrm{Var}(a) = 0, \tag{7.32}$$

$$\mathrm{Var}(bX) = b^2\,\mathrm{Var}(X), \tag{7.33}$$

$$\mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X). \tag{7.34}$$

The proof of rule (7.34) is given in Appendix C.2.

Example 7.6.2 In Examples 7.2.1, 7.3.1, 7.3.3, and 7.3.5, we evaluated a random variable describing the waiting time for a train. Now, suppose that a person first has to catch a bus to get to the train station. If this bus arrives only every 60 min, then the PDF of the random variable Y denoting the waiting time for the bus is

$$f(Y) = \begin{cases} \frac{1}{60} & \text{for} \quad 0 < x \leq 60 \\ 0 & \text{otherwise} . \end{cases}$$

We can use Eqs. (7.15) and (7.17) to determine both the expectation and variance of Y. However, the waiting time for the bus is governed by the relation $Y = 3X$ where X is the waiting time for the train. Therefore, we can calculate $\mathrm{E}(Y) = \mathrm{E}(3X) = 3\mathrm{E}(X) = 3 \cdot 10 = 30$ min by using rule (7.29) and the variance as $\mathrm{Var}(Y) = \mathrm{Var}(3X) = 3^2\,\mathrm{Var}(X) = 9 \cdot 33\frac{1}{3} = 300$ using rule (7.33). The total waiting time is the sum of the two waiting times.

7.6.1 Expectation and Variance of the Arithmetic Mean

Definition 7.6.1 We define the random variables $X_1, X_2, \ldots, X_n$ to be i.i.d. (independently identically distributed), if all X_i follow the same distribution and are stochastically independent of each other.

Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$. The arithmetic mean of these variables is given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which is again a random variable that follows a distribution with certain expectation and variance. A function of random variables is called a **statistic**. By using (7.29) and (7.31), we obtain

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \mu. \tag{7.35}$$

If we apply (7.34) and recall that the variables are independent of each other, we can also calculate the variance as

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{7.36}$$

Example 7.6.3 If we toss a coin, we obtain either head or tail, and therefore, $P(\text{“head”}) = P(\text{“tail”}) = \frac{1}{2}$. If we toss the coin n times, we have for each toss

$$X_i = \begin{cases} 0 & \text{for “tail”} \\ 1 & \text{for “head”} \end{cases}, \quad i = 1, \ldots, n.$$

It is straightforward to calculate the expectation and variance for each coin toss:

$$\begin{aligned} \mathrm{E}(X_i) &= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}, \\ \mathrm{Var}(X_i) &= (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

The arithmetic mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ describes the relative frequency of heads when the coin is tossed n times. We can now apply (7.35) and (7.36) to calculate

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} 1/2 = 1/2$$

and

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{4} = \frac{1}{4n}.$$

With this example, the interpretation of formulae (7.35) and (7.36) becomes clearer: if the probability of head is 0.5 for a single toss, then it is also 0.5 for the mean of all tosses. If we toss a coin many times, then the variance decreases when n increases. This means that a larger sample size yields a higher precision for the calculated arithmetic mean. This observation shows the basic conclusion of the next chapter: the higher the sample size, the more secure we are of our conclusions.

7.7 Covariance and Correlation

The variance measures the variability of a variable. Similarly, the covariance measures the covariation or association between X and Y.

7.7.1 Covariance

Definition 7.7.1 The **covariance** between X and Y is defined as

$$\varrho = \mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))]. \tag{7.37}$$

The covariance is positive if, on average, larger values of X correspond to larger values of Y; it is negative if, on average, greater values of X correspond to smaller values of Y.

The probability density function of any bivariate random variable (X, Y) is characterized by the expectation and variance of both X and Y,

$$\begin{aligned} \mathrm{E}(X) &= \mu_X, \quad \mathrm{Var}(X) = \sigma_X^2, \\ \mathrm{E}(Y) &= \mu_Y, \quad \mathrm{Var}(Y) = \sigma_Y^2, \end{aligned}$$

as well as their **covariance**. We can summarize these features by using the expectation vector

$$\mathrm{E}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mathrm{E}(X) \\ \mathrm{E}(Y) \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

and the **covariance matrix**

$$\mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(X, X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \mathrm{Cov}(Y, Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \varrho \\ \varrho & \sigma_Y^2 \end{pmatrix}.$$

Important properties of covariance are

(i) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$,
(ii) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$,
(iii) $\mathrm{Cov}(aX + b, cY + d) = ac\ \mathrm{Cov}(X, Y)$,
(iv) $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ where $\mathrm{E}(XY) = \int\int xyf(x, y)\mathrm{d}x\mathrm{d}y$ for continuous variables and $\mathrm{E}(XY) = \sum_i \sum_j x_i y_j p_{ij}$ for discrete variables,
(v) If X and Y are independent, it follows that $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) = \mu_X\mu_Y$, and therefore, $\mathrm{Cov}(X, Y) = \mu_X\mu_Y - \mu_X\mu_Y = 0$.

Theorem 7.7.1 (Additivity Theorem) *The variance of the sum (subtraction) of X and Y is given by*

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X, Y).$$

If X and Y are independent, it follows that $\mathrm{Cov}(X, Y) = 0$ and therefore $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. We omit the proof of this theorem.

Example 7.7.1 Recall Example 7.6.2 where we considered the waiting time Y for a bus to the train station and the waiting time X for the waiting time for a train. Suppose their joint bivariate probability density function can be written as

$$f_{XY}(x, y) = \begin{cases} \frac{1}{1200} & \text{for } 0 \le x \le 60, \quad 0 \le y \le 20 \\ 0 & \text{elsewhere.} \end{cases}$$

To calculate the covariance between X and Y, we need to calculate $\mathrm{E}(XY)$:

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\mathrm{d}x\,\mathrm{d}y = \int_0^{60}\int_0^{20} xy\frac{1}{1200}\mathrm{d}x\,\mathrm{d}y$$
$$= \int_0^{60}\left[\frac{x}{1200}\frac{y^2}{2}\right]_0^{20}\mathrm{d}y = \int_0^{60}\frac{400x}{2400}\mathrm{d}y = \left[\frac{1}{6}\frac{x^2}{2}\right]_0^{60} = \frac{3600}{12} = 300\,.$$

We know from Example 7.6.2 that $\mathrm{E}(X) = 10, \mathrm{E}(Y) = 30, \mathrm{Var}(X) = 33\frac{1}{3}$, and $\mathrm{Var}(Y) = 300$. The covariance is thus

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 300 - 30\cdot 10 = 0.$$

This makes sense as the waiting times for the train and the bus should be independent of each other. Using rule (7.31), we conclude that the total expected waiting time is

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) = 10 + 30 = 40\,\mathrm{min}.$$

The respective variance is

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y) = 33\frac{1}{3} + 300 - 2\cdot 0 = 333\frac{1}{3}$$

due to Theorem 7.7.1.

7.7.2 Correlation Coefficient

Definition 7.7.2 The **correlation coefficient** of X and Y is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}. \tag{7.38}$$

We already know from Chap. 4 that the correlation coefficient is a measure of the degree of linear relationship between X and Y. It can take values between -1 and 1, $-1 \le \rho(X,Y) \le 1$. However in Chap. 4, we considered the correlation of two samples, i.e. realizations of random variables; here, we describe the correlation coefficient of the population. If $\rho(X,Y) = 0$, then X and Y are said to be uncorrelated. If there is a perfect linear relationship between X and Y, then $\rho = 1$ for a positive relationship and $\rho = -1$ for a negative relationship, see Appendix C.2 for the proof.

Theorem 7.7.2 *If X and Y are independent, they are also uncorrelated. However, if they are uncorrelated then they are not necessarily independent.*

Example 7.7.2 In Example 7.6.2, we estimated the covariance between the waiting time for the bus and the waiting time for the train: $\mathrm{Cov}(X,Y) = 0$. The correlation coefficient is therefore also 0 indicating no linear relationship between the waiting times for bus and train.

7.8 Key Points and Further Issues

Note:

✓ Note that there is a difference between the empirical cumulative distribution function introduced in Chap. 2 and the CDF introduced in this chapter. In Chap. 2, we work with the cumulative distribution of observed values in a particular sample, whereas in this chapter, we deal with random variables modelling the distribution of a general population.

✓ The expectation and the variance of a random variable are defined as follows:

	Expectation	Variance
Discrete	$\sum_{i=1}^{n} x_i p_i$	$\sum_{i=1}^{n} (x_i - \mathrm{E}(X))^2 p_i$
Continuous	$\int_{-\infty}^{+\infty} x f(x) \mathrm{d}x$	$\int_{-\infty}^{+\infty} (x - \mathrm{E}(X))^2 f(x) \mathrm{d}x$

✓ Some important calculation rules are:

$$E(a + bX) = a + b\mathrm{E}(X); \quad \mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X);$$
$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y); \quad \mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2 \mathrm{Cov}(X, Y)$$

✓ Bivariate random variables (X, Y) have a joint CDF $F_{XY}(x, y)$ which specifies the probability $P(X \leq x; Y \leq y)$. The conditional distribution of $X|Y$ $[Y|X]$ is the PDF of X $[Y]$ for a given value $Y = y$ $[X = x]$. The marginal distribution of X $[Y]$ is the distribution of X $[Y]$ without referring to the values of Y $[X]$.

7.9 Exercises

Exercise 7.1 Consider the following cumulative distribution function of a random variable X:

$$F(x) = \begin{cases} 0 & \text{if } x < 2 \\ -\frac{1}{4}x^2 + 2x - 3 & \text{if } 2 \leq x \leq 4 \\ 1 & \text{if } x > 4. \end{cases}$$

(a) What is the PDF of X?
(b) Calculate $P(X < 3)$ and $P(X = 4)$.
(c) Determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Exercise 7.2 Joey manipulates a die to increase his chances of winning a board game against his friends. In each round, a die is rolled and larger numbers are generally an advantage. Consider the random variable X denoting the outcome of the rolled die and the respective probabilities $P(X = 1 = 2 = 3 = 5) = 1/9$, $P(X = 4) = 2/9$, and $P(X = 6) = 3/9$.

(a) Calculate and interpret the expectation and variance of X.
(b) Imagine that the board game contains an action which makes the players use $1/X$ rather than X. What is the expectation of $Y = 1/X$? Is $\mathrm{E}(Y) = \mathrm{E}(1/X) = 1/\mathrm{E}(X)$?

Exercise 7.3 An innovative winemaker experiments with new grapes and adds a new wine to his stock. The percentage sold by the end of the season depends on the weather and various other factors. It can be modelled using the random variable X with the CDF as

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 3x^2 - 2x^3 & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

(a) Plot the cumulative distribution function with R.
(b) Determine $f(x)$.
(c) What is the probability of selling at least one-third of his wine, but not more than two thirds?
(d) Define the CDF in R and calculate the probability of c) again.
(e) What is the variance of X?

Exercise 7.4 A quality index summarizes different features of a product by means of a score. Different experts may assign different quality scores depending on their experience with the product. Let X be the quality index for a tablet. Suppose the respective probability density function is given as follows:

$$f(x) = \begin{cases} cx(2 - x) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Determine c such that $f(x)$ is a proper PDF.
(b) Determine the cumulative distribution function.
(c) Calculate the expectation and variance of X.
(d) Use Tschebyschev's inequality to determine the probability that X does not deviate more than 0.5 from its expectation.

Exercise 7.5 Consider the joint PDF for the type of customer service X ($0 =$ telephonic hotline, $1 =$ Email) and of satisfaction score Y ($1 =$ unsatisfied, $2 =$ satisfied, $3 =$ very satisfied):

$X \backslash Y$	1	2	3
0	0	1/2	1/4
1	1/6	1/12	0

(a) Determine and interpret the marginal distributions of both X and Y.
(b) Calculate the 75 % quantile for the marginal distribution of Y.
(c) Determine and interpret the conditional distribution of satisfaction level for $X = 1$.
(d) Are the two variables independent?
(e) Calculate and interpret the covariance of X and Y.

Exercise 7.6 Consider a continuous random variable X with expectation 15 and variance 4. Determine the smallest interval $[15 - c, 15 + c]$ which contains at least 90 % of the values of X.

Exercise 7.7 Let X and Y be two random variables for which only 6 possible events—$A_1, A_2, A_3, A_4, A_5, A_6$—are defined:

i	1	2	3	4	5	6
$P(A_i)$	0.3	0.1	0.1	0.2	0.2	0.1
X_i	-1	2	2	-1	-1	2
Y_i	0	2	0	1	2	1

(a) What is the joint PDF of X and Y?
(b) Calculate the marginal distributions of X and Y.
(c) Are both variables independent?
(d) Determine the joint PDF for $U = X + Y$.
(e) Calculate $\mathrm{E}(U)$ and $\mathrm{Var}(U)$ and compare it with $\mathrm{E}(X) + \mathrm{E}(Y)$ and $\mathrm{Var}(X) + \mathrm{Var}(Y)$, respectively.

Exercise 7.8 Recall the urn model we introduced in Chap. 5. Consider an urn with eight balls: four of them are white, three are black, and one is red. Now, two balls are drawn from the urn. The random variables X and Y are defined as follows:

$$X = \begin{cases} 1 & \text{black ball} \\ 2 & \text{red ball in the first draw} \\ 3 & \text{white ball} \end{cases}$$

$$Y = \begin{cases} 1 & \text{black ball} \\ 2 & \text{red ball in the second draw} \\ 3 & \text{white ball.} \end{cases}$$

(a) When are X and Y independent—when the two balls are drawn with replacement or without replacement?
(b) Assume the balls are drawn such that X and Y are dependent. Use the conditional distribution $P(Y|X)$ to determine the joint PDF of X and Y.
(c) Calculate $\mathrm{E}(X)$, $\mathrm{E}(Y)$, and $\rho(X, Y)$.

Exercise 7.9 If X is the amount of money spent on food and other expenses during a day (in €) and Y is the daily allowance of a businesswoman, the joint density of these two variables is given by

$$f_{XY}(x, y) = \begin{cases} c\left(\frac{100-x}{x}\right) & \text{if } 10 \le x \le 100, \quad 40 \le y \le 100 \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Choose c such that $f_{XY}(x, y)$ is a probability density function.
(b) Find the marginal distribution of X.
(c) Calculate the probability that more than €75 are spent.
(d) Determine the conditional distribution of Y given X.

Exercise 7.10 Consider n i.i.d. random variables X_i with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ and the standardized variable $Y = \frac{X-\mu}{\sigma}$. Show that $\mathrm{E}(Y) = 0$ and $\mathrm{Var}(Y) = 1$.

→ Solutions to all exercises in this chapter can be found on p. 365

Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

Probability Distributions

8

We introduced the concept of probability density and probability mass functions of random variables in the previous chapter. In this chapter, we are introducing some common standard discrete and continuous probability distributions which are widely used for either practical applications or constructing statistical methods described later in this book. Suppose we are interested in determining the probability of a certain event. The determination of probabilities depends upon the nature of the study and various prevailing conditions which affect it. For example, the determination of the probability of a head when tossing a coin is different from the determination of the probability of rain in the afternoon. One can speculate that some mathematical functions can be defined which depict the behaviour of probabilities under different situations. Such functions have special properties and describe how probabilities are distributed under different conditions. We have already learned that they are called probability distribution functions. The form of such functions may be simple or complicated depending upon the nature and complexity of the phenomenon under consideration. Let us first recall and extend the definition of independent and identically distributed random variables:

Definition 8.0.1 The random variables $X_1, X_2, \ldots, X_n$ are called independent and identically distributed (i.i.d) if the $X_i (i = 1, 2, \ldots, n)$ have the same marginal cumulative distribution function $F(x)$ and if they are mutually independent.

Example 8.0.1 Suppose a researcher plans a survey on the weight of newborn babies in a country. The researcher randomly contacts 10 hospitals with a maternity ward and asks them to randomly select 20 of the newborn babies (no twins) born in the last 6 months and records their weights. The sample therefore consists of $10 \times 20 = 200$ baby weights. Since the hospitals and the babies are randomly selected, the babies' weights are therefore not known beforehand. The 200 weights can be denoted by the random variables $X_1, X_2, \ldots, X_{200}$. Note that the weights X_i are random variables

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_8

because, depending on the size of the population, different samples consisting of 200 babies can be randomly selected. Also, the babies' weights can be seen as stochastically independent (an example of stochastically dependent weights would be the weights of twins if they are included in the sample). After collecting the weights of 200 babies, the researcher has a sample of 200 realized values (i.e. the weights in grams). The values are now known and denoted by $x_1, x_2, \ldots, x_{200}$.

8.1 Standard Discrete Distributions

First, we discuss some standard distributions for discrete random variables.

8.1.1 Discrete Uniform Distribution

The discrete uniform distribution assumes that all possible outcomes have equal probability of occurrence. A more formal definition is given as follows:

Definition 8.1.1 A discrete random variable X with k possible outcomes $x_1, x_2, \ldots, x_k$ is said to follow a discrete **uniform** distribution if the probability mass function (PMF) of X is given by

$$P(X = x_i) = \frac{1}{k}, \quad \forall i = 1, 2, \ldots, k. \tag{8.1}$$

If the outcomes are the natural numbers $x_i = i$ $(i = 1, 2, \ldots, k)$, the mean and variance of X are obtained as

$$\mathrm{E}(X) = \frac{k+1}{2}, \tag{8.2}$$

$$\mathrm{Var}(X) = \frac{1}{12}(k^2 - 1). \tag{8.3}$$

Example 8.1.1 If we roll a fair die, the outcomes "1", "2", ..., "6" have equal probability of occurring, and hence, the random variable X "number of dots observed on the upper surface of the die" has a uniform discrete distribution with PMF

$$P(X = i) = \frac{1}{6}, \quad \text{for all} \quad i = 1, 2, \ldots, 6.$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{6+1}{2} = 3.5,$$

$$\mathrm{Var}(X) = \frac{1}{12}(6^2 - 1) = 35/12.$$

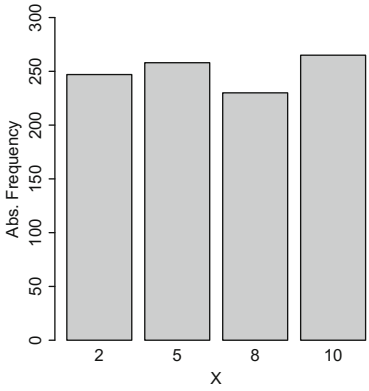


Fig. 8.1 Frequency distribution of 1000 generated discrete uniform random numbers with possible outcomes (2, 5, 8, 10)

Using the function `sample()` in *R*, it is easy to generate random numbers from a discrete uniform distribution. The following command generates a random sample of size 1000 from a uniform distribution with the four possible outcomes 2, 5, 8, 10 and draws a bar chart of the observed numbers. The use of the `set.seed()` function allows to reproduce the generated random numbers at any time. It is necessary to use the option `replace=TRUE` to simulate draws with replacement, i.e. to guarantee that a value can occur more than once.

```
set.seed(123789)
x <- sample(x=c(2,5,8,10), size=1000, replace=T,
prob=c(1/4,1/4,1/4,1/4))
barchart(table(x), ylim=c(0,300))
```

A bar chart of the frequency distribution of the 1000 sampled numbers with the possible outcomes (2, 5, 8, 10) using the discrete uniform distribution is given in Fig. 8.1. We see that the 1000 generated random numbers are not exactly uniformly distributed, e.g. the numbers 5 and 10 occur more often than the numbers 2 and 8. In fact, they are only approximately uniform. We expect that the deviance from a perfect uniform distribution is getting smaller as we generate more and more random numbers but will probably never be zero for a finite number of draws. The random numbers reflect the practical situation that a sample distribution is only an approximation to the theoretical distribution from which the sample was drawn. More details on how to work with random variables in *R* are given in Appendix A.3.

8.1.2 Degenerate Distribution

Definition 8.1.2 A random variable X has a **degenerate distribution** at a, if a is the only possible outcome with $P(X = a) = 1$. The CDF in such a case is given by

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } x \geq a. \end{cases}$$

Further, $\mathrm{E}(X) = a$ and $\mathrm{Var}(X) = 0$.

The degenerate distribution indicates that there is only one possible fixed outcome, and therefore, no randomness is involved. It follows that we need at least two different possible outcomes to have randomness in the observations of a random variable or random experiment. The Bernoulli distribution is such a distribution where there are only two outcomes, e.g. success and failure or male and female. These outcomes are usually denoted by the values "0" and "1".

8.1.3 Bernoulli Distribution

Definition 8.1.3 A random variable X has a Bernoulli distribution if the PMF of X is given as

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0. \end{cases}$$

The cumulative distribution function (CDF) of X is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

The mean (expectation) and variance of a Bernoulli random variable are calculated as

$$\mathrm{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p \tag{8.4}$$

and

$$\mathrm{Var}(X) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p), \tag{8.5}$$

respectively.

A Bernoulli distribution is useful when there are only two possible outcomes, and our interest lies in any of the two outcomes, e.g. whether a customer buys a certain product or not, or whether a hurricane hits an island or not. The outcome of an event A is usually coded as 1 which occurs with probability p. If the event of interest does not occur, i.e. the complementary event $\bar{A}$ occurs, the outcome is coded as 0 which occurs with probability $1 - p$. So p is the probability that the event of interest A occurs.

Example 8.1.2 A company organizes a raffle at an end-of-year function. There are 300 lottery tickets in total, and 50 of them are marked as winning tickets. The event A of interest is "ticket wins" (coded as $X = 1$), and the probability p of having a winning ticket is *a priori* (i.e. before any lottery ticket has been drawn)

$$P(X = 1) = \frac{50}{300} = \frac{1}{6} = p \quad \text{and} \quad P(X = 0) = \frac{250}{300} = \frac{5}{6} = 1 - p.$$

According to (8.4) and (8.5), the mean (expectation) and variance of X are

$$E(X) = \frac{1}{6} \quad \text{and} \quad \text{Var}(X) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36} \text{ respectively.}$$

8.1.4 Binomial Distribution

Consider n independent trials or repetitions of a Bernoulli experiment. In each trial or repetition, we may observe either A or $\bar{A}$. At the end of the experiment, we have thus observed A between 0 and n times. Suppose we are interested in the probability of A occurring k times, then the binomial distribution is useful.

Example 8.1.3 Consider a coin tossing experiment where a coin is tossed ten times and the event of interest is $A =$ "head". The random variable X "number of heads in 10 experiments" has the possible outcomes $k = 0, 1, \ldots, 10$. A question of interest may be: What is the probability that a head occurs in 7 out of 10 trials; or in 5 out of 10 trials? We assume that the order in which heads (and tails) appear is not of interest, only the total number of heads is of interest.

Questions of this kind are answered by the binomial distribution. This distribution can either be motivated as a repetition of n Bernoulli experiments (as in the above coin tossing example) or by the urn model (see Chap. 5): assume there are M white and $N - M$ black balls in the urn. Suppose n balls are drawn randomly from the urn, the colour of the ball is recorded and the ball is placed back into the urn (sampling with replacement). Let A be the event of interest that a white ball is drawn from the urn. The probability of A is $p = M/N$ (the probability of drawing a black ball is $1 - p = (N - M)/N$). Since the balls are drawn with replacement, these probabilities do not change from draw to draw. Further, let X be the random variable counting the number of white balls drawn from the urn in the n experiments. Since the order of the resulting colours of balls is not of interest in this case, there are $\binom{n}{k}$ combinations where k balls are white and $n - k$ balls are black. Since the balls are drawn with replacement, every outcome of the n experiments is independent of all others. The probability that $X = k$, $k = 0, 1, \ldots, n$, can therefore be calculated as

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (k = 0, 1, \ldots, n). \tag{8.6}$$

Please note that we can use the product $p^k(1 - p)^{n-k}$ because the draws are independent. The binomial coefficient $\binom{n}{k}$ is necessary to count the number of possible orders of the black and white balls.

Definition 8.1.4 A discrete random variable X is said to follow a binomial distribution with parameters n and p if its PMF is given by (8.6). We also write $X \sim B(n; p)$. The mean and variance of a binomial random variable X are given by

$$\mathrm{E}(X) = np, \tag{8.7}$$

$$\mathrm{Var}(X) = np(1-p). \tag{8.8}$$

Remark 8.1.1 A Bernoulli random variable is therefore $B(1; p)$ distributed.

Example 8.1.4 Consider an unfair coin where the probability of observing a tail (T) is $p(T) = 0.6$. Let us denote tails by "1" and heads by "0". Suppose the coin is tossed three times. In total, there are the $2^3 = 8$ following possible outcomes:

Outcome			$X = x$
1	1	1	3
1	1	0	2
1	0	1	2
0	1	1	2
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0

Note that the first outcome, viz. (1, 1, 1) leads to $x = 3$, the next 3 outcomes, viz., (1, 1, 0), (1, 0, 1), (0, 1, 1) obtained by $(= \binom{3}{2})$ lead to $x = 2$, the next 3 outcomes, viz., (1, 0, 0), ((0, 1, 0), (0, 0, 1) obtained by $(= \binom{3}{1})$ lead to $x = 1$, and the last outcome, viz. (0, 0, 0) obtained by $(= \binom{3}{0})$ leads to $x = 0$. We can, for example, calculate

$$P(X = 2) = \binom{3}{2} 0.6^2 (1 - 0.6)^1 = 0.432 \quad (\text{or} \quad 43.2\,\%).$$

Further, the mean and variance of X are

$$\mathrm{E}(X) = np = 3 \cdot 0.6 = 1.8, \quad \text{and} \quad \mathrm{Var}(X) = np(1-p) = 3 \cdot 0.6 \cdot 0.4 = 0.72.$$

Functions for the binomial distribution, as well as many other distributions, are implemented in *R*. For each of these distributions, we can easily determine the density function (PMF, PDF) for given values and parameters, determine the CDF, calculate quantiles and draw random numbers. Appendix A.3 gives more details. Nevertheless, we illustrate the concept of dealing with distributions in *R* in the following example.

Example 8.1.5 Suppose we roll an unfair die 50 times with the probability of a tail $p_{\text{tail}} = 0.6$. We thus deal with a $B(50, 0.6)$ distribution which can be plotted using the `dbinom` command. The prefix `d` stands for "density".

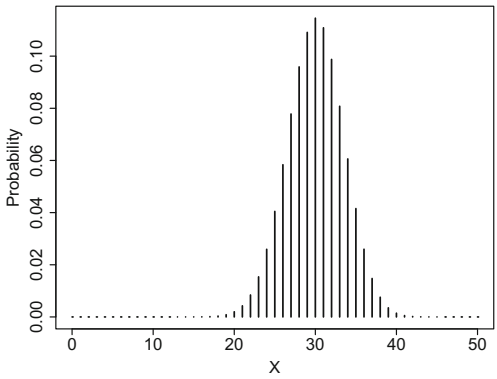


Fig. 8.2 PMF of a $B(50, 0.6)$ distribution

```
n <- 50
p <- 0.6
k <- 0:n
pmf <- dbinom(k,n,p)
plot(k,pmf, type=h)
```

A plot of the PMF of a binomial distribution with $n = 50$ and $p = 0.6$ (i.e. $B(50, 0.6)$) is given in Fig. 8.2.

Note that we can also calculate the CDF with *R*. We can use the `pbinom(x,n,p)` command, where the prefix `p` stands for probability, to calculate the CDF at any point. For example, suppose we are interested in $P(X \geq 30) = 1 - F(29)$, that is the probability of observing thirty or more tails; then we write

```
1-pbinom(29,50,0.6)
[1] 0.5610349
```

Similarly, we can determine quantiles. For instance, the 80 % quantile q which describes that $P(X \leq q) \geq 0.8$ can be obtained by the `qbinom(q,n,p)` command as follows:

```
qbinom(0.8,50,0.6)
[1] 33
```

If we want to generate 100 random realizations from a $B(50, 0.6)$ distribution we can use the `rbinom` command.

```
rbinom(100,50,0.6)
```

The binomial distribution has some nice properties. One of them is described in the following theorem:

Theorem 8.1.1 *Let $X \sim B(n; p)$ and $Y \sim B(m; p)$ and assume that X and Y are (stochastically) independent. Then*

$$X + Y \sim B(n + m; p). \tag{8.9}$$

This is intuitively clear since we can interpret this theorem as describing the additive combination of two independent binomial experiments with n and m trials, with equal probability p, respectively. Since every binomial experiment is a series of independent Bernoulli experiments, this is equivalent to a series of $n + m$ independent Bernoulli trials with constant success probability p which in turn is equivalent to a binomial distribution with $n + m$ trials.

8.1.5 Poisson Distribution

Consider a situation in which the number of events is very large and the probability of success is very small: for example, the number of alpha particles emitted by a radioactive substance entering a particular region in a given short time interval. Note that the number of emitted alpha particles is very high but only a few particles are transmitted through the region in a given short time interval. Some other examples where Poisson distributions are useful are the number of flu cases in a country within one year, the number of tropical storms within a given area in one year, or the number of bacteria found in a biological investigation.

Definition 8.1.5 A discrete random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ if its PMF is given by

$$P(X = x) = \frac{\lambda^x}{x!} \exp(-\lambda) \quad (x = 0, 1, 2, \ldots). \tag{8.10}$$

We also write $X \sim Po(\lambda)$. The mean and variance of a Poisson random variable are identical:

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda.$$

Example 8.1.6 Suppose a country experiences $X = 4$ tropical storms on average per year. Then the probability of suffering from only two tropical storms is obtained by using the Poisson distribution as

$$P(X = 2) = \frac{\lambda^x}{x!} \exp(-\lambda) = \frac{4^2}{2!} \exp(-4) = 0.146525.$$

If we are interested in the probability that not more than 2 storms are experienced, then we can apply rules (7.7)–(7.13) from Chap. 7: $P(X \leq 2) = P(X = 2) + P(X = 1) + P(X = 0) = F(2) = 0.2381033$. We can calculate $P(X = 1)$ and $P(X = 0)$ from (8.10) or using *R*. Similar to Example 8.1.5, we use the prefix `d` to obtain the PMF and the prefix `p` to work with the CDF, i.e. we can use `dpois(x,`λ`)` and `ppois(x,`λ`)` to determine $P(X = x)$ and $P(X \leq x)$, respectively.

```
dpois(2,4) + dpois(1,4) + dpois(0,4)
[1] 0.2381033
ppois(2,4)
[1] 0.2381033
```

8.1.6 Multinomial Distribution

We now consider random experiments where k distinct or disjoint events $A_1, A_2, \ldots, A_k$ can occur with probabilities $p_1, p_2, \ldots, p_k$, respectively, with the restriction $\sum_{j=1}^{k} p_j = 1$. For example, if eight parties compete in a political election, we may be interested in the probability that a person votes for party A_j, $j = 1, 2, \ldots, 8$. Similarly one might be interested in the probability whether tuberculosis is detected in the lungs (A_1), in other organs (A_2), or both (A_3). Practically, we often use the multinomial distribution to model the distribution of categorical variables. This can be interpreted as a generalization of the binomial distribution (where only two distinct events can occur) to the situation where more than two events or outcomes can occur. If the experiment is repeated n times independently, we are interested in the probability that

$$A_1 \text{ occurs } n_1\text{-times}, A_2 \text{ occurs } n_2\text{-times}, \ldots, A_k \text{ occurs } n_k\text{-times}$$

with $\sum_{j=1}^{k} n_j = n$. Since several events can occur, the outcome of one (of the n) experiments is conveniently described by binary indicator variables. Let $V_{ij}, i = 1, \ldots, n, j = 1, \ldots, k$, denote the event "$A_j$ is observed in experiment i", i.e.

$$V_{ij} = \begin{cases} 1 & \text{if } A_j \text{ occurs in experiment } i \\ 0 & \text{if } A_j \text{ does not occur in experiment } i \end{cases}$$

with probabilities $P(V_{ij} = 1) = p_j$, $j = 1, 2, \ldots, k$; then, the outcome of one experiment is a vector of length k,

$$V_i = (V_{i1}, \ldots, V_{ij}, \ldots, V_{ik}) = (0, \ldots, 1, \ldots, 0),$$

with "1" being present in only one position, i.e. in position j, if A_j occurs in experiment i. Now, define (for each $j = 1, \ldots, k$) $X_j = \sum_{i=1}^{n} V_{ij}$. Then, X_j is counting how often event A_j was observed in the n independent experiments (i.e. how often V_{ij} was 1 in the n experiments).

Definition 8.1.6 The random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ is said to follow a **multinomial distribution** if its PMF is given as

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} \cdot p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \tag{8.11}$$

with the restrictions $\sum_{j=1}^{k} n_j = n$ and $\sum_{j=1}^{k} p_j = 1$. We also write $\mathbf{X} \sim M(n; p_1, \ldots, p_k)$. The mean of $\mathbf{X}$ is the (component-wise) vector

$$\begin{aligned} \mathrm{E}(\mathbf{X}) &= (\mathrm{E}(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_k)) \\ &= (np_1, np_2, \ldots, np_k). \end{aligned}$$

The (i, j)th element of the covariance matrix $V(\mathbf{X})$ is

$$\text{Cov}(X_i, X_j) = \begin{cases} np_i(1-p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j. \end{cases}$$

Remark 8.1.2 Due to the restriction that $\sum_{j=1}^{k} n_j = \sum_{j=1}^{k} X_j = n$, $X_1, \ldots, X_k$ are not stochastically independent which is reflected by the negative covariance. This is also intuitively clear: if one X_j gets higher, another $X_{j'}$, $j \neq j'$, has to become lower to satisfy the restrictions.

We use the multinomial distribution to describe the randomness of categorical variables. Suppose we are interested in the variable "political party"; there might be eight political parties, and we could thus summarize this variable by eight binary variables, each of them describing the event of party A_j, $j = 1, 2, \ldots, 8$, being voted for. In this sense, $\mathbf{X} = (X_1, X_2, \ldots, X_8)$ follows a multinomial distribution.

Example 8.1.7 Consider a simple example of the urn model. The urn contains 50 balls of three colours: 25 red balls, 15 white balls, and 10 black balls. The balls are drawn from the urn with replacement. The balls are placed back into the urn after every draw, which means the draws are independent. Therefore, the probability of drawing a red ball in every draw is $p_1 = \frac{25}{50} = 0.5$. Analogously, $p_2 = 0.3$ (for white balls) and $p_3 = 0.2$ (for black balls). Consider $n = 4$ draws. The probability of the random event of drawing "2 red balls, 1 white ball, and 1 black ball" is:

$$P(X_1 = 2, X_2 = 1, X_3 = 1) = \frac{4!}{2!1!1!}(0.5)^2(0.3)^1(0.2)^1 = 0.18. \quad (8.12)$$

We would have obtained the same result in *R* using the `dmultinom` function:

```
dmultinom(c(2,1,1),prob=c(0.5,0.3,0.2))
```

This example demonstrates that the multinomial distribution relates to an experiment with replacement and without considering the order of the draws. Instead of the urn model, consider another example where we may interpret these three probabilities as probabilities of voting for candidate A_j, $j = 1, 2, 3$, in an election. Now, suppose we ask four voters about their choice, then the probability of candidate A_1 receiving 2 votes, candidate A_2 receiving 1 vote, and candidate A_3 receiving 1 vote is 18 % as calculated in (8.12).

Remark 8.1.3 In contrast to most of the distributions, the CDF of the multinomial distribution, i.e. the function calculating $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k)$, is not contained in the base *R*-distribution. Please note that for $k = 2$, the multinomial distribution reduces to the binomial distribution.

8.1.7 Geometric Distribution

Consider a situation in which we are interested in determining how many independent Bernoulli trials are needed until the event of interest occurs for the first time. For instance, we may be interested in how many tickets to buy in a raffle until we win for the first time, or how many different drugs to try to successfully tackle a severe migraine, etc. The geometric distribution can be used to determine the probability that the event of interest happens at the kth trial for the first time.

Definition 8.1.7 A discrete random variable X is said to follow a geometric distribution with parameter p if its PMF is given by

$$P(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \ldots \tag{8.13}$$

The mean (expectation) and variance are given by $\mathrm{E}(X) = 1/p$ and $\mathrm{Var}(X) = 1/p(1/p - 1)$, respectively.

Example 8.1.8 Let us consider an experiment where a coin is tossed until "head" is obtained for the first time. The probability of getting a head is $p = 0.5$ for each toss. Using (8.13), we can determine the following probabilities:

$$\begin{aligned} P(X = 1) &= 0.5 \\ P(X = 2) &= 0.5(1 - 0.5) = 0.25 \\ P(X = 3) &= 0.5(1 - 0.5)^2 = 0.125 \\ P(X = 4) &= 0.5(1 - 0.5)^3 = 0.0625 \\ &\ldots \quad \ldots \end{aligned}$$

Using the command structure for obtaining PMF's in R (Appendix A as well as Examples 8.1.5 and 8.1.6), we can determine the latter probability of $P(X = 4)$ as follows:

```
dgeom(3,0.5)
```

Note that the definition of X in R slightly differs from our definition. In R, k is the number of failures before the first success. This means we need to specify $k - 1$ in the `dgeom` function rather than k. The mean and variance for this setting are

$$\mathrm{E}(X) = \frac{1}{0.5} = 2; \quad \mathrm{Var}(X) = \frac{1}{0.5}\left(\frac{1}{0.5} - 1\right) = 2.$$

8.1.8 Hypergeometric Distribution

We can again use the urn model to motivate another distribution, the hypergeometric distribution. Consider an urn with We randomly draw n balls without replacement,

M	white balls
$N-M$	black balls
N	total balls

i.e. we do not place a ball back into the urn once it is drawn. The order in which the balls are drawn is assumed to be of no interest; only the number of drawn white balls is of relevance. We define the following random variable

$$X: \text{“number of white balls } (x) \text{ among the } n \text{ drawn balls”}.$$

To be more precise, among the n drawn balls, x are white and $n-x$ are black. There are $\binom{M}{x}$ possibilities to choose x white balls from the total of M white balls, and analogously, there are $\binom{N-M}{n-x}$ possibilities to choose $(n-x)$ black balls from the total of $N-M$ black balls. In total, we draw n out of N balls. Recall the probability definition of Laplace as the number of simple favourable events divided by all possible events. The number of combinations for all possible events is $\binom{N}{n}$; the number of favourable events is $\binom{M}{x}\binom{N-M}{n-x}$ because we draw, independent of each other, x out of M balls and $n-x$ out of $N-M$ balls. Hence, the PMF of the hypergeometric distribution is

$$P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{8.14}$$

for $x \in \{\max(0, n-(N-M)), \ldots, \min(n, M)\}$.

Definition 8.1.8 A random variable X is said to follow a **hypergeometric** distribution with parameters n, M, N, i.e. $X \sim H(n, M, N)$, if its PMF is given by (8.14).

Example 8.1.9 The German national lottery draws 6 out of 49 balls from a rotating bowl. Each ball is associated with a number between 1 and 49. A simple bet is to choose 6 numbers between 1 and 49. If 3 or more chosen numbers correspond to the numbers drawn in the lottery, then one wins a certain amount of money. What is the probability of choosing 4 correct numbers? We can utilize the hypergeometric distribution with $x=4$, $M=6$, $N=49$, and $n=6$ to calculate such probabilities. The interpretation is that we “draw” (i.e. bet on) 4 out of the 6 winning balls and “draw” (i.e. bet on) another 2 out of the remaining 43 $(49-6)$ losing balls. In total, we draw 6 out of 49 balls. Calculating the number of the favourable combinations and all possible combinations leads to the application of the hypergeometric distribution as follows:

$$P(X=4) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} \approx 0.001 \text{ (or } \quad 0.1\,\%).$$

We would have obtained the same results using the `dhyper` command. Its arguments are `x, M, N, n`, and thus, we specify

```
dhyper(4,6,43,6)
```

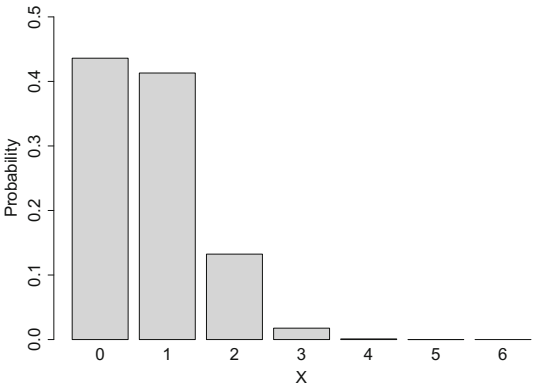


Fig. 8.3 The $H(6, 43, 6)$ distribution

The $H(6, 43, 6)$ distribution is also visualized in Fig. 8.3. It is evident that the cumulative probability of choosing 2 or fewer correct numbers is greater than 0.9 (or 90 %), but it is very unlikely to have 3 or more numbers right. This may explain why the national lottery pays out money only for 3 or more correct numbers.

8.2 Standard Continuous Distributions

Now, we discuss some standard probability distributions of (absolute) continuous random variables. Characteristics of continuous random variables are that the number of possible outcomes is uncountably infinite and that they have a continuous distribution function $F(x)$. It follows that the point probabilities are zero, i.e. $P(X = x) = 0$. Further, we assume a unique density function f exists, such that $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$.

8.2.1 Continuous Uniform Distribution

A continuous analogue to the discrete uniform distribution is the continuous uniform distribution on a closed interval in $\mathbb{R}$.

Definition 8.2.1 A continuous random variable X is said to follow a (continuous) **uniform distribution** in the interval $[a, b]$, i.e. $X \sim U(a, b)$, if its probability density function (PDF) is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad a \leq x \leq b \quad (a < b) \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance of $X \sim U(a, b)$ are

$$\mathrm{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12},$$

respectively.

Example 8.2.1 Suppose a train arrives at a subway station regularly every 10 min. If a passenger arrives at the station without knowing the timetable, then the waiting time to catch the train is uniformly distributed with density

$$f(x) = \begin{cases} \frac{1}{10} & \text{if } 0 \le x \le 10 \\ 0 & \text{otherwise.} \end{cases}$$

The "average" waiting time is $\mathrm{E}(X) = (10+0)/2 = 5$ min. The probability of waiting for the train for less than 3 min is obviously 0.3 (or 30 %) which can be calculated in *R* using the `punif(x,a,b)` command (see also Appendix A.3):

```
punif(3,0,10)
```

8.2.2 Normal Distribution

The normal distribution is one of the most important distributions used in statistics. The name was given by Carl Friedrich Gauss (1777–1855), a German mathematician, astronomer, geodesist, and physicist who observed that measurements in geodesy and astronomy randomly deviate in a symmetric way from their true values. The normal distribution is therefore also often called a Gaussian distribution.

Definition 8.2.2 A random variable X is said to follow a **normal distribution** with parameters μ and σ^2 if its PDF is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \qquad -\infty < x < \infty, -\infty < \mu < \infty, \sigma^2 > 0. \tag{8.15}$$

We write $X \sim N(\mu, \sigma^2)$. The mean and variance of X are

$$\mathrm{E}(X) = \mu; \quad \text{and} \quad \mathrm{Var}(X) = \sigma^2,$$

respectively. If $\mu = 0$ and $\sigma^2 = 1$, then X is said to follow a ***standard*** **normal distribution**, $X \sim N(0, 1)$. The PDF of a standard normal distribution is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}); \quad -\infty < x < \infty.$$

The density of a normal distribution has its maximum (see Fig. 8.4) at $x = \mu$. The density is also symmetric around μ. The inflexion points of the density are at $(\mu - \sigma)$ and $(\mu + \sigma)$ (Fig. 8.4). A lower σ indicates a higher concentration around the mean μ. A higher σ indicates a flatter density (Fig. 8.5).

The cumulative distribution function of $X \sim N(\mu, \sigma^2)$ is

$$F(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t \tag{8.16}$$

which is often denoted as $\Phi(x)$. The value of $\Phi(x)$ for various values of x can be obtained in *R* following the rules introduced in Appendix A.3. For example,

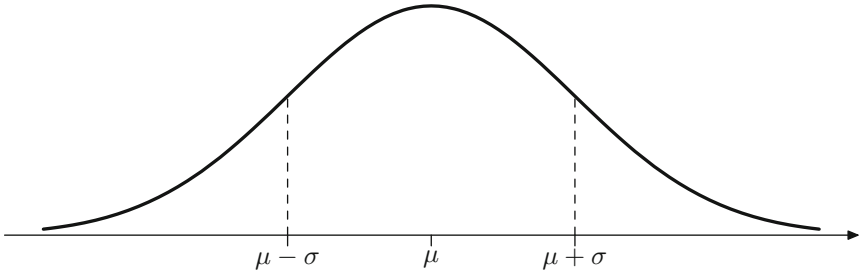


Fig. 8.4 PDF of a normal distribution*

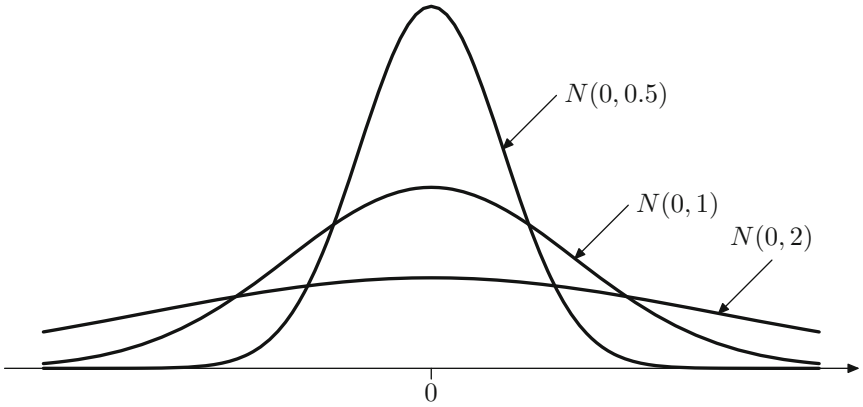


Fig. 8.5 PDF of $N(0, 2)$, $N(0, 1)$ and $N(0, 0.5)$ distributions*

```
pnorm(1.96, mean = 0, sd = 1)
```

calculates $\Phi(1.96)$ as approximately 0.975. This means, for a standard normal distribution the probability $P(X \leq 1.96) \approx 0.975$.

Remark 8.2.1 There is no explicit formula to solve the integral in Eq. (8.16). It has to be solved by numerical (or computational) methods. This is the reason why CDF tables are presented in almost all statistical textbooks, see Table C.1 in the Appendix C.

Example 8.2.2 An orange farmer sells his oranges in wooden boxes. The weights of the boxes vary and are assumed to be normally distributed with $\mu = 15$ kg and $\sigma^2 = \frac{9}{4}$ kg^2. The farmer wants to avoid customers being unsatisfied because the boxes are too low in weight. He therefore asks the following question: What is the probability that a box with a weight of less than 13 kg is sold? Using the `pnorm(x, `μ`, `σ`)` command in *R*, we get

```
pnorm(13,15,sqrt(9/4))
[1] 0.09121122
```

To calculate the probability in Example 8.2.2 manually, we first have to introduce some theoretical results.

Calculation rules for normal random variables.

Let $X \sim N(\mu, \sigma^2)$. Using the transformation

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \tag{8.17}$$

every normally distributed random variable can be transformed into a *standard* normal random variable. We call this transformation the Z-transformation. We can use this transformation to derive convenient calculation rules. The probability for $X \leq b$ is

$$P(X \leq b) = P\left(\frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = P\left(Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right). \tag{8.18}$$

Consequently, the probability for $X > a$ is

$$P(X > a) = 1 - P(X \leq a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right). \tag{8.19}$$

The probability that X realizes a value in the interval $[a, b]$ is

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \tag{8.20}$$

Because of the symmetry of the probability density function $\phi(x)$ around its mean 0, the following equation holds for the distribution function $\Phi(x)$ of a standard normal random variable for any value a:

$$\Phi(-a) = 1 - \Phi(a). \tag{8.21}$$

It follows that $P(-a < Z < a) = 2 \cdot \Phi(a) - 1$, see also Fig. 8.6.

Example 8.2.3 Recall Example 8.2.2 where a farmer sold his oranges. He was interested in $P(X \leq 13)$ for $X \sim N(15, 9/4)$. Using (8.17), we get

$$\begin{aligned} P(X \leq 13) &= \Phi\left(\frac{13 - 15}{\frac{3}{2}}\right) \\ &= \Phi\left(-\frac{4}{3}\right) = 1 - \Phi\left(\frac{4}{3}\right) \approx 0.091 \text{ (or } 9.1\,\%). \end{aligned}$$

To obtain $\Phi(4/3) \approx 90.9\,\%$, we could either use *R* (`pnorm(4/3)`) or use Table C.1.

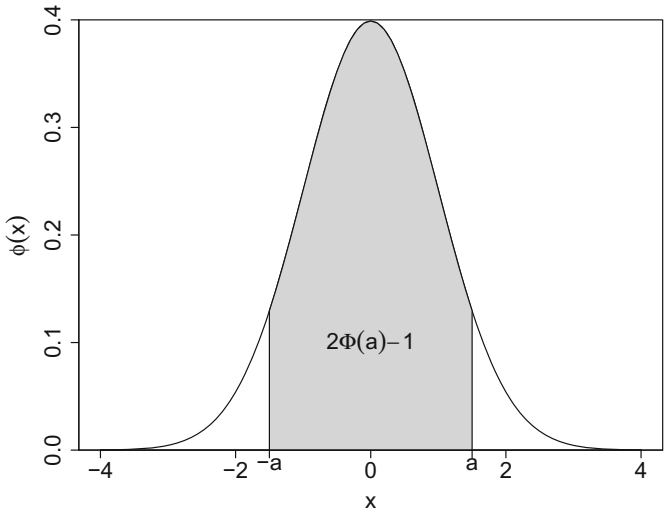


Fig. 8.6 Distribution function of the standard normal distribution

Distribution of the Arithmetic Mean.

Assume that $X \sim N(\mu, \sigma^2)$. Consider a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of independent and identically distributed random variables X_i with $X_i \sim N(\mu, \sigma^2)$. Then, the arithmetic mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ follows a normal distribution with mean

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \mu$$

and variance

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n} \tag{8.22}$$

where $\mathrm{Cov}(X_i, X_j) = 0 \quad \text{for} \quad i \neq j$. In summary, we get

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Remark 8.2.2 In fact, in Eq. (8.22), we have used the fact that the sum of normal random variables also follows a normal distribution, i.e.

$$(X_1 + X_2) \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

This result can be generalized to n (not necessarily identically distributed but independent) normal random variables. In fact, it holds that if $X_1, X_2, \ldots, X_n$ are independent normal variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then for any real numbers $a_1, a_2, \ldots, a_n$, it holds that

$$(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \sim N\left(a_1\mu_1 + a_2\mu_2 + \cdots a_n\mu_n, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots a_n^2\sigma_n^2\right).$$

In general, it cannot be taken for granted that the sum of two random variables follows the same distribution as the two variables themselves. As an example, consider the sum of two independent uniform distributions with $X_1 \sim U[0, 10]$ and $X_2 \sim U[20, 30]$. It holds that $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$ and $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$, but $X_1 + X_2$ is obviously not uniformly distributed.

8.2.3 Exponential Distribution

The exponential distribution is useful in many situations, for example when one is interested in the waiting time, or lifetime, until an event of interest occurs. If we assume that the future lifetime is independent of the lifetime that has already taken place (i.e. no "ageing" process is working), the waiting times can be considered to be exponentially distributed.

Definition 8.2.3 A random variable X is said to follow an exponential distribution with parameter $\lambda > 0$ if its PDF is given by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } \ x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8.23}$$

We write $X \sim Exp(\lambda)$. The mean and variance of an exponentially distributed random variable X are

$$\mathrm{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2},$$

respectively. The CDF of the exponential distribution is given as

$$F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{if } \ x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8.24}$$

Note, that $P(X > x) = 1 - F(x) = \exp(-\lambda x)$ $(x \geq 0)$. An interesting property of the exponential distribution is its **memorylessness**: if time t has already been reached, the probability of reaching a time greater than $t + \Delta$ does not depend on t. This can be written as

$$P(X > t + \Delta | X > t) = P(X > \Delta) \quad t, \Delta > 0.$$

The result can be derived using basic probability rules as follows:

$$\begin{aligned} P(X > t + \Delta | X > t) &= \frac{P(X > t + \Delta \text{ and } X > t)}{P(X > t)} = \frac{P(X > t + \Delta)}{P(X > t)} \\ &= \frac{\exp[-\lambda(t + \Delta)]}{\exp[-\lambda t]} = \exp[-\lambda \Delta] \\ &= 1 - F(\Delta) = P(X > \Delta). \end{aligned}$$

For example, suppose someone stands in a supermarket queue for t minutes. Say the person forgot to buy milk, so she leaves the queue, gets the milk, and stands in the queue again. If we use the exponential distribution to model the waiting time, we say that it does not matter what time it is: the random variable "waiting time from standing in the queue until paying the bill" is not influenced by how much

time has elapsed already; it does not matter if we queued before or not. Please note that the memorylessness property is shared by the geometric and the exponential distributions.

There is also a relationship between the Poisson and the exponential distribution:

Theorem 8.2.1 *The number of events Y occurring within a continuum of time is Poisson distributed with parameter λ if and only if the time between two events is exponentially distributed with parameter λ.*

The continuum of time depends on the problem at hand. It may be a second, a minute, 3 months, a year, or any other time period.

Example 8.2.4 Let Y be the random variable which counts the "number of accesses per second for a search engine". Assume that Y is Poisson distributed with parameter $\lambda = 10$ ($\mathrm{E}(Y) = 10$, $\mathrm{Var}(Y) = 10$). The random variable X, "waiting time until the next access", is then exponentially distributed with parameter $\lambda = 10$. We therefore get

$$\mathrm{E}(X) = \frac{1}{10}, \quad \mathrm{Var}(X) = \frac{1}{10^2}.$$

In this example, the continuum is 1 s. The expected number of accesses per second is therefore $\mathrm{E}(Y) = 10$, and the expected waiting time between two accesses is $\mathrm{E}(X) = 1/10$ s. The probability of experiencing a waiting time of less than 0.1 s is

$$F(0.1) = 1 - \exp(-\lambda x) = 1 - \exp(-10 \cdot 0.1) \approx 0.63.$$

In R, we can obtain the same result as

```
pexp(0.1,10)
[1] 0.6321206
```

8.3 Sampling Distributions

All the distributions introduced in this chapter up to now are motivated by practical applications. However, there are theoretical distributions which play an important role in the construction and development of various statistical tools such as those introduced in Chaps. 9–11. We call these distributions "sampling distributions". Now, we discuss the χ^2-, t-, and F-distributions.

8.3.1 χ^2-Distribution

Definition 8.3.1 Let $Z_1, Z_2 \ldots, Z_n$ be n independent and identically $N(0, 1)$-distributed random variables. The sum of their squares, $\sum_{i=1}^{n} Z_i^2$, is then **χ^2-distributed** with n degrees of freedom and is denoted as χ_n^2. The PDF of the χ^2-distribution is given in Eq. (C.7) in Appendix C.3.

The χ^2-distribution is not symmetric. A χ^2-distributed random variable can only realize values greater than or equal to zero. Figure 8.7a shows the χ^2_1- , χ^2_2- , and χ^2_5-distributions. It can be seen that the "degrees of freedom" specify the shape of the distribution. Their interpretation and meaning will nevertheless become clearer in the following chapters. The quantiles of the CDF of different χ^2-distributions can be obtained in *R* using the `qchisq(p,df)` command. They are also listed in Table C.3 for different values of n.

Theorem 8.3.1 *Consider two independent random variables which are χ^2_m- and χ^2_n-distributed, respectively. The sum of these two random variables is χ^2_{n+m}-distributed.*

An important example of a χ^2-distributed random variable is the sample variance (S^2_X) of an i.i.d. sample of size n from a normally distributed population, i.e.

$$\frac{(n-1)S^2_X}{\sigma^2} \sim \chi^2_{n-1}. \tag{8.25}$$

8.3.2 *t*-Distribution

Definition 8.3.2 Let X and Y be two independent random variables where $X \sim N(0, 1)$ and $Y \sim \chi^2_n$. The ratio

$$\frac{X}{\sqrt{Y/n}} \sim t_n$$

follows a ***t*-distribution** (Student's t-distribution) with n degrees of freedom. The PDF of the t-distribution is given in Eq. (C.8) in Appendix C.3.

Figure 8.7b visualizes the t_1-, t_5-, and t_{30}-distributions. The quantiles of different t-distributions can be obtained in *R* using the `qt(p,df)` command. They are also listed in Table C.2 for different values of n.

An application of the t-distribution is the following: if we draw a sample of size n from a normal population $N(\mu, \sigma^2)$ and calculate the arithmetic mean $\bar{X}$ and the sample variance S^2_X, then the following theorem holds:

Theorem 8.3.2 (Student's theorem) *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *with* $X_i \overset{iid.}{\sim} N(\mu, \sigma^2)$. *The ratio*

$$\frac{(\bar{X}-\mu)\sqrt{n}}{S_X} = \frac{(\bar{X}-\mu)\sqrt{n}}{\sqrt{\frac{1}{n-1}\sum_{i+1}^{n}(X_i-\bar{X})^2}} \sim t_{n-1} \tag{8.26}$$

is then t-distributed with $n-1$ degrees of freedom.

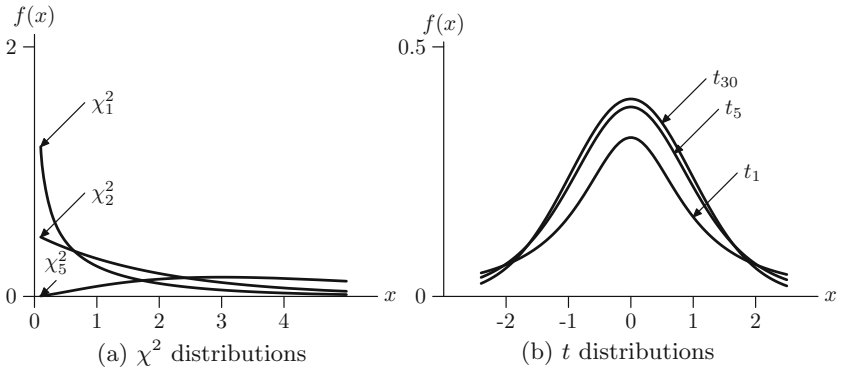


Fig. 8.7 Probability density functions of χ^2 and t distributions*

8.3.3 F-Distribution

Definition 8.3.3 Let X and Y be independent χ^2_m and χ^2_n-distributed random variables, then the distribution of the ratio

$$\frac{X/m}{Y/n} \sim F_{m,n} \tag{8.27}$$

follows the **Fisher F-distribution** with (m, n) degrees of freedom. The PDF of the F-distribution is given in Eq. (C.9) in Appendix C.3.

If X is a χ^2_1-distributed random variable, then the ratio (8.27) is $F_{1,n}$-distributed. The square root of this ratio is t_n-distributed since the square root of a χ^2_1-distributed random variable is $N(0, 1)$-distributed. If W is F-distributed, $F_{m,n}$, then $1/W$ is $F_{n,m}$-distributed. Figure 8.8 visualizes the $F_{5,5}$, $F_{5,10}$ and $F_{5,30}$ distributions. The

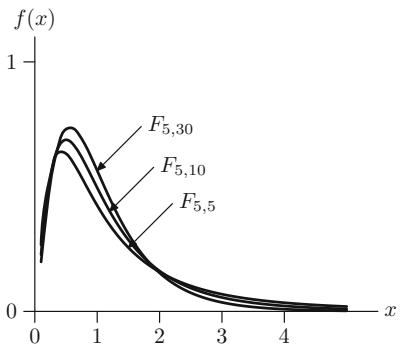


Fig. 8.8 Probability density functions for different F-distributions*

quantiles of different F-distributions can be obtained in *R* using the `qf(p,df1,df2)` command.

One application of the F-distribution relates to the ratio of two sample variances of two independent samples of size m and n, where each sample is an i.i.d. sample from a normal population, i.e. $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$. For the sample variances $S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ from the two populations, the ratio

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1}$$

is F-distributed with $(m-1)$ degrees of freedom in the numerator and $(n-1)$ degrees of freedom in the denominator.

8.4 Key Points and Further Issues

Note:

✓ Examples of different distributions are:

Distribution	Example
Uniform	Rolling a die (discrete)
	Waiting for a train (continuous)
Bernoulli	Any binary variable such as gender
Binomial	Number of "heads" when tossing a coin n times
Poisson	Number of particles emitted by a radioactive source entering a small area in a given time interval
Multinomial	Categorical variables such as "party voted for"
Geometric	Number of raffle tickets until first ticket wins
Hypergeometric	National lotteries; Fisher's test, see p. 428
Normal	Height or weight of women (men)
Exponential	Survival time of a PC
χ^2	Sample variance; χ^2 tests, see p. 235 ff
t	Confidence interval for the mean, see p. 197
F	Tests in the linear model, see p. 272

Note:

✓ One can use *R* to determine values of densities (PDF/PMF), cumulative probability distribution functions (CDF), quantiles of the CDF, and random numbers:

First letter	Function	Further letters	Example
`d`	Density	`distribution name`	`dnorm`
`p`	Probability	`distribution name`	`pnorm`
`q`	Quantiles	`distribution name`	`qnorm`
`r`	Random number	`distribution name`	`rnorm`

We encourage the use of *R* to obtain quantiles of sampling distributions, but Tables C.1–C.3 also list some of them.

✓ In this chapter, we assumed the parameters such as μ, σ, λ, and others to be known. In Chap. 9, we will propose how to estimate these parameters from the data. In Chap. 10, we test statistical hypotheses about these parameters.

✓ For n i.i.d. random variables $X_1, X_2, \ldots, X_n$, the arithmetic mean $\bar{X}$ converges to a $N(\mu, \sigma^2/n)$ distribution as n tends to infinity. See Appendix C.3 as well as Exercise 8.11 for the Theorem of Large Numbers and the Central Limit Theorem, respectively.

8.5 Exercises

Exercise 8.1 A company producing cereals offers a toy in every sixth cereal package in celebration of their 50th anniversary. A father immediately buys 20 packages.

(a) What is the probability of finding 4 toys in the 20 packages?
(b) What is the probability of finding no toy at all?
(c) The packages contain three toys. What is the probability that among the 5 packages that are given to the family's youngest daughter, she finds two toys?

Exercise 8.2 A study on breeding birds collects information such as the length of their eggs (in mm). Assume that the length is normally distributed with $\mu = 42.1$ mm and $\sigma^2 = 20.8^2$. What is the probability of

(a) finding an egg with a length greater than 50 mm?
(b) finding an egg between 30 and 40 mm in length?

Calculate the results both manually and by using *R*.

Exercise 8.3 A dodecahedron is a die with 12 sides. Suppose the numbers on the die are 1–12. Consider the random variable X which describes which number is shown after rolling the die once. What is the distribution of X? Determine $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Exercise 8.4 Felix states that he is able to distinguish a freshly ground coffee blend from an ordinary supermarket coffee. One of his friends asks him to taste 10 cups of coffee and tell him which coffee he has tasted. Suppose that Felix has actually no clue about coffee and simply guesses the brand. What is the probability of at least 8 correct guesses?

Exercise 8.5 An advertising board is illuminated by several hundred bulbs. Some of the bulbs are fused or smashed regularly. If there are more than 5 fused bulbs on a day, the owner of the board replaces them, otherwise not. Consider the following data collected over a month which captures the number of days (n_i) on which i bulbs were broken:

Fused bulbs	0	1	2	3	4	5
n_i	6	8	8	5	2	1

(a) Suggest an appropriate distribution for X: "number of broken bulbs per day".
(b) What is the average number of broken bulbs per day? What is the variance?
(c) Determine the probabilities $P(X = x)$ using the distribution you chose in (a) and using the average number of broken bulbs you calculated in (b). Compare the probabilities with the proportions obtained from the data.
(d) Calculate the probability that at least 6 bulbs are fused, which means they need to be replaced.
(e) Consider the random variable Y: "time until next bulb breaks". What is the distribution of Y?
(f) Calculate and interpret $\mathrm{E}(Y)$.

Exercise 8.6 Marco's company organizes a raffle at an end-of-year function. There are 4000 raffle tickets to be sold, of which 500 win a prize. The price of each ticket is €1.50. The value of the prizes, which are mostly electrical appliances produced by the company, varies between €80 and €250, with an average value of €142.

(a) Marco wants to have a 99 % guarantee of receiving three prizes. How much money does he need to spend? Use R to solve the question.
(b) Use R to plot the function which describes the relationship between the number of tickets bought and the probability of winning at least three prizes.
(c) Given the value of the prizes and the costs of the tickets, is it worth taking part in the raffle?

Exercise 8.7 A country has a ratio between male and female births of 1.05 which means that 51.22 % of babies born are male.

(a) What is the probability for a mother that the first girl is born during the first three births?
(b) What is the probability of getting 2 girls among 4 babies?

Exercise 8.8 A fishermen catches, on average, three fish in an hour. Let Y be a random variable denoting the number of fish caught in one hour and let X be the time interval between catching two fishes. We assume that X follows an exponential distribution.

(a) What is the distribution of Y?
(b) Determine $\mathrm{E}(Y)$ and $\mathrm{E}(X)$.
(c) Calculate $\mathrm{P}(Y = 5)$ and $\mathrm{P}(Y < 1)$.

Exercise 8.9 A restaurant sells three different types of dessert: chocolate, brownies, yogurt with seasonal fruits, and lemon tart. Years of experience have shown that the probabilities with which the desserts are chosen are 0.2, 0.3, and 0.5, respectively.

(a) What is the probability that out of 5 guests, 2 guests choose brownies, 1 guest chooses yogurt, and the remaining 2 guests choose lemon tart?
(b) Suppose two out of the five guests are known to always choose lemon tart. What is the probability of the others choosing lemon tart as well?
(c) Determine the expectation and variance assuming a group of 20 guests.

Exercise 8.10 A reinsurance company works on a premium policy for natural disasters. Based on experience, it is known that $W =$ "number of natural disasters from October to March" (winter) is Poisson distributed with $\lambda_W = 4$. Similarly, the random variable $S =$ "number of natural disasters from April to September" (summer) is Poisson distributed with $\lambda_S = 3$. Determine the probability that there is at least 1 disaster during both summer and winter based on the assumption that the two random variables are independent.

Exercise 8.11 Read Appendix C.3 to learn about the Theorem of Large Numbers and the Central Limit Theorem.

(a) Draw 1000 realizations from a standard normal distribution using *R* and calculate the arithmetic mean. Repeat this process 1000 times. Evaluate the distribution of the arithmetic mean by drawing a kernel density plot and by calculating the mean and variance of it.

(b) Repeat the procedure in (a) with an exponential distribution with $\lambda = 1$. Interpret your findings in the light of the Central Limit Theorem.
(c) Repeat the procedure in (b) using 10,000 rather than 1000 realizations. How do the results change and why?

→ Solutions to all exercises in this chapter can be found on p. 375

*Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg

Part III
Inductive Statistics

Inference

9

9.1 Introduction

The first four chapters of this book illustrated how one can summarize a data set both numerically and graphically. The validity of interpretations made from such a descriptive analysis is valid only for the data set under consideration and cannot necessarily be generalized to other data. However, it is desirable to make conclusions about the entire population of interest and not only about the sample data. In this chapter, we describe the framework of **statistical inference** which allows us to infer from the sample data about the population of interest–at a given, prespecified uncertainty level–and knowledge about the random process generating the data.

Consider an example where the objective is to forecast an election outcome. This requires us to determine the proportion of votes that each of the k participating parties is going to receive, i.e. to calculate or estimate $p_1, p_2, \ldots, p_k$. If it is possible to ask every voter about their party preference, then one can simply calculate the proportions $p_1, p_2, \ldots, p_k$ for each party. However, it is logistically impossible to ask all eligible voters (which form the population in this case) about their preferred party. It seems more realistic to ask only a small fraction of voters and infer from their responses to the responses of the whole population. It is evident that there might be differences in responses between the sample and the population—but the more voters are asked, the closer we are to the population's preference, i.e. the higher the precision of our estimates for $p_1, p_2, \ldots, p_k$ (the meaning of "precision" will become clearer later in this chapter). Also, it is intuitively clear that the sample must be a representative sample of the voters' population to avoid any discrepancy or bias in the forecasting. When we speak of a representative sample, we mean that all the characteristics present in the population are contained in the sample too. There are many ways to get representative random samples. In fact, there is a branch of statistics, called sampling theory, which studies this subject [see, e.g. Groves et al. (2009) or Kauermann and Küchenhoff (2011) for more details]. A simple random sample is one where each voter has an equal probability of being selected in the sample and

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_9

each voter is independently chosen from the same population. In the following, we will assume that all samples are simple random samples. To further formalize the election forecast problem, assume that we are interested in the true proportions which each party receives on the election day. It is practically impossible to make a perfect prediction of these proportions because there are too many voters to interview, and moreover, a voter may possibly make their final decision possibly only when casting the vote and not before. The voter may change his/her opinion at any moment and may differ from what he/she claimed earlier. In statistics, we call these true proportions *parameters of the population*. The task is then to estimate these parameters on the basis of a sample. In the election example, the intuitive estimates for the proportions in the population are the proportions in the sample and we call them *sample estimates*. How to find good and precise estimates are some of the challenges that are addressed by the concept of *statistical inference*. Now, it is possible to describe the election forecast problem in a statistical and operational framework: estimate the parameters of a population by calculating the sample estimates. An important property of every good statistical inference procedure is that it provides not only estimates for the population parameters but also information about the precision of these estimates.

Consider another example in which we would like to study the distribution of weight of children in different age categories and get an understanding of the "normal" weight. Again, it is not possible to measure the weight of all the children of a specific age in the entire population of children in a particular country. Instead, we draw a random sample and use methods of statistical inference to estimate the weight of children in each age group. More specifically, we have several populations in this problem. We could consider all boys of a specific age and all girls of a specific age as two different populations. For example, all 3-year-old boys will form one possible population. Then, a random sample is drawn from this population. It is reasonable to assume that the distribution of the weight of k-year-old boys follows a normal distribution with some unknown parameters μ_{kb} and σ^2_{kb}. Similarly, another population of k-year-old girls is assumed to follow a normal distribution with some unknown parameters μ_{kg} and σ^2_{kg}. The indices kb and kg are used to emphasize that the parameters may vary by age and gender. The task is now to calculate the estimates of the unknown parameters (in the population) of the normal distributions from the samples. Using quantiles, a range of "normal" weights can then be specified, e.g. the interval from the 1 % quantile to the 99 % quantile of the estimated normal distribution or, alternatively, all weights which are not more than twice the standard deviation away from the mean. Children with weights outside this interval may be categorized as underweight or overweight. Note that we make a specific assumption for the distribution class; i.e. we assume a normal distribution for the weights and estimate its parameters. We call this a **parametric** estimation problem because it is based on distributional assumptions. Otherwise, if no distributional assumptions are made, we speak of a **nonparametric** estimation problem.

9.2 Properties of Point Estimators

As we discussed in the introduction, the primary goal in statistical inference is to find a good estimate of (a) population parameter(s). The parameters are associated with the probability distribution which is believed to characterize the population; e.g. μ and σ^2 are the parameters in a normal distribution $N(\mu, \sigma^2)$. If these parameters are known, then one can characterize the entire population. In practice, these parameters are unknown, so the objective is to estimate them. One can attempt to obtain them based on a function of the sample values. But what does this function look like; and if there is more than one such function, then which is the best one? What is the best approach to estimate the population parameters on the basis of a given sample of data? The answer is given by various statistical concepts such as bias, variability, consistency, efficiency, sufficiency, and completeness of the estimates. We are going to introduce them now.

Assume $x = (x_1, x_2 \ldots, x_n)$ are the observations of a random sample from a population of interest. The random sample represents the realized values of a random variable X. It can be said that $x_1, x_2 \ldots, x_n$ are the n observations collected on the random variable X. Any function of random variables is called a **statistic**. For example, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\max(X_1, X_2, \ldots, X_n)$ etc. are functions of $X_1, X_2, \ldots, X_n$, so they are a statistic. It follows that a statistic is also a random variable. Consider a statistic $T(X)$ which is used to estimate a population parameter θ (which may be either a scalar or a vector). We say $T(X)$ is an **estimator** of θ. To indicate that we estimate θ using $T(X)$, we use the "hat" ($\hat{}$) symbol, i.e. we write $\hat{\theta} = T(X)$. When T is calculated from the sample values $x_1, x_2 \ldots, x_n$, we write $T(x)$ and call it an **estimate** of θ. It becomes clear that $T(X)$ is a random variable but $T(x)$ is its observed value (dependent on the actual sample). For example, $T(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator and a statistic, but $T(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is its estimated value from the realized sample values $x_1, x_2, \ldots, x_n$. Since the sample values are realizations from a random variable, each sample leads to a different value of the estimate of the population parameter. The population parameter is assumed to be a fixed value. Parameters can also be assumed to be random, for example in Bayesian statistics, but this is beyond the scope of this book.

9.2.1 Unbiasedness and Efficiency

Definition 9.2.1 An estimator $T(X)$ is called an *unbiased* estimator of θ if

$$\mathrm{E}_\theta(T(X)) = \theta \,. \tag{9.1}$$

The index θ denotes that the expectation is calculated with respect to the distribution whose parameter is θ.

The bias of an estimator $T(X)$ is defined as

$$\text{Bias}_\theta(T(X)) = \text{E}_\theta(T(X)) - \theta \,. \tag{9.2}$$

It follows that an estimator is said to be unbiased if its bias is zero.

Definition 9.2.2 The variance of $T(X)$ is defined as

$$\text{Var}_\theta(T(X)) = \text{E}\left\{[T(X) - \text{E}(T(X))]^2\right\} \,. \tag{9.3}$$

Both bias and variance are measures which characterize the properties of an estimator. In statistical theory, we search for "good" estimators in the sense that the bias and the variance are as small as possible and therefore the accuracy is as high as possible. Readers interested in a practical example may consult Examples 9.2.1 and 9.2.2, or the explanations for Fig. 9.1.

It turns out that we cannot minimize both measures simultaneously as there is always a so-called bias–variance tradeoff. A measure which combines bias and variance into one measure is the mean squared error.

Definition 9.2.3 The mean squared error (MSE) of $T(X)$ is defined as

$$\text{MSE}_\theta(T(X)) = \text{E}\left\{[T(X) - \theta]^2\right\} \,. \tag{9.4}$$

The expression (9.4) can be partitioned into two parts: the variance and the squared bias, i.e.

$$\text{MSE}_\theta(T(X)) = \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 \,. \tag{9.5}$$

This can be proven as follows:

$$\begin{aligned}
\text{MSE}_\theta(T(X)) &= \text{E}[T(X) - \theta]^2 \\
&= \text{E}[(T(X) - \text{E}_\theta(T(X)) + (\text{E}_\theta(T(X) - \theta)]^2 \\
&= \text{E}[T(X) - \text{E}_\theta(T(X))]^2 + [\text{E}_\theta(T(X)) - \theta]^2 \\
&= \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 \,.
\end{aligned}$$

Note that the calculation is based on the result that the cross product term is zero. The mean squared error can be used to compare different biased estimators.

Definition 9.2.4 An estimator $T_1(X)$ is said to be MSE-better than another estimator $T_2(X)$ for estimating θ if

$$\text{MSE}_\theta(T_1(X)) < \text{MSE}_\theta(T_2(X)) \,,$$

where $\theta \in \Theta$ and Θ is the parameter space, i.e. the set of all possible values of θ. Often, Θ is $\mathbb{R}$ or all positive real values $\mathbb{R}_+$. For example, for a normal distribution, $N(\mu, \sigma^2)$, μ can be any real value and σ^2 has to be a number greater than zero.

Unfortunately, we cannot find an MSE-optimal estimator in the sense that an estimator is MSE-better than all other possible estimators for all possible values of θ. This becomes clear if we define the constant estimator $T(x) = c$ (independent of the actual sample): if $\theta = c$, i.e. if the constant value equals the true population parameter we want to estimate, then the MSE of this constant estimator is zero (but it will be greater than zero for all other values of θ, and the bias increases more as we move c far away from the true θ). Usually, we can only find estimators which are locally best (in a certain subset of Θ). This is why classical statistical inference restricts the search for best estimators to the class of unbiased estimators. For unbiased estimators, the MSE is equal to the variance of an estimator. In this context, the following definition is used for comparing two (unbiased) estimators.

Definition 9.2.5 An unbiased estimator $T_1(X)$ is said to be more efficient than another unbiased estimator $T_2(X)$ for estimating θ if

$$\text{Var}_\theta(T_1(X)) \leq \text{Var}_\theta(T_2(X)) \ , \ \forall \theta \in \Theta \ ,$$

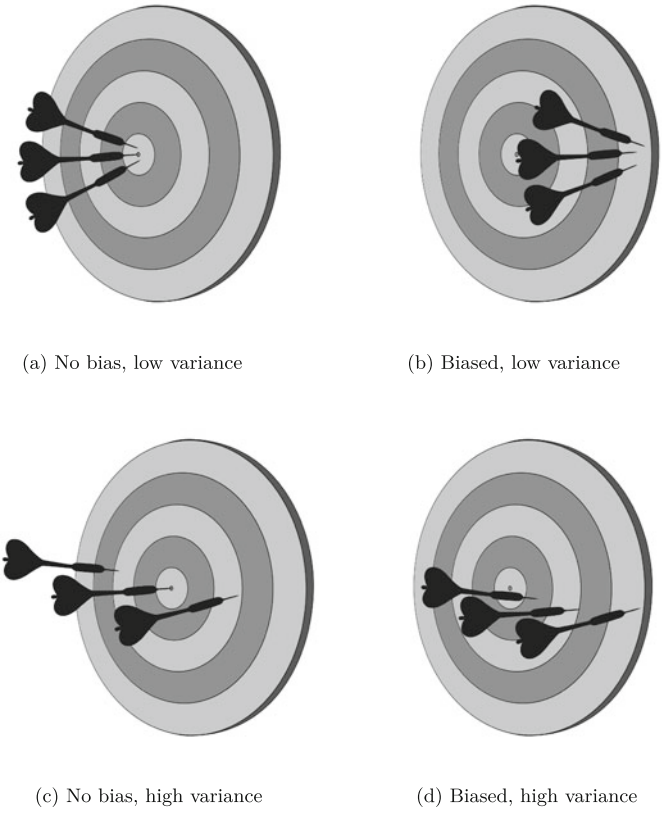
and

$$\text{Var}_\theta(T_1(X)) < \text{Var}_\theta(T_2(X))$$

for at least one $\theta \in \Theta$. It turns out that restricting our search of best estimators to unbiased estimators is sometimes a successful strategy; i.e. for many problems, a best or most efficient estimate can be found. If such an estimator exists, it is said to be UMVU (uniformly minimum variance unbiased). Uniformly means that it has the lowest variance among all other unbiased estimators for estimating the population parameter(s) θ.

Consider the illustration in Fig. 9.1 to better understand the introduced concepts. Suppose we throw three darts at a target and the goal is to hit the centre of the target, i.e. the innermost circle of the dart board. The centre represents the population parameter θ. The three darts play the role of three estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ (based on different realizations of the sample) of the population parameter θ. Four possible situations are illustrated in Fig. 9.1. For example, in Fig. 9.1b, we illustrate the case of an estimator which is biased but has low variance: all three darts are "far" away from the centre of the target, but they are "close" together. If we look at Fig. 9.1a, c, we see that all three darts are symmetrically grouped around the centre of the target, meaning that there is no bias; however, in Fig. 9.1a there is much higher precision than in Fig. 9.1c. It is obvious that Fig. 9.1a presents an ideal situation: an estimator which is unbiased and has minimum variance.

Theorem 9.2.1 *Let $X = (X_1, X_2 \ldots, X_n)$ be an i.i.d. (random) sample of a random variable X with population mean $E(X_i) = \mu$ and population variance $Var(X_i) = \sigma^2$, for all $i = 1, 2, \ldots, n$. Then the arithmetic mean $\bar{X} = \sum_{i=1}^n X_i$ is an unbiased estimator of μ and the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator of σ^2.*



(a) No bias, low variance

(b) Biased, low variance

(c) No bias, high variance

(d) Biased, high variance

Fig. 9.1 Illustration of bias and variance

Note that the theorem holds, in general, for i.i.d. samples, irrespective of the choice of the distribution of the X_i's. Note again that we are looking at the situation *before* we have any observations on X. Therefore, we again use capital letters to denote that the X_i's are random variables which are not known beforehand (i.e. before we actually record the observations on our selected sampling units).

Remark 9.2.1 The empirical variance $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimate of σ^2 and its bias is $-\frac{1}{n}\sigma^2$.

Example 9.2.1 Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables whose population mean is μ and population variance is σ^2. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of μ. This can be shown as follows:

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(7.29)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

The variance of $\bar{X}$ can be calculated as follows:

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)\,, \quad \left[\mathrm{Cov}(X_i, X_j) = 0 \text{ using independence of } X_i\text{'s}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}\,.\end{aligned}$$

We conclude that $\bar{X}$ is an unbiased estimator of μ and its variance is $\frac{\sigma^2}{n}$ irrespective of the choice of the distribution of X. We have learned about the distribution of $\bar{X}$ already in Chap. 8, see also Appendix C.3 for the Theorem of Large Numbers and the Central Limit Theorem; however, we would like to highlight the property of "unbiasedness" in the current context.

Now, we consider another example to illustrate that estimators may not always be unbiased but may have the same variance.

Example 9.2.2 Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables whose population mean is μ and population variance is σ^2. Then $\tilde{X} = \bar{X} + 1 = \frac{1}{n}\sum_{i=1}^{n}(X_i + 1)$ is a biased estimator of μ. This can be shown as follows:

$$\begin{aligned}\mathrm{E}(\tilde{X}) &\overset{(7.31)}{=} \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) + \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}1\right) \\ &\overset{(7.29)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) + \frac{1}{n}\cdot n = \frac{1}{n}\sum_{i=1}^{n}\mu + 1 \\ &= \mu + 1 \neq \mu\,.\end{aligned}$$

However, the variance of $\tilde{X}$ is

$$\mathrm{Var}(\tilde{X}) = \mathrm{Var}\left(\bar{X} + 1\right) \overset{(7.34)}{=} \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\,.$$

If we compare the two estimators $\tilde{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i + 1)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i)$, we see that both have the same variance but the former ($\tilde{X}$) is biased. The efficiency of both estimators is thus the same. It further follows that the mean squared error of $\bar{X}$ is smaller than the mean squared error of $\tilde{X}$ because the MSE consists of the sum of the variance and the squared bias. Therefore $\bar{X}$ is MSE-better than $\tilde{X}$. The comparison of bias, variance and MSE tells us that we should prefer $\bar{X}$ over $\tilde{X}$ when estimating the population mean. This is intuitive, but the argument we make is a purely statistical one.

Theorem 9.2.1 contains the following special cases:

- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a normally distributed population $N(\mu, \sigma^2)$ is an unbiased point estimator of μ.

- The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a normally distributed population $N(\mu, \sigma^2)$ is an unbiased point estimator of σ^2. The sample variance $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator for σ^2, but it is asymptotically unbiased in the sense that its bias tends to zero as the sample size n tends to infinity.
- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli distributed population $B(1, p)$ is an unbiased point estimator of the probability p.

For illustration, we show the validity of the third statement. Let us consider an i.i.d. random sample X_i, $i = 1, 2 \ldots, n$, from a Bernoulli distribution, where $X_i = 1$ if an event occurs and $X_i = 0$ otherwise. Here, p is the probability of occurrence of an event in the population, i.e. $p = P(X_i = 1)$. Note that p is also the population mean: $\mathrm{E}(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p$, $i = 1, 2, \ldots, n$. The arithmetic mean (relative frequency) is an unbiased estimator of p because

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = p,$$

and thus, we can write the estimate of p as

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{9.6}$$

Example 9.2.3 Suppose a random sample of size $n = 20$ of the weight of 10-year-old children in a particular city is drawn. Let us assume that the children's weight in the population follows a normal distribution $N(\mu, \sigma^2)$. The sample provides the following values of weights (in kg):

40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3
34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

To obtain an estimate of the population mean μ, we calculate the arithmetic mean of the observations as

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}(40.2 + 32.8 + \cdots + 34.2) = 41.97,$$

because it is an unbiased estimator of μ. Similarly, we use S^2 to estimate σ^2 because it is unbiased in comparison to $\tilde{S}^2$. Using s_X^2 as an estimate for σ^2 for the given observations, we get

$$\begin{aligned}\hat{\sigma}^2 = s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{19}((40.2 - 41.97)^2 + \cdots + (34.2 - 41.97)^2) \approx 36.85.\end{aligned}$$

The square root of 36.85 is approximately 6.07 which is the standard deviation. Note that the standard deviation based on the sample values divided by the square root of the sample size, i.e. $\hat{\sigma}/\sqrt{20}$, is called the **standard error** of the mean $\bar{X}$ (SEM). As already introduced in Chap. 3, we obtain these results in *R* using the `mean` and `var` commands.

Example 9.2.4 A library draws a random sample of size $n = 100$ members from the members' database to see how many members have to pay a penalty for returning books late, i.e. $x_i = 1$. It turns out that 39 members in the sample have to pay a penalty. Therefore, an unbiased estimator of the population proportion of all members of the library who return books late is

$$\hat{p} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{100} \cdot 39 = \frac{39}{100} = 0.39.$$

Remark 9.2.2 Unbiasedness and efficiency can also be defined asymptotically: we say, for example, that an estimator is asymptotically unbiased, if the bias approaches zero when the sample size tends to infinity. The concept of asymptotic efficiency involves some mathematical knowledge which is beyond the intended scope of this book. Loosely speaking, an asymptotic efficient estimator is an estimator which achieves the lowest possible (asymptotic) variance under given distributional assumptions. The estimators introduced in Sect. 9.3.1, which are based on the maximum likelihood principle, have these properties (under certain mathematically defined regularity conditions).

Next, we illustrate the properties of consistency and sufficiency of an estimator.

9.2.2 Consistency of Estimators

For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated. This property of estimators is referred to as consistency.

Definition 9.2.6 Let $T_1, T_2, \ldots, T_n$, be a sequence of estimators for the parameter θ where $T_n = T_n(X_1, X_2, \ldots, X_n)$ is a function of $X_1, X_2, \ldots, X_n$. The sequence $\{T_n\}$ is a **consistent** sequence of estimators for θ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left[|T_n - \theta| < \epsilon\right] = 1$$

or equivalently

$$\lim_{n\to\infty} P\left[|T_n - \theta| \geq \epsilon\right] = 0.$$

This definition says that as the sample size n increases, the probability that T_n is getting closer to θ is approaching 1. This means that the estimator T_n is getting closer to the parameter θ as n grows larger. Note that there is no information on how *fast* T_n is converging to θ in the sense of convergence defined above.

Example 9.2.5 Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables with expectation μ and variance σ^2. Then for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. For any $\epsilon > 0$, we can write the following:

$$P\left[|\bar{X}_n - \mu| \geq \epsilon\right] = P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right]$$

where $\epsilon = c\sigma/\sqrt{n}$. Using Tschebyschev's inequality (Theorem 7.4.1, p. 139), we get $\frac{1}{c^2} = \sigma^2/n\epsilon^2$, and therefore

$$P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \frac{1}{c^2} = \frac{\sigma^2}{n\epsilon^2}$$

and

$$\lim_{n\to\infty} P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0,$$

provided σ^2 is finite. Hence $\bar{X}_n, n = 1, 2, \ldots,$ converges to μ and therefore $\bar{X}_n$ is a consistent estimator of μ.

Remark 9.2.3 We call this type of consistency *weak* consistency. Another definition is *MSE* consistency, which says that an estimator is MSE consistent if $MSE \longrightarrow 0$ as $n \to \infty$. If the estimator is unbiased, it is sufficient that $\mathrm{Var} \longrightarrow 0$ as $n \to \infty$. If $T_n(X)$ is MSE consistent, it is also weakly consistent. Therefore, it follows that an unbiased estimator with its variance approaching zero as the sample size approaches infinity is both MSE consistent and weakly consistent.

In Example 9.2.5, the variance of $T_n(X) = \bar{X}_n$ is σ^2/n which goes to zero as n goes to ∞ and therefore $\bar{X}_n$ is both weakly consistent and MSE consistent.

9.2.3 Sufficiency of Estimators

Sufficiency is another criterion to judge the quality of an estimator. Before delving deeper into the subject matter, we first try to understand some basic concepts.

Consider two independent random variables X and Y, each following a $N(\mu, 1)$ distribution. We conclude that both X and Y contain information about μ. Consider two estimators of μ as $\hat{\mu}_1 = X + Y$ and $\hat{\mu}_2 = X - Y$. Suppose we want to know whether to use $\hat{\mu}_1$ or $\hat{\mu}_2$ to estimate μ. We notice that $\mathrm{E}(\hat{\mu}_1) = \mathrm{E}(X) + \mathrm{E}(Y) = \mu + \mu = 2\mu$, $\mathrm{E}(\hat{\mu}_2) = \mathrm{E}(X) - \mathrm{E}(Y) = \mu - \mu = 0$, $\mathrm{Var}(\hat{\mu}_1) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2$ and $\mathrm{Var}(\hat{\mu}_2) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2$. Using the additivity property of the normal distribution, which was introduced in Remark 8.2.2, we can say that $\hat{\mu}_1 \sim N(2\mu, 2)$ and $\hat{\mu}_2 \sim N(0, 2)$. So $\hat{\mu}_1$ contains information about μ, whereas $\hat{\mu}_2$

does not contain any information about μ. In other words, $\hat{\mu}_2$ loses the information about μ. We call this property "loss of information".

If we want to make conclusions about μ using both X and Y, we need to acknowledge that the dimension of them is 2. On the other hand, if we use $\hat{\mu}_1$ or equivalently $\hat{\mu}_1/2 \sim N(\mu, \frac{1}{2})$, then we need to concentrate only on one variable and we say that it has dimension 1. It follows that $\hat{\mu}_1$ and $\hat{\mu}_1/2$ provide the same information about μ as provided by the entire sample on both X and Y. So we can say that either $\hat{\mu}_1$ or $\hat{\mu}_1/2$ is sufficient to provide the same information about μ that can be obtained on the basis of the entire sample. This is the idea behind the concept of sufficiency and it results in the reduction of dimension. In general, we can say that if all the information about μ contained in the sample of size n can be obtained, for example, through the sample mean then it is sufficient to use this one-dimensional summary statistic to make inference about μ.

Definition 9.2.7 Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f(x, \theta)$. A statistic T is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ is independent of θ.

The Neyman–Fisher Factorization Theorem provides a practical way to find sufficient statistics.

Theorem 9.2.2 (Neyman–Fisher Factorization Theorem (NFFT)) *Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f(x, \theta)$. A statistic $T = T(x_1, x_2, \ldots, x_n)$ is said to be sufficient for θ if and only if the joint density of $X_1, X_2, \ldots, X_n$ can be factorized as*

$$f(x_1, x_2, \ldots, x_n; \theta) = g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

where $h(x_1, x_2, \ldots, x_n)$ is nonnegative and does not involve θ; and $g(t, \theta)$ is a nonnegative function of θ which depends on $x_1, x_2, \ldots, x_n$ only through t, which is a particular value of T.

This theorem holds for discrete random variables too. Any one-to-one function of a sufficient statistic is also sufficient. A function f is called one-to-one if whenever $f(a) = f(b)$ then $a = b$.

Example 9.2.6 Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 1)$ where μ is unknown. We attempt to find a sufficient statistic for μ. Consider the following function as the joint distribution of $x_1, x_2, \ldots, x_n$ (whose interpretation will become clearer in the next section):

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \mu) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{n\mu^2}{2} + \mu\sum_{i=1}^{n} x_i\right)\exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right). \end{aligned}$$

Here

$$g(t, \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{n\mu^2}{2} + \mu \sum_{i=1}^{n} x_i\right),$$

$$h(x_1, x_2, \ldots, x_n) = \exp\left(-\frac{1}{2} \sum_{i=1}^{n} x_i^2\right),$$

$$t = t(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i.$$

Using the Neyman–Fisher Factorization Theorem, we conclude that $T = T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ. Also, $T = T(X_1, X_2, \ldots, X_n) = \bar{X}$ is sufficient for μ as it is a one-to-one statistic of $\sum_{i=1}^{n} X_i$. On the other hand, $T = \bar{X}^2$ is not sufficient for μ as it is not a one-to-one function of $\sum_{i=1}^{n} X_i$. The important point here is that $\bar{X}$ is a function of the sufficient statistic and hence a good estimator for μ. It is thus summarizing the sample information about the parameter of interest in a complete yet parsimonious way. Another, multivariate, example of sufficiency is given in Appendix C.4.

9.3 Point Estimation

In the previous section, we introduced and discussed various properties of estimators. In this section, we want to show how one can find estimators with good properties. In the general case, properties such as unbiasedness and efficiency cannot be guaranteed for a finite sample. But often, the properties can be shown to hold asymptotically.

9.3.1 Maximum Likelihood Estimation

We have used several estimators throughout the book without stating explicitly that they are estimators. For example, we used the sample mean ($\bar{X}$) to estimate μ in a $N(\mu, \sigma^2)$ distribution; we also used the sample proportion (relative frequency) to estimate p in a $B(1, p)$ distribution, etc. The obvious question is how to obtain a good statistic to estimate an unknown parameter, for example how to determine that the sample mean can be used to estimate μ. We need a general framework for parameter estimation. The method of **maximum likelihood** provides such an approach. For the purpose of illustration, we introduce the method of maximum likelihood estimation with an example using the Bernoulli distribution.

Example 9.3.1 Consider an i.i.d. random sample $X = (X_1, X_2, \ldots, X_n)$ from a Bernoulli population with $p = P(X_i = 1)$ and $(1 - p) = P(X_i = 0)$. The joint probability mass function for a given set of realizations $x_1, x_2, \ldots, x_n$ (i.e. the data) is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|p) = P(X_1 = x_1|p) \cdot \cdots \cdot P(X_n = x_n|p)$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \tag{9.7}$$

This is a function of $(x_1, x_2, \ldots, x_n)$ given the parameter p. The product results from the fact that the draws are independent and the fact that $p^{x_i}(1-p)^{1-x_i} = p$ if $x_i = 1$ and $p^{x_i}(1-p)^{1-x_i} = 1-p$ if $x_i = 0$. That is, the term $p^{x_i}(1-p)^{1-x_i}$ covers results from both possible outcomes. Now, consider a random sample where the values $x = (x_1, x_2, \ldots, x_n)$ are known, for example $x = (0, 1, 0, 0, \ldots, 1)$. Then, (9.7) can be seen as a function of p because $(x_1, x_2, \ldots, x_n)$ is known. In this case, after obtaining a sample of data, the function is called the likelihood function and can be written as

$$L(x_1, x_2, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_1}(1-p)^{1-x_i}\ . \tag{9.8}$$

The joint density function of $X_1, X_2, \ldots, X_n$ is called the **likelihood function**. For better understanding, consider a sample of size 5 with $x = (x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 0)$. The likelihood (function) is

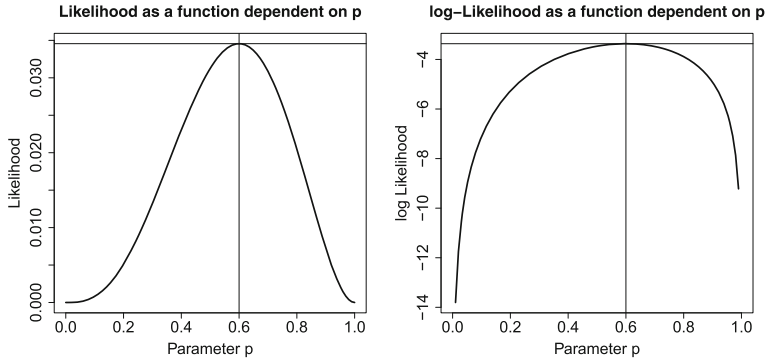
$$L(1, 1, 0, 1, 0|p) = p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3(1-p)^2\ . \tag{9.9}$$

The maximum likelihood estimation principle now says that the estimator $\hat{p}$ of p is the value of p which maximizes the likelihood (9.8) or (9.9). In other words, the maximum likelihood estimate is the value which maximizes the probability of observing the realized sample from the likelihood function. In general, i.e. for any sample, we have to maximize the likelihood function (9.9) with respect to p. We use the well-known principle of maxima–minima to maximize the likelihood function in this case. In principle, any other optimization procedure can also be used, for example numerical algorithms such as the Newton–Raphson algorithm. If the likelihood is differentiable, the first-order condition for the maximum is that the first derivative with respect to p is zero. For maximization, we can transform the likelihood by a strictly monotone increasing function. This guarantees that the potential maximum is taken at the same point as in the original likelihood. A good and highly common choice is the *natural logarithm* since it transforms products in sums and sums are easy to differentiate by differentiating each term in the sum. The log-likelihood in our example is therefore

$$l(1, 1, 0, 1, 0|p) = \ln L(1, 1, 0, 1, 0, |p) = \ln\left\{p^3(1-p)^2\right\} \tag{9.10}$$
$$= 3\ln(p) + 2\ln(1-p) \tag{9.11}$$

where ln denotes the natural logarithm function and we use the rules

$$\ln(a \cdot b) = \ln(a) + \ln(b)\ , a > 0, b > 0$$
$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)\ , a > 0, b > 0$$
$$\ln\left(a^b\right) = b\ln(a)\ , a > 0\ .$$



(a) Likelihood (b) Log-likelihood

Fig. 9.2 Illustration of the likelihood and log-likelihood function of a binomial distribution

Taking the first derivative of (9.10) with respect to p results in

$$\frac{\partial l(1,1,0,1,0|p)}{\partial p} = \frac{3}{p} - \frac{2}{1-p} \,. \tag{9.12}$$

Setting (9.12) to zero and solving for p leads to

$$\begin{aligned} \frac{3}{p} - \frac{2}{1-p} &= 0 \\ \frac{3}{p} &= \frac{2}{1-p} \\ 3(1-p) &= 2p \\ 5p &= 3 \end{aligned}$$

$$\hat{p}_{\text{ML}} = \frac{3}{5} = \frac{1}{5}(1+1+0+1+0) = \bar{x} \,.$$

The value of the second-order partial derivative of (9.9) with respect to p at $p = \hat{p}_{\text{ML}}$ is negative which ensures that $\hat{p}_{\text{ML}}$ maximizes the likelihood function. It follows from this example that the maximum likelihood estimate for p leads to the well-known arithmetic mean. Figure 9.2 shows the likelihood function and the log-likelihood function as functions of p, where $p \in [0, 1]$. The figures show that the likelihood function and the log-likelihood function have the same maxima at $p = 3/5 = 0.6$.

Maximum likelihood estimators have some important properties: they are usually consistent, asymptotically unbiased, asymptotically normally distributed, asymptotically efficient, and sufficient. Even if they are not, a function of a sufficient statistic can always be found which has such properties. This is the reason why maximum likelihood estimation is popular. By "asymptotically" we mean that the properties hold as n tends to infinity, i.e. as the sample size increases. There might be other good estimators in a particular context, for example estimators that are efficient and not only asymptotically efficient; however, in general, the ML principle is a great

choice in many circumstances. We are going to use it in the following sections and chapters, for instance for general point estimation and in the linear regression model (Chap. 11).

Remark 9.3.1 More examples of maximum likelihood estimators are given in Exercises 9.1–9.3.

9.3.2 Method of Moments

The **method of moments** is another well-known method to derive the estimators for population parameters. Below, we outline this principle briefly by way of example.

The idea is that the population parameters of interest can be related to the moments (e.g. expectation, variance) of the distribution of the considered random variables.

A simple case is the estimator for the expected value $\mathrm{E}(X) = \mu$ of a population using an i.i.d. random sample $X = (X_1, \ldots, X_n)$. In this case, $\hat{\mu} = \bar{X}$ is the natural moment estimator of μ. Further, since $\mathrm{E}(X^2) = \sigma^2 + \mu^2$, an estimator of $\sigma^2 + \mu^2$ is $\frac{1}{n}\sum_{i=1}^n X_i^2$. Using $\bar{X}^2$ as an estimator for μ^2, this results in the biased, but asymptotically unbiased estimate

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \,.$$

An extension of this method is the *generalized method of moments* (GMM). GMM estimators have interesting properties: under relatively week conditions (not further discussed here), they are consistent and asymptotically normal, as well as efficient in the class of those estimators that do not use any additional information besides the information included in the moment conditions. Usually, they require a two-step estimating approach or an iterative estimating procedure.

The **least squares estimator** for a linear regression model with i.i.d. random errors, discussed in detail in Chap. 11, can be seen as a special case of a GMM estimator.

9.4 Interval Estimation

9.4.1 Introduction

Let us first consider an example to understand what we mean by interval estimation. Consider a situation in which a lady wants to know the time taken to travel from her home to the train station. Suppose she makes 20 trips and notes down the time taken. To get an estimate of the expected time, one can use the arithmetic mean. Let us say $\bar{x} = 25$ min. This is the point estimate for the expected travelling time. It may not be appropriate to say that she will always take exactly 25 min to reach the train station.

Rather the time may vary by a few minutes each time. To take this into account, the time can be estimated in the form of an interval: it may then be found that the time varies mostly between 20 and 30 min. Such a statement is more informative. Both expectation and variation of the data are taken into account. The interval (20, 30 min) provides a range in which most of the values are expected to lie. We call this concept interval estimation.

A point estimate on its own does not take into account the precision of the estimate. The deviation between the point estimate and the true parameter (e.g. $|\bar{x} - \mu|$) can be substantial, especially when the sample size is small. To incorporate the information about the precision of an estimate in the estimated value, a **confidence interval** can be constructed. It is a **random interval** with **lower and upper bounds**, $I_l(\mathbf{X})$ and $I_u(\mathbf{X})$, such that the unknown parameter θ is covered by a prespecified probability of at least $1 - \alpha$:

$$P_\theta(I_l(\mathbf{X}) \leq \theta \leq I_u(\mathbf{X})) \geq 1 - \alpha. \tag{9.13}$$

The probability $1 - \alpha$ is called the **confidence level** or **confidence coefficient**, $I_l(\mathbf{X})$ is called the **lower confidence bound** or **lower confidence limit** and $I_u(\mathbf{X})$ is called the **upper confidence bound** or **upper confidence limit**. It is important to note that the bounds are random and the parameter is a fixed value. This is the reason why we say that the true parameter is covered by the interval with probability $1 - \alpha$ and **not** that the probability that the interval contains the parameter is $1 - \alpha$. Please note that some software packages use the term "error bar" when referring to confidence intervals.

Frequency interpretation of the confidence interval: Suppose N independent samples $\mathbf{X}^{(j)}$, $j = 1, 2, \ldots, N$, of size n are sampled from the same population and N confidence intervals of the form $[I_l(\mathbf{X}^{(j)}), I_u(\mathbf{X}^{(j)})]$ are calculated. If N is large enough, then on an average $N(1 - \alpha)$ of the intervals (9.13) cover the true parameter.

Example 9.4.1 Let a random variable follow a normal distribution with $\mu = 10$ and $\sigma^2 = 1$. Suppose we draw a sample of $n = 10$ observations repeatedly. The sample will differ in each draw, and hence, the mean and the confidence interval will also differ. The data sets are realizations from random variables. Have a look at Fig. 9.3 which illustrates the mean and the 95 % confidence intervals for 6 random samples. They vary with respect to the mean and the confidence interval width. Most of the means are close to $\mu = 10$, but not all. Similarly, most confidence intervals, but not all, include μ. This is the idea of the frequency interpretation of the confidence interval: different samples will yield different point and interval estimates. Most of the times the interval will cover μ, but not always. The coverage probability is specified by $1 - \alpha$, and the frequency interpretation means that we expect that (approximately) $(1 - \alpha) \cdot 100\,\%$ of the intervals to cover the true parameter μ. In that sense, the location of the interval will give us some idea about where the true but unknown population parameter μ lies, while the length of the interval reflects our uncertainty about μ: the wider the interval is, the higher is our uncertainty about the location of μ.

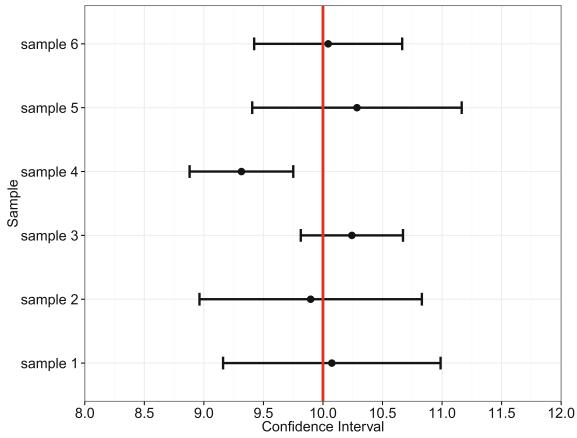


Fig. 9.3 Frequency interpretation of confidence intervals

We now introduce the following confidence intervals:

- Confidence interval for the mean μ of a normal distribution.
- Confidence interval for the probability p of a binomial random variable.
- Confidence interval for the odds ratio.

9.4.2 Confidence Interval for the Mean of a Normal Distribution

Confidence Interval for μ When $\sigma^2 = \sigma_0^2$ is Known.

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a $N(\mu, \sigma_0^2)$ distribution where σ_0^2 is assumed to be known. We use the point estimate $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ and construct a confidence interval around the mean μ. Using the Central Limit Theorem (Appendix C.3, p. 426), it follows that $\bar{X}$ follows a $N(\mu, \sigma_0^2/n)$ distribution. Therefore $\sqrt{n}(\bar{X} - \mu)/\sigma_0 \sim N(0, 1)$, and it follows that

$$P_\mu\left(\left|\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}\right| \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{9.14}$$

where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ quantile of the standard normal distribution $N(0, 1)$. We solve this inequality for the unknown μ and get the desired confidence interval as follows:

$$P_\mu\left[-z_{1-\frac{\alpha}{2}} \leq \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}\right) \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

or

$$P_\mu\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] = 1 - \alpha\,.$$

The confidence interval for μ is thus obtained as

$$[I_l(\mathbf{X}), I_u(\mathbf{X})] = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]. \tag{9.15}$$

This is known as $(1-\alpha)\%$ confidence interval for μ or the confidence interval for μ with confidence coefficient α.

We can use the *R* function `qnorm` or Table C.1 to obtain $z_{1-\frac{\alpha}{2}}$, see also Sects. 8.4, A.3, and C.7. For example, for $\alpha = 0.05$ and $\alpha = 0.01$ we get $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ and $z_{1-\frac{\alpha}{2}} = z_{0.995} = 2.576$ using `qnorm(0.975)` and `qnorm(0.995)`. This gives us the quantiles we need to determine a 95 % and 99 % confidence interval, respectively.

Example 9.4.2 We consider again Example 9.2.3 where we evaluated the weight of 10-year-old children. Assume that the variance is known to be 36; then the upper and lower limits of a 95 % confidence interval for the expected weight μ can be calculated as follows:

$$I_l(X) = \bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} = 41.97 - 1.96\frac{\sqrt{36}}{\sqrt{20}} \approx 39.34\,,$$

$$I_u(X) = \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} = 41.97 + 1.96\frac{\sqrt{36}}{\sqrt{20}} \approx 44.59.$$

We get the confidence interval $[I_u(X), I_o(X)] = [39.34, 44.59]$. With 95 % confidence, the true parameter μ is covered by the interval [39.34, 44.59].

Confidence Interval for μ When σ^2 is Unknown.

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ where σ^2 is assumed to be unknown and is being estimated by the sample variance S_X^2. We know from Sect. 8.3.1 that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}\,.$$

It can be shown that $\bar{X}$ and S_X^2 are stochastically independent. Thus, we know that

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S_X} \sim t_{n-1}$$

follows a t-distribution with $n-1$ degrees of freedom. We can use this result to determine the confidence interval for μ as

$$P_\mu\left[-t_{1-\frac{\alpha}{2},n-1} \le \left(\frac{\sqrt{n}(\bar{X}-\mu)}{S_X}\right) \le t_{1-\frac{\alpha}{2},n-1}\right] = 1-\alpha$$

or

$$P_\mu\left[\bar{X} - t_{1-\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}} \le \mu \le \bar{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}}\right] = 1-\alpha\,.$$

The confidence interval for μ is thus obtained as

$$[I_l(\mathbf{X}), I_u(\mathbf{X})] = \left[\bar{X} - t_{n-1;1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}},\ \bar{X} + t_{n-1;1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}}\right] \tag{9.16}$$

which is the $100(1-\alpha)\%$ confidence interval for μ or the confidence interval for μ with confidence coefficient α.

The interval (9.16) is, in general, wider than the interval (9.15) for identical α and identical sample size n, since the unknown parameter σ^2 is estimated by S_X^2 which induces additional uncertainty. The quantiles for the t-distribution can be obtained using the R command `qt` or Table C.2.

Example 9.4.3 Consider Example 9.4.2 where we evaluated the weight of 10-year-old children. We have already calculated the point estimate of μ as $\bar{x} = 41.97$. With $t_{19;0.975} = 2.093$, obtained via `qt(0.975,19)` or Table C.2, the upper and lower limits of a 95 % confidence interval for μ are obtained as

$$I_u(X) = \bar{x} - t_{19;0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 - 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 39.12\ ,$$
$$I_o(X) = \bar{x} + t_{19;0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 + 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 44.81\ .$$

Therefore, the confidence interval is $[I_l(X), I_u(X)] = [39.13, 44.81]$. In R, we can use the `conf.int` value of the `t.test` command to get a confidence interval for the mean (see also Example 10.3.3 for more details on `t.test`). The default is a 95 % confidence interval, but it can be changed easily if desired:

```
x <- c(40.2, 32.8, 38.2, 43.5, ..., 49.9, 34.2)
t.test(x,conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
```

There is no unique best way to draw the calculated confidence intervals in R. Among many other options, one can simply work with the `plot` functionality or use `geom_errorbar` in conjunction with a `ggplot` object created with the library `ggplot2`, or use the `plotCI` command in the library `plotrix`.

9.4.3 Confidence Interval for a Binomial Probability

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Bernoulli distribution $B(1, p)$. Then $Y = \sum_{i=1}^{n} X_i$ has a binomial distribution $B(n, p)$.

We have already introduced $\hat{p}$ as an estimator for p:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} Y.$$

From (8.8), we know that $\mathrm{Var}(Y) = np(1-p)$. Applying rule (7.33), the variance of the estimator $\hat{p}$ is

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

and it can be estimated by

$$S^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n}.$$

Nowadays, the exact confidence intervals of the binomial distribution function can be easily calculated using computer implementations. Nevertheless, (i) for a sufficiently large sample size n, (ii) if p is not extremely low or high, and (iii) if the condition $np(1-p) \geq 9$ is fulfilled, we can use an approximation based on the normal distribution to calculate confidence intervals. To be more specific, one can show that

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{approx.}{\sim} N(0,1). \tag{9.17}$$

This gives us

$$P\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \approx 1-\alpha, \tag{9.18}$$

and we get a confidence interval for p as

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \tag{9.19}$$

Example 9.4.4 We look again at Example 9.2.4 where we evaluated the proportion of members who had to pay a penalty. Out of all borrowers, 39 % brought back their books late and thus had to pay a fee. A 95 % confidence interval for the probability p of bringing back a book late can be constructed using the normal approximation, since $n\hat{p}(1-\hat{p}) = 100 \cdot 0.39 \cdot 0.61 = 23.79 > 9$. With $z_{1-\alpha/2} = z_{0.975} = 1.96$ and $\hat{p} = 0.39$, we get the 95 % confidence interval as

$$\left[0.39 - 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}},\ 0.39 + 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}}\right] = [0.294, 0.486].$$

In *R*, an exact confidence interval can be found using the function `binom.test`:

```
binom.test(x=39,n=100)$conf.int
[1] 0.2940104 0.4926855
```

One can see that the exact and approximate confidence limits differ slightly due to the normal approximation which approximates the exact binomial probabilities.

9.4.4 Confidence Interval for the Odds Ratio

In Chap. 4, we introduced the odds ratio to determine the strength of association between two binary variables. One may be interested in the dispersion of the odds ratio and hence calculate a confidence interval for it. Recall the notation for 2×2 contingency tables:

		Y		Total (row)
		y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
Total (column)		$a+c$	$b+d$	n

In the spirit of the preceding sections, we can interpret the entries in this contingency table as population parameters. For example, a describes the absolute frequency of observations in the population for which $Y = y_1$ and $X = x_1$. If we have a sample then we can estimate a by the number of *observed* observations n_{11} for which $Y = y_1$ and $X = x_1$. We can thus view n_{11} to be an estimator for a, n_{12} to be an estimator for b, n_{21} to be an estimator for c, and n_{22} to be an estimator for d. It follows that

$$\widehat{\text{OR}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{9.20}$$

serves as the point estimate for the population odds ratio $\text{OR} = ad/bc$. To construct a confidence interval for the odds ratio, we need to work on a log-scale. The log odds ratio,

$$\theta_0 = \ln \text{ OR} = \ln a - \ln b - \ln c + \ln d, \tag{9.21}$$

takes the natural logarithm of the odds ratio. It is evident that it can be estimated using the observed absolute frequencies of the joint frequency distribution of X and Y:

$$\hat{\theta}_0 = \ln \widehat{\text{OR}} = \ln \frac{n_{11}n_{22}}{n_{12}n_{21}}\,. \tag{9.22}$$

It can be shown that $\hat{\theta}_0$ follows approximately a normal distribution with expectation θ_0 and standard deviation

$$\hat{\sigma}_{\hat{\theta}_0} = \left(\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}\right)^{\frac{1}{2}}\,. \tag{9.23}$$

Following the reasoning explained in the earlier section on confidence intervals for binomial probabilities, we can calculate the $100(1-\alpha)\%$ confidence interval for θ_0 under a normal approximation as follows:

$$\left[\hat{\theta}_0 - z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0}, \hat{\theta}_0 + z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0}\right] = [I_u, I_o]\,. \tag{9.24}$$

Since we are interested in the confidence interval of the odds ratio, and not the log odds ratio, we need to transform back the lower and upper bound of the confidence interval as

$$\left[\exp(I_u), \exp(I_o)\right] . \tag{9.25}$$

Example 9.4.5 Recall Example 4.2.5 from Chap. 4 where we were interested in the association of smoking with a particular disease. The data is summarized in the following 2×2 contingency table:

		Smoking		Total (row)
		Yes	No	
Disease	Yes	34	66	100
	No	22	118	140
Total (column)		56	184	240

The odds ratio was estimated to be 2.76, and we therefore concluded that the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. To calculate a 95 % confidence intervals, we need $\hat{\theta}_0 = \ln(2.76)$, $z_{1-\frac{\alpha}{2}} \approx 1.96$ and

$$\begin{aligned}\hat{\sigma}_{\hat{\theta}_0} &= \left(\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{34} + \frac{1}{118} + \frac{1}{66} + \frac{1}{22}\right)^{\frac{1}{2}} \approx 0.314.\end{aligned}$$

The confidence interval for the log odds ratio is

$$[\ln(2.76) - 1.96 \cdot 0.314, \ln(2.76) + 1.96 \cdot 0.314] \approx [0.40, 1.63] .$$

Exponentiation of the confidence interval bounds yields the 95 % confidence interval for the odds ratio as

$$[1.49, 5.11] .$$

There are many ways to obtain the same results in *R*. One option is to use the `oddsratio` function of the library `epitools`. Note that we need to specify "wald" under the methods option to get confidence intervals which use the normal approximation as we did in this case.

```
library(epitools)
smd <- matrix(c(34,22,66,118),ncol=2,nrow=2)  #data
oddsratio(smd,method='wald')
```

9.5 Sample Size Determinations

Confidence intervals help us estimating the precision of point estimates. What if we are required to adhere to a prespecified precision level? We know that the variance decreases as the sample size increases. In turn, confidence intervals become narrower. On the other hand, increasing the sample size has its own consequences. For example, the cost and time involved in setting up experiments, or conducting a survey, increases. In these situations it is important to find a balance between the variability of the estimates and the sample size. We cannot control the variability in the data in most of the situations, but it is possible to control the sample size and therefore the precision of our estimates. For example, we can control the number of people to be interviewed in a survey—given the resources which are available. We discuss how to determine the number of observations needed to get a particular precision (length) of the confidence interval. We find the answers to such questions using the formulae for confidence intervals.

Sample Size Calculation for μ.

Let us consider the situation where we are interested in estimating the population mean μ. The length of the confidence interval (9.15) for the point estimate $\bar{X}$ is

$$2z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}. \tag{9.26}$$

We would now like to fix the width of the confidence interval and come up with a sample size which is required to achieve this width. Let us fix the length of the confidence interval as

$$\Delta = 2z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}. \tag{9.27}$$

Assume we have knowledge of σ_0. The knowledge about σ_0 can be obtained, for example, through a pilot study or past experience with the experiment. We are interested in obtaining the value of n for which a confidence interval has a fixed confidence width of Δ or less. Rearranging (9.27) gives us

$$n \geq \left[2\frac{z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2. \tag{9.28}$$

This means a minimum or optimum sample size is

$$n_{opt} = \left[2\frac{z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2. \tag{9.29}$$

The sample size n_{opt} ensures that the $1-\alpha$ confidence interval for μ has at most length Δ. But note that we have assumed that σ_0 is known. If we do not know σ_0 (which is more likely in practice), we have to make an assumption about it, e.g. by using an estimate from a former study, a pilot study, or other external information. Practically, (9.28) is used in the case of known and unknown σ_0^2.

Example 9.5.1 A call centre is interested in determining the expected length of a telephone call as precisely as possible. The requirements are that the 95 % confidence interval for μ should have a width of 1 min. Suppose that the call centre has developed a pilot study in which σ_0 was estimated to be 5 min. The sample size n that is needed to estimate the expected length of the phone calls with the desired precision is:

$$n \geq \left[\frac{2z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2 = \left[\frac{2 \times 1.96 \times 5}{1}\right]^2 \approx 384.$$

This means that at least 384 calls are required to get the desired confidence interval width.

Sample Size Calculation for p.

We can follow the earlier reasoning and determine the optimum sample size for a specific confidence interval width using the confidence interval definition (9.19). Since the width of the confidence interval is

$$\Delta = 2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

we get

$$n \geq \left[2\frac{z_{1-\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p}). \tag{9.30}$$

Example 9.5.2 A factory may be interested in the probability of an error in an operating process. The length of the confidence interval should be ± 2 %, i.e. $\Delta = 0.04$. Suppose it is speculated that the error probability is 10 %; we may then use $\hat{p} = 0.1$ as our prior judgment for the true value of p. This yields

$$n \geq \left[2\frac{z_{1-\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p}) = \left[2 \times \frac{1.96}{0.04}\right]^2 0.1 \cdot (1-0.1) \approx 865. \tag{9.31}$$

This means we need a sample size of at least 865 to obtain the desired width of the confidence interval for p.

The above examples for both μ and p have shown us that without external knowledge about the research question of interest, it is difficult to come up with an appropriate sample size. Results may vary considerably depending on what type of information is assumed to be known. With limited knowledge, it can be useful to report results for different widths of confidence intervals and hypothesized values of p or σ_0.

Sample size calculations can be highly complex in many practical situations and may not remain as simple as in the examples considered here. For example, Chap. 10 uses additional concepts in the context of hypothesis testing, such as the power, which can be taken into consideration when estimating sample sizes. However,

in this case, calculations and interpretations become more difficult and complex. A detailed overview of sample size calculations can be found in Chow et al. (2007) and Bock (1997).

9.6 Key Points and Further Issues

Note:

✓ We have introduced important point estimates for the parameters of a normal and a binomial distribution:

$$\bar{x} \text{ for } \mu, \qquad S^2 \text{ for } \sigma^2, \qquad \bar{x} \text{ for } p.$$

In general, the choice of these point estimates is not arbitrary but follows some principles of statistical inference such as maximum likelihood estimation, or least squares estimation (introduced in Chap. 11).

✓ The maximum likelihood estimator is usually consistent, asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient.

✓ The validity of all results in this chapter depends on the assumption that the data is complete and has no missing values. Incomplete data may yield different conclusions.

✓ A confidence interval is defined in terms of upper and lower confidence limits and covers the true target parameter with probability $1 - \alpha$. Confidence intervals are often constructed as follows:

$$\text{point estimate} \pm \text{quantile} \cdot \underbrace{\sqrt{\text{variance of point estimate}}}_{\text{standard error}} .$$

✓ More detailed introductions to inference are presented in Casella and Berger (2002) and Young and Smith (2005).

9.7 Exercises

Exercise 9.1 Consider an i.i.d. sample of size n from a Po(λ) distributed random variable X.

(a) Determine the maximum likelihood estimate for λ.
(b) What does the log-likelihood function look like for the following realizations: $x_1 = 4, x_2 = 3, x_3 = 8, x_4 = 6, x_5 = 6$? Plot the function using *R*. Hint: The `curve` command can be used to plot functions.
(c) Use the Neyman–Fisher Factorization Theorem to argue that the maximum likelihood estimate obtained in (a) is a sufficient statistic for λ.

Exercise 9.2 Consider an i.i.d. sample of size n from a $N(\mu, \sigma^2)$ distributed random variable X.

(a) Determine the maximum likelihood estimator for μ under the assumption that $\sigma^2 = 1$.
(b) Now determine the maximum likelihood estimator for μ for an arbitrary σ^2.
(c) What is the maximum likelihood estimate for σ^2?

Exercise 9.3 Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random variables which follow a uniform distribution, $U(0, \theta)$. Write down the likelihood function and argue, without differentiating the function, what the maximum likelihood estimate of θ is.

Exercise 9.4 Let $X_1, X_2, \ldots, X_n$ be n i.i.d. random variables which follow an exponential distribution. An intelligent statistician proposes to use the following two estimators to estimate $\mu = 1/\lambda$:

(i) $T_n(X) = nX_{\min}$ with $X_{\min} = \min(X_1, \ldots, X_n)$ and $X_{\min} \sim \text{Exp}(n\lambda)$,
(ii) $V_n(X) = n^{-1} \sum_{i=1}^{n} X_i$.

(a) Are both $T_n(X)$ and $V_n(X)$ (asymptotically) unbiased for μ?
(b) Calculate the mean squared error of both estimators. Which estimator is more efficient?
(c) Is $V_n(X)$ MSE consistent, weakly consistent, both, or not consistent at all?

Exercise 9.5 A national park in Namibia determines the weight (in kg) of a sample of common eland antelopes:

450 730 700 600 620 660 850 520 490 670 700 820
910 770 760 620 550 520 590 490 620 660 940 790

Calculate

(a) the point estimate of μ and σ^2 and
(b) the confidence interval for μ ($\alpha = 0.05$).

under the assumption that the weight is normally distributed.

(c) Use *R* to reproduce the results from (b).

Exercise 9.6 We are interested in the heights of the players of the two basketball teams "Brose Baskets Bamberg" and "Bayer Giants Leverkusen" as well as the football team "SV Werder Bremen". The following summary statistics are given:

	N	Minimum	Maximum	Mean	Std. dev.
Bamberg	16	185	211	199.06	7.047
Leverkusen	14	175	210	196.00	9.782
Bremen	23	178	195	187.52	5.239

Calculate a 95 % confidence interval for μ for all three teams and interpret the results.

Exercise 9.7 A married couple tosses a coin after each dinner to determine who has to wash the dishes. If the coin shows "head", then the husband has to wash the dishes, and if the coin shows "tails", then the wife has to wash the dishes. After 98 dinners, the wife notes that the coin has shown head 59 times.

(a) Estimate the probability that the wife has to wash the dishes.
(b) Calculate and interpret the 95 % confidence interval for p.
(c) How many dinners are needed to estimate the true probability for the coin showing "head" with a precision of ± 0.5 % under the assumption that the coin is fair?

Exercise 9.8 Suppose 93 out of 104 pupils have passed the final examination at a certain school.

(a) Calculate a 95 % confidence interval for the probability of failing the examination both by manual calculations and by using R, and compare the results.
(b) At county level 3.2 % of pupils failed the examination. Are the school's pupils worse than those in the whole county?

Exercise 9.9 To estimate the audience rate for several TV stations, 3000 households are asked to allow a device, which records which TV station is watched, to be installed on their TVs. 2500 agreed to participate. Assume it is of interest to estimate the probability of someone switching on the TV and watching the show "Germany's next top model".

(a) What is the precision with which the probability can be estimated?
(b) What source of bias could potentially influence the estimates?

Exercise 9.10 An Olympic decathlon athlete is interested in his performance compared with the performance of other athletes. He is a good runner and interested in his 100 m results compared with those of other athletes.

(a) He uses the decathlon data from this book (Appendix A.2) to come up with $\hat{\sigma} = s = 0.233$. What sample size does he need to calculate a 95 % confidence interval for the mean running time which is precise to ± 0.1 s?
(b) Calculate a 95 % confidence interval for the mean running time ($\bar{x} = 10.93$) of the 30 athletes captured in the data set in Chap. A.2. Interpret the width of this interval compared with the width determined in a).
(c) The runner's own best time is 10.86 s. He wants to be among the best 10 % of all athletes. Calculate an appropriate confidence interval to compare his time with the 10 % best times.

Exercise 9.11 Consider the pizza delivery data described in Chap. A.4. We distinguish between pizzas delivered on time (i.e. in less than 30 min) and not delivered on time (i.e. in more than 30 min). The contingency table for delivery time and operator looks as follows:

	Operator		Total
	Laura	Melissa	
<30 min	163	151	314
≥30 min	475	477	952
Total	638	628	1266

(a) Calculate and interpret the odds ratio and its 95 % confidence interval.
(b) Reproduce the results from (a) using *R*.

→ Solutions to all exercises in this chapter can be found on p. 384

Hypothesis Testing

10

10.1 Introduction

We introduced point and interval estimation of parameters in the previous chapter. Sometimes, the research question is less ambitious in the sense that we are not interested in precise estimates of a parameter, but we only want to examine whether a statement about a parameter of interest or the research hypothesis is true or not (although we will see later in this chapter that there is a connection between confidence intervals and statistical tests, called *duality*). Another related issue is that once an analyst estimates the parameters on the basis of a random sample, (s)he would like to infer something about the value of the parameter in the population. Statistical hypothesis tests facilitate the comparison of estimated values with hypothetical values.

Example 10.1.1 As a simple example, consider the case where we want to find out whether the proportion of votes for a party P in an election will exceed 30 % or not. Typically, before the election, we will try to get representative data about the election proportions for different parties (e.g. by telephone interviews) and then make a statement like "yes", we expect that P will get more than 30 % of the votes or "no", we do not have enough evidence that P will get more than 30 % of the votes. In such a case, we will only know after the election whether our statement was right or wrong. Note that the term representative data only means that the sample is similar to the population with respect to the distributions of some key variables, e.g. age, gender, and education. Since we use one sample to compare it with a fixed value (30 %), we call it a **one-sample problem**.

Example 10.1.2 Consider another example in which a clinical study is conducted to compare the effectiveness of a new drug (B) to an established standard drug (A) for a specific disease, for example too high blood pressure. Assume that, as a first step, we want to find out whether the new drug causes a higher reduction in blood

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_10

pressure than the already established older drug. A frequently used study design for this question is a randomized (i.e. patients are randomly allocated to one of the two treatments) controlled clinical trial (double blinded, i.e. neither the patient nor the doctor know which of the drugs a patient is receiving during the trial), conducted in a fixed time interval, say 3 months. A possible hypothesis is that the average change in the blood pressure in group B is higher than in group A, i.e. $\delta_B > \delta_A$ where $\delta_j = \mu_{j0} - \mu_{j3}$, $j = A, B$ and μ_{j0} is the average blood pressure at baseline before measuring the blood pressure again after 3 months (μ_{j3}). Note that we expect both the differences δ_A and δ_B to be positive, since otherwise we would have some doubt that either drug is effective at all. As a second step (after statistically proving our hypothesis), we are interested in whether the improvement of B compared to A is relevant in a medical or biological sense and is valid for the entire population or not. This will lead us again to the estimation problems of the previous chapter, i.e. quantifying an effect using point and interval estimation. Since we are comparing two drugs, we need to have two samples from each of the drugs; hence, we have a **two-sample problem**. Since the patients receiving A are different from those receiving B in this example, we refer to it as a "two-*independent*-samples problem".

Example 10.1.3 In another example, we consider an experiment in which a group of students receives extra mathematical tuition. Their ability to solve mathematical problems is evaluated before and after the extra tuition. We are interested in knowing whether the ability to solve mathematical problems increases after the tuition, or not. Since the same group of students is used in a pre–post experiment, this is called a "two-*dependent*-samples problem" or a "paired data problem".

10.2 Basic Definitions

10.2.1 One- and Two-Sample Problems

In one-sample problems, the data is usually assumed to arise as *one* sample from a defined population. In two-sample problems, the data originates in the form of *two samples* possibly from two different populations. The heterogeneity is often modelled by assuming that the two populations only differ in some parameters or key quantities such as expectation (i.e. mean), median, or variance. As in our introductory example, the samples can either be independent (as in the drug Example 10.1.2) or dependent (as in the evaluation Example 10.1.3).

10.2.2 Hypotheses

A researcher may have a research question for which the truth about the population of interest is unknown. Suppose data can be obtained using a survey, observation, or

an experiment: if, given a prespecified uncertainty level, a statistical test based on the data supports the hypothesis about the population, we say that this hypothesis is statistically proven. Note that the research question has to be operationalized before it can be tested by a statistical test. Consider the drug Example 10.1.2: we want to examine whether the new drug B has a greater blood pressure lowering effect than the standard drug A. We have several options to operationalize this research question into a statistical set-up. One is to test whether the *average* reduction (from baseline to 3 months) of the blood pressure is higher (and positive) for drug B than drug A. We then state our hypotheses in terms of expected values (i.e. μ). Why do we have to use the expected values μ and not simply compare the arithmetic means $\bar{x}$? The reason is that the superiority of B shown in the sample will only be valid for this sample and not necessarily for another sample. We need to show the superiority of B in the entire population, and hence, our hypothesis needs to reflect this. Another option would be, for example, to use median changes in blood pressure values instead of mean changes in blood pressure values. An important point is that the research hypothesis which we want to prove has to be formulated as the statistical alternative hypothesis, often denoted by H_1. The reason for this will become clearer later in this chapter. The opposite of the research hypothesis has to be formulated as the statistical null hypothesis, denoted by H_0. In the drug example, the alternative and null hypotheses are, respectively,

$$H_1 : \delta_B > \delta_A$$

and

$$H_0 : \delta_B \leq \delta_A.$$

We note that the two hypotheses are disjoint and the union of them covers all possible differences of δ_B and δ_A. There is a boundary value ($\delta_B = \delta_A$) which separates the two hypotheses. Since we want to show the superiority of B, the hypothesis was formulated as a one-sided hypothesis. Note that there are different ways to formulate two-sample hypotheses; for example, $H_1 : \delta_B > \delta_A$ is equivalent to $H_1 : \delta_B - \delta_A > 0$. In fact, it is very common to formulate two-sample hypotheses as differences, which we will see later in this chapter.

10.2.3 One- and Two-Sided Tests

We distinguish between one-sided and two-sided hypotheses and tests. In the previous section, we gave an example of a one-sided test.

For an unknown population parameter θ (e.g. μ) and a fixed value θ_0 (e.g. 5), the following three cases have to be distinguished:

Case	Null hypothesis	Alternative hypothesis	
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	Two-sided test problem
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	One-sided test problem
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	One-sided test problem

Example 10.2.1 One-sample problems often test whether a target value is achieved or not. For example, consider the null hypothesis as

- H_0 : average filling weight of packages of flour = 1 kg
- H_0 : average body height (men) = 178 cm.

The alternative hypothesis H_1 is formulated as deviation from the target value. If deviations in both directions are interesting, then H_1 is formulated as a two-sided hypothesis,

- H_1 : average body height (men) $\neq$ 178 cm.

If deviations in a specific direction are the subject of interest, then H_1 is formulated as a one-sided hypothesis, for example,

- H_1 : average filling weight of flour packages is lower than 1 kg.
- H_1 : average filling weight of flour packages is greater than 1 kg.

Two-sample problems often examine differences of two samples. Suppose the null hypothesis H_0 is related to the average weight of flour packages filled by two machines, say 1 and 2. Then, the null hypothesis is

- H_0 : average weight of flour packages filled by machine 1 = average weight of flour packages filled by machine 2.

Then, H_1 can be formulated as a one-sided or two-sided hypothesis. If we want to prove that machine 1 and machine 2 have different filling weights, then H_1 would be formulated as a two-sided hypothesis

- H_1 : average filling weight of machine 1 $\neq$ average filling weight of machine 2.

If we want to prove that machine 1 has lower average filling weight than machine 2, H_1 would be formulated as a one-sided hypothesis

- H_1 : average filling weight of machine 1 $<$ average filling weight of machine 2.

If we want to prove that machine 2 has lower filling weight than machine 1, H_1 would be formulated as a one-sided hypothesis

- H_1 : average filling weight of machine 1 $>$ average filling weight of machine 2.

Remark 10.2.1 Note that we have *not* considered the following situation: $H_0 : \theta \neq \theta_0$, $H_1 : \theta = \theta_0$. In general, with the tests described in this chapter, we cannot prove the equality of a parameter to a predefined value and neither can we prove the equality of two parameters, as in $H_0 : \theta_1 \neq \theta_2$, $H_1 : \theta_1 = \theta_2$. We can, for example,

not prove (statistically) that machines 1 and 2 in the previous example provide equal filling weight. This would lead to the more complex class of equivalence tests, which is a topic beyond the scope of this book.

10.2.4 Type I and Type II Error

If we undertake a statistical test, two types of error can occur.

- The hypothesis H_0 is true but is rejected; this error is called **type I error**.
- The hypothesis H_0 is not rejected although it is wrong; this is called **type II error**.

When a hypothesis is tested, then the following four situations are possible:

	H_0 is true	H_0 is not true
H_0 is not rejected	Correct decision	Type II error
H_0 is rejected	Type I error	Correct decision

The significance level is the probability of type I error, $P(H_1|H_0) = \alpha$, which is the probability of rejecting H_0 (accepting H_1) if H_0 is true. If we construct a test, the significance level α is prespecified, e.g. $\alpha = 0.05$. A significance test is constructed such that the probability of a type I error does not exceed α while the probability of a type II error depends on the true but unknown parameter values in the population(s) and the sample size. Therefore, the two errors are not symmetrically treated in a significance test. In fact, the type II error β, $P(H_0|H_1) = \beta$ is not controlled by the construction of the test and can become very high, sometimes up to $1 - \alpha$. This is the reason why a test not rejecting H_0 is not a (statistical) proof of H_0. In mathematical statistics, one searches for the best test which maintains α and minimizes β. Minimization of both α and β simultaneously is not possible. The reason is that when α increases then β decreases and vice versa. So one of the errors needs to be fixed and the other error is minimized. Consequently, the error which is considered more serious is fixed and then the other error is minimized. The tests discussed in the below sections are obtained based on the assumption that the type I error is more serious than the type II error. So the test statistics are obtained by fixing α and then minimizing β. In fact, the null hypothesis is framed in such a way that it implies that the type I error is more serious than the type II error. The probability $1 - \beta = P(H_1|H_1)$ is called the **power** of the test. It is the probability of making a decision in favour of the research hypothesis H_1, if it is true, i.e. the probability of detecting a correct research hypothesis.

10.2.5 How to Conduct a Statistical Test

In general, we can follow the steps described below to test a hypothesis about a population parameter based on a sample of data.

(1) Define the distributional assumptions for the random variables of interest, and specify them in terms of population parameters (e.g. θ or μ and σ). This is necessary for parametric tests. There are other types of tests, so-called nonparametric tests, where the assumptions can be relaxed in the sense that we do not have to specify a particular distribution, see Sect. 10.6ff. Moreover, for some tests the distributional assumptions can be relaxed if the sample size is large.
(2) Formulate the null hypothesis and the alternative hypothesis as described in Sects. 10.2.2 and 10.2.3.
(3) Fix a significance value (often called type I error) α, for example $\alpha = 0.05$, see also Sect. 10.2.4.
(4) Construct a test statistic $T(\mathbf{X}) = T(X_1, X_2, \ldots, X_n)$. The distribution of T has to be known under the null hypothesis H_0. We note again that $(X_1, X_2, \ldots, X_n)$ refers to the random variables before drawing the actual sample and $x_1, x_2, \ldots, x_n$ are the realized values (observations) in the sample.
(5) Construct a critical region K for the statistic T, i.e. a region where—if T falls in this region—H_0 is rejected, such that

$$P_{H_0}(T(\mathbf{X}) \in K) \leq \alpha \, .$$

The notation $P_{H_0}(\cdot)$ means that this inequality must hold for all parameter values θ that belong to the null hypothesis H_0. Since we assume that we know the distribution of $T(\mathbf{X})$ under H_0, the critical region is defined by those values of $T(\mathbf{X})$ which are unlikely (i.e. with probability of less than α) to be observed under the null hypothesis. Note that although $T(X)$ is a random variable, K is a well-defined region, see Fig. 10.1 for an example.
(6) Calculate $t(x) = T(x_1, x_2, \ldots, x_n)$ based on the realized sample values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.
(7) Decision rule: if $t(x)$ falls into the critical region K, the null hypothesis H_0 is rejected. The alternative hypothesis is then statistically proven. If $t(x)$ falls outside the critical region, H_0 is not rejected.

$$t(x) \in K : H_0 \text{ rejected} \Rightarrow H_1 \text{ is statistically significant,}$$
$$t(x) \notin K : H_0 \text{ not rejected and therefore accepted.}$$

The next two paragraphs show how to arrive at the test decisions from step 7 in a different way. Readers interested in an example of a statistical test may jump to Sect. 10.3.1 and possibly also Example 10.3.1.

10.2.6 Test Decisions Using the p-Value

Statistical software usually does not show us all the steps of hypothesis testing as outlined in Sect. 10.2.5. It is common that instead of calculating and reporting the critical values, the test statistic is printed together with the so-called p-value. It is possible to use the p-value instead of critical regions for making test decisions. The p-value of the test statistic $T(\mathbf{X})$ is defined as follows:

$$\begin{aligned} \text{two-sided case:} \quad & P_{H_0}(|T| \geq t(x)) = p\text{-value} \\ \text{one-sided case:} \quad & P_{H_0}(T \geq t(x)) = p\text{-value} \\ & P_{H_0}(T \leq t(x)) = p\text{-value} \end{aligned}$$

It can be interpreted as the probability of observing results equal to, or more extreme than those actually observed if the null hypothesis was true. Then, the decision rule is

H_0 is rejected if the p-value is smaller than the prespecified significance level α.
Otherwise, H_0 cannot be rejected.

Example 10.2.2 Assume that we are dealing with a two-sided test and assume further that the test statistic $T(x)$ is $N(0, 1)$-distributed under H_0. The significance level is $\alpha = 0.05$. If we observe, for example, $t = 3$, then the p-value is $P_{H_0}(|T| \geq 3)$. This can be calculated in R as

```
2*(1-pnorm(3))
```

because `pnorm()` is used to calculate $P(X \leq x)$, and therefore, `1-pnorm()` can be used to calculate $P(X > x)$. We have to multiply with two because we are dealing with a two-sided hypothesis. The result is $p = 0.002699796$. Therefore, H_0 is rejected. The one-sided p-value is half of the two-sided p-value, i.e. $P(T \geq 3) = P(T \leq 3) = 0.001349898$, and is not necessarily reported by R. It is therefore important to look carefully at the R output when dealing with one-sided hypotheses.

The p-value is sometimes also called the *significance*, although we prefer the term p-value. We use the term *significance* only in the context of a test result: a test is (statistically) significant if (and only if) H_0 can be rejected.

Unfortunately, the p-value is often over-interpreted: both a test and the p-value can only provide a yes/no decision: either H_0 is rejected or not. Interpreting the p-value as the probability that the null hypothesis is true is wrong! It is also incorrect to say that the p-value is the probability of making an error during the test decision. In our (frequentist) context, hypotheses are true or false and no probability is assigned to them. It can also be misleading to speak of "highly significant" results if the p-value is very small. A last remark: the p-value itself is a random variable: under the null hypothesis, it follows a uniform distribution, i.e. $p \sim U(0, 1)$.

10.2.7 Test Decisions Using Confidence Intervals

There is an interesting and useful relationship between confidence intervals and hypothesis tests. If the null hypothesis H_0 is rejected at the significance level α, then there exists a $100(1-\alpha)\,\%$ confidence interval which yields the same conclusion as the test: if the appropriate confidence interval does not contain the value θ_0 targeted in the hypothesis, then H_0 is rejected. We call this **duality**. For example, recall Example 10.1.2 where we were interested in whether the average change in blood pressure for drug B is higher than for drug A, i.e. $H_1 : \delta_B > \delta_A$. This hypothesis is equivalent to $H_1 : \delta_B - \delta_A > \delta_0 = 0$. In the following section, we develop tests to decide whether H_1 is statistically significant or not. Alternatively, we could construct a $100(1-\alpha)\%$ confidence interval for the difference $\delta_B - \delta_A$ and evaluate whether the interval contains $\delta_0 = 0$ or not; if yes, we accept H_0; otherwise, we reject it. For some of the tests introduced in following section, we refer to the confidence intervals which lead to the same results as the corresponding test.

10.3 Parametric Tests for Location Parameters

10.3.1 Test for the Mean When the Variance is Known (One-Sample Gauss Test)

We develop a hypothesis test to test whether the unknown mean (expectation) μ of a $N(\mu, \sigma^2)$-distributed random variable X either differs from a specific value $\mu = \mu_0$ or is smaller (or greater) than μ_0. We assume that the variance $\sigma^2 = \sigma_0^2$ is known. We apply the scheme of Sect. 10.2.5 step by step to develop the test procedure and then give an illustrative example.

1. Distributional assumption: The random variable X follows a $N(\mu, \sigma_0^2)$- distribution with known variance σ_0^2. We assume that an i.i.d. random sample is drawn from $X_1, X_2, \ldots, X_n$ where the X_is follow the same distribution as X, $i = 1, 2, \ldots, n$.

2. Define any of the following set of hypotheses H_0 and H_1:

$$\begin{aligned} H_0 &: \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0, && \text{(two-sided test)} \\ H_0 &: \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0, && \text{(one-sided test)} \\ H_0 &: \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0, && \text{(one-sided test).} \end{aligned}$$

3. Specify the probability of a type I error α: Often $\alpha = 0.05 = 5\,\%$ is chosen.

4. Construct a test statistic: The unknown mean, i.e. the expectation μ, is usually estimated by the sample mean $\bar{x}$. We already know that if the X_is are i.i.d., then the sample mean is normally distributed. Under the assumption that H_0 is true,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{H_0}{\sim} N(\mu_0, \sigma_0^2/n),$$

where $\overset{H_0}{\sim}$ means the "distribution under H_0". If we standardize the mean under H_0, we get a $N(0, 1)$-distributed test statistic

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} \overset{H_0}{\sim} N(0, 1),$$

see also Theorem 7.3.2. Note that $T(\mathbf{X})$ follows a normal distribution even if the X_is are *not* normally distributed and if n is large enough which follows from the Central Limit Theorem (Appendix C.3). One can conclude that the distributional assumption from step 1 is thus particularly important for small samples, but not necessarily important for large samples. As a rule of thumb, $n \geq 30$ is considered to be a large sample. This rule is based on the knowledge that a t-distribution with more than 30 degrees of freedom gets very close to a $N(0, 1)$-distribution.

5. Critical region: Since the test statistic $T(\mathbf{X})$ is $N(0, 1)$-distributed, we get the following critical regions, depending on the hypothesis:

Case	H_0	H_1	Critical region K
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$
(b)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (z_{1-\alpha}, \infty)$
(c)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, z_\alpha = -z_{1-\alpha})$

For case (a) with H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$, we are interested in extreme values of the test statistic on both tails: very small values and very large values of the test statistic give us evidence that H_0 is wrong (because the statistic is mainly driven by the difference of the sample mean and the test value μ_0 for a fixed variance), see Fig. 10.1. In such a two-sided test, when the distribution of the test statistic is symmetric, we divide the critical region into two equal parts and assign each region of size $\alpha/2$ to the left and right tails of the distribution. For $\alpha = 0.05$, 2.5 % of the most extreme values towards the right end of the distribution and 2.5 % of the most extreme values towards the left end of the distribution give us enough evidence that H_0 is wrong and can be rejected and that H_1 is accepted. It is also clear why α is

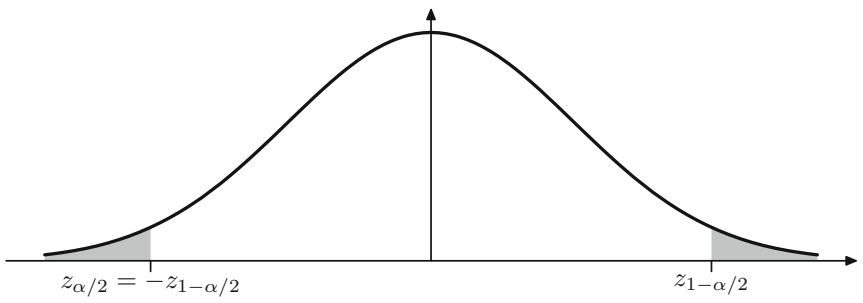


Fig. 10.1 Critical region of a two-sided one-sample Gauss-test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$. The critical region $K = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, \infty)$ has probability mass α if H_0 is true*

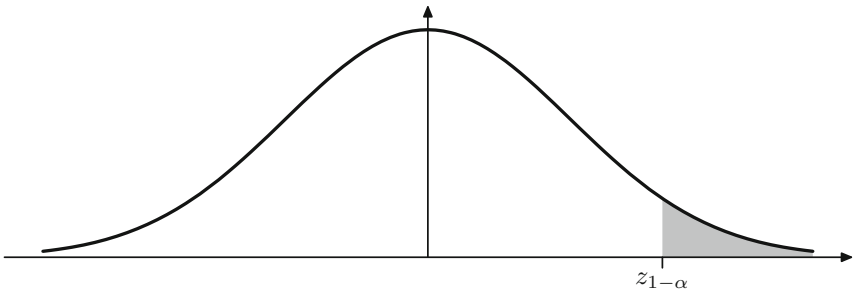


Fig. 10.2 Critical region of a one-sided one-sample Gauss test H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$. The critical region $K = (z_{1-\alpha}, \infty)$ has probability mass α if H_0 is true*

the probability of a type I error: the most extreme values in the two tails together have 5 % probability and are just the probability that the test statistic falls into the critical region although H_0 is true. Also, these areas are those which have the least probability of occurring if H_0 is true. For $\alpha = 0.05$, we get $z_{1-\frac{\alpha}{2}} = 1.96$.

For case (b), only one direction is of interest. The critical region lies on the right tail of the distribution of the test statistic. A very large value of the test statistic has a low probability of occurrence if H_0 is true. An illustration is given in Fig. 10.2: for $\alpha = 0.05$, we get $z_{1-\alpha} = 1.64$ and any values greater than 1.64 are unlikely to be observed under H_0. Analogously, the critical region for case (c) is constructed. Here, the shaded area (critical region) is on the left-hand side. In this case, for $\alpha = 0.05$, we get $z_\alpha = -z_{1-\alpha} = -1.64$.

6. Realization of the test statistic: For an observed sample $x_1, x_2, \ldots, x_n$, the arithmetic mean

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

is used to calculate the realized (observed) test statistic $t(x) = T(x_1, x_2, \ldots, x_n)$ as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0} \sqrt{n}.$$

7. Test decision: If the realized test statistic from step 6 falls into the critical region, H_0 is rejected (and therefore, H_1 is statistically proven). Table 10.1 summarizes the test decisions depending on $t(x)$ and the quantiles defining the appropriate critical regions.

Example 10.3.1 A bakery supplies loaves of bread to supermarkets. The stated selling weight (and therefore the required minimum expected weight) is $\mu = 2$ kg. However, not every package weighs exactly 2 kg because there is variability in the weights. It is therefore important to find out if the average weight of the loaves

Table 10.1 Rules to make test decisions for the one-sample Gauss test (and the two-sample Gauss test, the one-sample approximate binomial test, and the two-sample approximate binomial test—which are all discussed later in this chapter)

Case	H_0	H_1	Reject H_0 if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t(x)\rvert > z_{1-\alpha/2}$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$t(x) < z_\alpha$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$t(x) > z_{1-\alpha}$

is significantly smaller than 2 kg. The weight X (measured in kg) of the loaves is assumed to be normally distributed. We assume that the variance $\sigma_0^2 = 0.1^2$ is known from experience. A supermarket draws a sample of $n = 20$ loaves and weighs them. The average weight is calculated as $\bar{x} = 1.97$ kg. Since the supermarket wants to be sure that the weights are, on average, not lower than 2 kg, a one-sided hypothesis is appropriate and is formulated as H_0: $\mu \geq \mu_0 = 2$ kg versus H_1: $\mu < \mu_0 = 2$ kg. The significance level is specified as $\alpha = 0.05$, and therefore, $z_{1-\alpha} = 1.64$. The test statistic is calculated as

$$t(x) = \frac{\bar{x} - \mu_0}{\sigma_0}\sqrt{n} = \frac{1.97 - 2}{0.1}\sqrt{20} = -1.34.$$

The null hypothesis is not rejected, since $t(x) = -1.34 > -1.64 = -z_{1-0.05} = z_{0.05}$.

Interpretation: The sample average $\bar{x} = 1.97$ kg is below the target value of $\mu = 2$ kg. But there is not enough evidence to reject the hypothesis that the sample comes from a $N(2, 0.1^2)$-distributed population. The probability to observe a sample of size $n = 20$ with an average of at most 1.97 in a $N(2, 0.1^2)$-distributed population is greater than $\alpha = 0.05 = 5\,\%$. The difference between $\bar{x} = 1.97$ kg and the target value $\mu = 2$ kg is not statistically significant.

Remark 10.3.1 The Gauss test assumes the variance to be known, which is often not the case in practice. The t-test (Sect. 10.3.2) assumes that the variance needs to be estimated. The t-test is therefore commonly employed when testing hypotheses about the mean. Its usage is outlined below. In *R*, the command `Gauss.test` from the library `compositions` offers an implementation of the Gauss test.

10.3.2 Test for the Mean When the Variance is Unknown (One-Sample t-Test)

If the variance σ^2 is unknown, hypotheses about the mean μ of a normal random variable $X \sim N(\mu, \sigma^2)$ can be tested in a similar way to the one-sample Gauss test. The difference is that the unknown variance is estimated from the sample. An

unbiased estimator of σ^2 is the sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The test statistic is therefore

$$T(\mathbf{X}) = \frac{\bar{X} - \mu_0}{S_X}\sqrt{n},$$

which follows a t-distribution with $n-1$ degrees of freedom if H_0 is true, as we know from Theorem 8.3.2.

Critical regions and test decisions
Since $T(\mathbf{X})$ follows a t-distribution under H_0, the critical regions refer to the regions of the t-distribution which are unlikely to be observed under H_0:

Case	H_0	H_1	Critical region K
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha/2}) \cup (t_{n-1;1-\alpha/2}, \infty)$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$K = (-\infty, -t_{n-1;1-\alpha})$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$K = (t_{n-1;1-\alpha}, \infty)$

The hypothesis H_0 is rejected if the realized test statistic, i.e.

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n},$$

falls into the critical region. The critical regions are based on the appropriate quantiles of the t-distribution with $(n-1)$ degrees of freedom, as outlined in Table 10.2.

Example 10.3.2 We again consider Example 10.3.1. Now we assume that the variance of the loaves is unknown. Suppose a random sample of size $n = 20$ has an arithmetic mean of $\bar{x} = 1.9668$ and a sample variance of $s^2 = 0.0927^2$. We want to test whether this result contradicts the two-sided hypothesis H_0: $\mu = 2$, that is case (a). The significance level is fixed at $\alpha = 0.05$. For the realized test statistic $t(x)$, we calculate

$$t(x) = \frac{\bar{x} - \mu_0}{s_X}\sqrt{n} = \frac{1.9668 - 2}{0.0927}\sqrt{20} = -1.60.$$

Table 10.2 Rules to make test decisions for the one-sample t-test (and the two-sample t-test, and the paired t-test, both explained below)

Case	H_0	H_1	Reject H_0, if
(a)	$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t(x)\rvert > t_{n-1;1-\alpha/2}$
(b)	$\mu \geq \mu_0$	$\mu < \mu_0$	$t(x) < -t_{n-1;1-\alpha}$
(c)	$\mu \leq \mu_0$	$\mu > \mu_0$	$t(x) > t_{n-1;1-\alpha}$

H_0 is not rejected since $|t| = 1.60 < 2.09 = t_{19;0.975}$, where the quantiles ± 2.09 are defining the critical region (see Table C.2 or use *R*: `qt(0.975,19)`). The same results can be obtained in *R* using the `t.test()` function, see Example 10.3.3 for more details. Or, we can directly calculate the (two-sided) p-value as

```
2*(1-pt(abs(1.6), df=19))
```

This yields a p-value of 0.1260951 which is not smaller than α, and therefore, H_0 is not rejected.

10.3.3 Comparing the Means of Two Independent Samples

In a two-sample problem, we may be interested in comparing the means of two *independent* samples. Assume that we have two samples of two normally distributed variables $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ of size n_1 and n_2, i.e. $X_1, X_2, \ldots, X_{n_1}$ are i.i.d. with the same distribution as X and $Y_1, Y_2, \ldots, Y_{n_2}$ are i.i.d. with the same distribution as Y. We can specify the following hypotheses:

Case	Null hypothesis	Alternative hypothesis	
(a)	$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	Two-sided test problem
(b)	$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	One-sided test problem
(c)	$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	One-sided test problem

We distinguish another three cases:

1. σ_X^2 and σ_Y^2 are known.
2. σ_X^2 and σ_Y^2 are unknown, but they are assumed to be equal, i.e. $\sigma_X^2 = \sigma_Y^2$.
3. Both σ_X^2 and σ_Y^2 are unknown and unequal ($\sigma_X^2 \neq \sigma_Y^2$).

Case 1: The variances are known (two-sample Gauss test).
If the null hypothesis H_0: $\mu_X = \mu_Y$ is true, then, using the usual rules for the normal distribution and the independence of the samples,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_1}\right),$$

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right),$$

and

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_Y - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

It follows that the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \tag{10.1}$$

follows a standard normal distribution, $T(\mathbf{X}, \mathbf{Y}) \sim N(0, 1)$. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}. \tag{10.2}$$

The test procedure is identical to the procedure of the one-sample Gauss test introduced in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

Case 2: The variances are unknown, but equal (two-sample t-test).
We denote the unknown variance of both distributions as σ^2 (i.e. both the populations are assumed to have variance σ^2). We estimate σ^2 by using the pooled sample variance where each sample is assigned weights relative to the sample size:

$$S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}. \tag{10.3}$$

The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \tag{10.4}$$

with S as in (10.3) follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if H_0 is true. The realized test statistic is

$$t(x, y) = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \tag{10.5}$$

The test procedure is identical to the procedure of the one-sample t-test; that is, the test decision is based on Table 10.2.

Case 3: The variances are unknown and unequal (Welch test).
We test H_0: $\mu_X = \mu_Y$ versus H_1: $\mu_X \neq \mu_Y$ given $\sigma_X^2 \neq \sigma_Y^2$ and both σ_X^2 and σ_Y^2 are unknown. This problem is also known as the Behrens–Fisher problem and is the most frequently used test when comparing two means in practice. The test statistic can be written as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}, \tag{10.6}$$

which is approximately t-distributed with v degrees of freedom:

$$v = \left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2 \Bigg/ \left(\frac{\left(s_x^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_y^2/n_2\right)^2}{n_2 - 1}\right) \tag{10.7}$$

where s_x^2 and s_y^2 are the estimated values of $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, respectively. The test procedure, using the observed test statistic

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}, \tag{10.8}$$

is identical to the procedure of the one-sample t-test; that is, the test decision is based on Table 10.2 except that the degrees of freedom are not $n - 1$ but v. If v is not an integer, it can be rounded off to an integer value.

Example 10.3.3 A small bakery sells cookies in packages of 500 g. The cookies are handmade and the packaging is either done by the baker himself or his wife. Some customers conjecture that the wife is more generous than the baker. One customer does an experiment: he buys packages of cookies packed by the baker and his wife on 16 different days and weighs the packages. He gets the following two samples (one for the baker, one for his wife).

Weight (wife) (X)	512	530	498	540	521	528	505	523
Weight (baker) (Y)	499	500	510	495	515	503	490	511

We want to test whether the complaint of the customers is justified. Let us start with the following simple hypotheses:

$$H_0 : \mu_x = \mu_y \quad \text{versus} \quad H_1 : \mu_x \neq \mu_y,$$

i.e. we only want to test whether the weights are different, not that the wife is making heavier cookie packages. Since the variances are unknown, we assume that case 3 is the right choice. We calculate and obtain $\bar{x} = 519.625$, $\bar{y} = 502.875$, $s_X^2 = 192.268$, and $s_Y^2 = 73.554$. The test statistic is:

$$t(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} = \frac{|519.625 - 502.875|}{\sqrt{\frac{192.268}{8} + \frac{73.554}{8}}} \approx 2.91.$$

The degrees of freedom are:

$$v = \left(\frac{192.268}{8} + \frac{73.554}{8}\right)^2 \Big/ \left(\frac{(192.268/8)^2}{7} + \frac{(73.554/8)^2}{7}\right) \approx 11.67 \approx 12.$$

Since $|t(x)| = 2.91 > 2.18 = t_{12;0.975}$, it follows that H_0 is rejected. Therefore, H_1 is statistically significant. This means that the mean weight of the wife's packages is different from the mean weight of the baker's packages. Let us refine the hypothesis and try to find out whether the wife's packages have a higher mean weight. The hypotheses are now:

$$H_0 : \mu_x \leq \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y.$$

The test statistic remains the same but the critical region and the degrees of freedom change. Thus, H_0 is rejected if $t(x, y) > t_{v;1-\alpha}$. Using $t_{v;1-\alpha} = t_{12;0.95} \approx 1.78$ and $t(x, y) = 2.91$, it follows that the null hypothesis can be rejected. The mean weight of the wife's packages is greater than the mean weight of the baker's packages.

In *R*, we would have obtained the same result using the `t.test` command:

```
x <- c(512,530,498,540,521,528,505,523)
y <- c(499,500,510,495,515,503,490,511)
t.test(x,y,alternative='greater')
```

```
        Welch Two-Sample t-test

data:  x and y
t = 2.9058, df = 11.672, p-value = 0.006762
alternative hypothesis: true difference in means is greater
than 0...
```

Note that we have to specify the *alternative* hypothesis under the option `alternative`. The output shows us the test statistic (2.9058), the degrees of freedom (11.672), the alternative hypothesis—but not the decision rule. We know that H_0 is rejected if $t(x, y) > t_{v;1-\alpha}$, so the decision is easy in this case: we simply have to calculate $t_{12;0.95}$ using `qt(0.95,12)` in *R*. A simpler way to arrive at the same decision is to use the p-value. We know that H_0 is rejected if $p < \alpha$ which is the case in this example. It is also worthwhile mentioning that *R* displays the hypotheses slightly differently from ours: our alternative hypothesis is $\mu_x > \mu_y$ which is identical to the statement $\mu_x - \mu_y > 0$, as shown by *R*, see also Sect. 10.2.2.

If we specify `two.sided` as an alternative (which is the default), a confidence interval for the mean *difference* is also part of the output:

```
t.test(x,y,alternative='two.sided')
```

```
...
95 % confidence interval:
  4.151321 29.348679
```

It can be seen that the confidence interval of the difference does not cover the "0". Therefore, the null hypothesis is rejected. This is the duality property referred to earlier in this chapter: the test decision is the same, no matter whether one evaluates (i) the confidence interval, (ii) the test statistic, or (iii) the p-value.

Any kind of t-test can be calculated with the `t.test` command: for example, the two-sample t-test requires to specify the option `var.equal=TRUE` while the Welch test is calculated when the (default) option `var.equal=FALSE` is set. We can also conduct a one-sample t-test. Suppose we are interested in whether the mean

weight of the wife's packages of cookies is greater than 500 g; then, we could test the hypotheses:

$$H_0 : \mu_x \leq 500 \quad \text{versus} \quad H_1 : \mu_x > 500.$$

In *R*, we simply have to specify μ_0:

```
t.test(x,mu=500,alternative='greater')
```

which gives us

```
        One-Sample t-test

data:  x
t = 4.0031, df = 7, p-value = 0.002585
alternative hypothesis: true mean is greater than 500
...
```

10.3.4 Test for Comparing the Means of Two Dependent Samples (Paired *t*-Test)

Suppose there are two dependent continuous random variables X and Y with $\mathrm{E}(X) = \mu_X$ and $\mathrm{E}(Y) = \mu_Y$. They could be dependent because we measure the same variable twice on the same subjects at different times. Typically, this is the case in pre–post experiments, for example when we measure the weight of a person before starting a special diet and after finishing the diet; or when evaluating household expenditures on electronic appliances in two consecutive years. We then say that the samples are *paired*, or dependent. Since the same variable is measured twice on the same subject, it makes sense to calculate a difference between the two respective values. Let $D = X - Y$ denote the random variable "difference of X and Y". If H_0: $\mu_X = \mu_Y$ is true, then the expected difference is zero, and we get $\mathrm{E}(D) = \mu_D = 0$. This means testing $H_0 : \mu_X = \mu_Y$ is identical to testing $\mu_X - \mu_Y = \mu_D = 0$. We further assume that D is normally distributed if H_0: $\mu_X = \mu_Y$ is true (or equivalently if H_0: $\mu_D = 0$ is true), i.e. $D \sim N(0, \sigma_D^2)$. For a random sample $(D_1, D_2, \ldots, D_n)$ of the differences, the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = T(\mathbf{D}) = \frac{\bar{D}}{S_D}\sqrt{n} \tag{10.9}$$

is t-distributed with $n - 1$ degrees of freedom. The sample mean is $\bar{D} = \sum_{i=1}^{n} / D_i n$ and the sample variance is

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

which is an estimator of σ_D^2. The realized test statistic is thus

$$t(d) = \frac{\bar{d}}{s_d}\sqrt{n} \tag{10.10}$$

where $\bar{d} = \sum_{i=1}^{n} d_i/n$ and $s_d^2 = \sum_{i=1}^{n}(d_i - \bar{d})^2/n - 1$.

The two-sided test H_0: $\mu_D = 0$ versus H_1: $\mu_D \neq 0$ and the one-sided tests H_0: $\mu_D \leq 0$ versus H_1: $\mu_D > 0$ or H_0: $\mu_D \geq 0$ versus H_1: $\mu_D < 0$ can be derived as in Sect. 10.3.2; that is, the test decision is based on Table 10.2. In fact, the paired t-test is a one-sample t-test on the differences of X and Y.

Example 10.3.4 In an experiment, $n = 10$ students have to solve different tasks before and after drinking a cup of coffee. Let Y and X denote the random variables "number of points before/after drinking a cup of coffee". Assume that a higher number of points means that the student is performing better. Since the test is repeated on the same students, we have a paired sample. The data is given in the following table:

i	y_i (before)	x_i (after)	$d_i = x_i - y_i$	$(d_i - \bar{d})^2$
1	4	5	1	0
2	3	4	1	0
3	5	6	1	0
4	6	7	1	0
5	7	8	1	0
6	6	7	1	0
7	4	5	1	0
8	7	8	1	0
9	6	5	−1	4
10	2	5	3	4
Total			10	8

We calculate

$$\bar{d} = 1 \quad \text{and} \quad s_d^2 = \frac{8}{9} = 0.943^2,$$

respectively. For the realized test statistic $t(d)$, using $\alpha = 0.05$, we get

$$t(d) = \frac{1}{0.943}\sqrt{10} = 3.35 > t_{9;0.95} = 1.83,$$

such that H_0: $\mu_X \leq \mu_Y$ is rejected and H_1: $\mu_X > \mu_Y$ is accepted. We can conclude (for this example) that drinking coffee significantly increased the problem-solving capacity of the students.

In *R*, we would have obtained the same results using the `t.test` function and specifying the option `paired=TRUE`:

```
yp <- c(4,3,5,6,7,6,4,7,6,2)
xp <- c(5,4,6,7,8,7,5,8,5,5)
t.test(xp,yp,paired=TRUE)
```

```
        Paired t-test

data:  xp and yp
t = 3.3541, df = 9, p-value = 0.008468
alternative hypothesis: true difference in means != 0
95 % confidence interval:
 0.325555 1.674445
sample estimates:
mean of the differences
                      1
```

We can make the test decision using the *R* output in three different ways:

(i) We compare the test statistic ($t = -3.35$) with the critical value (1.83, obtained via `qt(0.95,9)`).
(ii) We evaluate whether the p-value (0.008468) is smaller than the significance level $\alpha = 0.05$.
(iii) We evaluate whether the confidence interval for the mean difference covers "0" or not.

10.4 Parametric Tests for Probabilities

10.4.1 One-Sample Binomial Test for the Probability p

Test construction and hypotheses.
Let X be a Bernoulli $B(1; p)$ random variable with the two possible outcomes 1 and 0, which indicate occurrence and non-occurrence of an event of interest A. The probability for A in the population is p. From the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of independent $B(1; p)$-distributed random variables, we calculate the mean (relative frequency) as $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ which is an unbiased estimate of p. The following hypotheses may thus be of interest:

Case	Null hypothesis	Alternative hypothesis	
(a)	$p = p_0$	$p \neq p_0$	Two-sided problem
(b)	$p \geq p_0$	$p < p_0$	One-sided problem
(c)	$p \leq p_0$	$p > p_0$	One-sided problem

In the following, we describe two possible solutions, one exact approach and an approximate solution. The approximate solution is based on the approximation of the binomial distribution by the normal distribution, which is appropriate if n is sufficiently large and the condition $np(1-p) \geq 9$ holds (i.e. p is neither too small nor too large). First, we present the approximate solution and then the exact one.

Test statistic and test decisions.

(a) **Approximate binomial test**. We define the standardized test statistic as

$$T(\mathbf{X}) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}. \tag{10.11}$$

It holds approximately that $T(\mathbf{X}) \sim N(0,1)$, given that the conditions that (i) n is sufficiently large and (ii) $np(1-p) \geq 9$ are satisfied. The test can then be conducted along the lines of the Gauss test in Sect. 10.3.1; that is, the test decision is based on Table 10.1.

Example 10.4.1 We return to Example 10.1.1. Let us assume that a representative sample of size $n = 2000$ has been drawn from the population of eligible voters, from which 700 (35 %) have voted for the party of interest P. The research hypothesis (which has to be stated as H_1) is that more than 30 % (i.e. $p_0 = 0.3$) of the eligible voters cast their votes for party P. The sample is in favour of H_1 because $\hat{p} = 35\,\%$, but to draw conclusions for the proportion of voters of party P in the population, we have to conduct a binomial test. Since n is large and $np(1-p) = 2000 \cdot 0.35 \cdot 0.65 = 455 \geq 9$, the assumptions for the use of the test statistic (10.11) are satisfied. We can write down the realized test statistic as

$$t(x) = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{0.35 - 0.3}{\sqrt{0.3(1-0.3)}}\sqrt{2000} = 4.8795.$$

Using $\alpha = 0.05$, it follows that $T(X) = 4.8795 > z_{1-\alpha} = 1.64$, and thus, the null hypothesis $H_0 : p \leq 0.3$ can be rejected. Therefore, $H_1 : p > 0.3$ is statistically significant; that is, the proportion of votes for party P is greater than 30 %.

(b) The **exact binomial test** can be constructed using the knowledge that under H_0, $Y = \sum_{i=1}^{n} X_i$ (i.e. the number of successes) follows a binomial distribution. In fact, we can use Y directly as the test statistic:

$$T(\mathbf{X}) = Y \sim B(n, p_0)\ .$$

The observed test statistic is $t(x) = \sum_i x_i$. For the two-sided case (a), the two critical numbers c_l and c_r ($c_l < c_r$) which define the critical region, have to be found such that

$$P_{H_0}(Y \leq c_l) \leq \frac{\alpha}{2} \qquad \text{and} \qquad P_{H_0}(Y \geq c_r) \leq \frac{\alpha}{2}.$$

The null hypothesis is rejected if the test statistic, i.e. Y, is greater than or equal to c_r or less than or equal to c_l. For the one-sided case, a critical number c has to be found such that

$$P_{H_0}(Y \leq c) \leq \alpha$$

for hypotheses of type (b) and

$$P_{H_0}(Y \geq c) \leq \alpha$$

for hypotheses of type (c). If Y is less than the critical value c (for case (b)) or greater than the critical value (for case (c)), the null hypothesis is rejected.

Example 10.4.2 We consider again Example 10.1.1 where we looked at the population of eligible voters, from which 700 (35 %) have voted for the party of interest P. The observed test statistic is $t(x) = \sum_i x_i = 700$ and the alternative hypothesis is $H_1 : p \geq 0.3$, as in case (c). There are at least two ways in which we can obtain the results:

(i) *Long way*: We can calculate the test statistic and compare it to the critical region. To get the critical region, we search c such that

$$P_{p=0.3}(Y \geq c) \leq 0.05\,,$$

which equates to

$$P_{p=0.3}(Y < c) \geq 0.95$$

and can be calculated in *R* as:

```
qbinom(p=0.95, prob=0.3, size=2000)
[1] 634
```

Since $Y = 700 > c = 634$ we reject the null hypothesis. As in Example 10.4.1, we conclude that there is enough evidence that the proportion of votes for party P is greater than 30 %.

(ii) *Short way*: The above result can be easily obtained in *R* using the `binom.test()` command. We need to specify the number of "successes" (here: 700), the number of "failures" ($2000 - 700 = 1300$), and the alternative hypothesis:

```
binom.test(c(700,1300),p=0.3,alternative='greater')
```

```
data:  c(700, 1300)
number of successes = 700, number of trials = 2000,
p-value = 8.395e-07
alternative hypothesis: true probability of success
is greater than 0.3
95 % confidence interval:
 0.332378 1.000000
probability of success
                  0.35
```

Both the p-value (which is smaller than $\alpha = 0.05$) and the confidence interval (for which we do not show the calculation) confirm the rejection of the null hypothesis.
Note that

```
binom.test(x=700,n=2000,p=0.3,
alternative='greater')
```

returns the same result.

10.4.2 Two-Sample Binomial Test

Test construction and hypotheses.
We consider now the case of two independent i.i.d. samples from Bernoulli distributions with parameters p_1 and p_2.

$$\begin{aligned} \mathbf{X} &= (X_1, X_2, \ldots, X_{n_1}), \quad X_i \sim B(1; p_1) \\ \mathbf{Y} &= (Y_1, Y_2, \ldots, Y_{n_2}), \quad Y_i \sim B(1; p_2). \end{aligned}$$

The sums

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2)$$

follow binomial distributions. One of the following hypotheses may be of interest:

Case	Null hypothesis	Alternative hypothesis	
(a)	$p_1 = p_2$	$p_1 \neq p_2$	Two-sided problem
(b)	$p_1 \geq p_2$	$p_1 < p_2$	One-sided problem
(c)	$p_1 \leq p_2$	$p_1 > p_2$	One-sided problem

Similar to the one-sample case, both exact and approximate tests exist. Here, we only present the approximate test. The **exact test of Fisher** is presented in Appendix C.5, p. 428. Let n_1 and n_2 denote the sample sizes. Then, X/n_1 and Y/n_2 are approximately normally distributed:

$$\begin{aligned} \frac{X}{n_1} &\overset{approx.}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \\ \frac{Y}{n_2} &\overset{approx.}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right). \end{aligned}$$

Their difference D

$$D \overset{approx.}{\sim} N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

is normally distributed too under H_0 (given $p = p_1 = p_2$ holds). Since the probabilities p_1 and p_2 are identical under H_0, we can pool the two samples and estimate p by

$$\hat{p} = \frac{X + Y}{n_1 + n_2}\,. \tag{10.12}$$

Test statistic and test decision.
The test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{D}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{10.13}$$

follows a $N(0, 1)$-distribution if n_1 and n_2 are sufficiently large and p is not near the boundaries 0 and 1 (one could use, for example, again the condition $np(1-p) > 9$ with $n = n_1 + n_2$). The realized test statistic can be calculated using the observed difference $\hat{d} = \hat{p}_1 - \hat{p}_2$. The test can be conducted for the one-sided and the two-sided case as the Gauss test introduced in Sect. 10.3.1; that is, the decision rules from Table 10.1 can be applied.

Example 10.4.3 Two competing lotteries claim that every fourth lottery ticket wins. Suppose we want to test whether the probabilities of winning are different for the two lotteries, i.e. $H_0 : p_1 = p_2$ and $H_1 : p_1 \neq p_2$. We have the following data

	n	Winning	Not winning
Lottery A	63	14	49
Lottery B	45	13	32

We can estimate the probabilities of a winning ticket for each lottery, as well as the respective difference, as

$$\hat{p}_A = \frac{14}{63}, \quad \hat{p}_B = \frac{13}{45}, \quad \hat{d} = \hat{p}_A - \hat{p}_B = -\frac{1}{15}.$$

Under H_0, an estimate for p following (10.12) is

$$\hat{p} = \frac{14 + 13}{63 + 45} = \frac{27}{108} = 0.25.$$

The test statistic can be calculated as

$$t(x, y) = \frac{-\frac{1}{15}}{\sqrt{0.25(1-0.25)\left(\frac{1}{63} + \frac{1}{45}\right)}} = -0.79.$$

H_0 is not rejected since $|t(x, y)| = 0.79 < 1.96 = z_{1-0.05/2}$. Thus, there is no statistical evidence for different winning probabilities for the two lotteries. These hypotheses can be tested in *R* using the Test of Fisher, see Appendix C.5, p. 428, for more details.

10.5 Tests for Scale Parameters

There are various tests available to test hypotheses about scale parameters. Such tests are useful when one is interested in the dispersion of a variable, for example in quality control where the variability of a process may be of interest. One-sample tests of hypotheses for the variance of a normal distribution, e.g. hypotheses such as $H_0 : \sigma^2 = \sigma_0^2$, can be tested by the χ^2-test for the variance, see Appendix C.5, p. 430. Two-sample problems can be addressed by the F-test (which is explained in Appendix C.5, p. 431); or by other tests such as the Levene test or Bartlett's test, which are also available in *R* (`leveneTest` in the package `car`, `bartlett` in the base distribution of *R*).

10.6 Wilcoxon–Mann–Whitney (WMW) U-Test

Test construction and hypotheses.

The WMW U-test is often proposed as an alternative to the t-test because it also focuses on location but not on the expected value μ. It is a *nonparametric* test and useful in situations where skewed distributions are compared with each other. We consider two independent random samples $\mathbf{X} = (X_1, X_2, \ldots, X_{n_1})$ and $\mathbf{Y} = (Y_1, Y_2, , \ldots, Y_{n_2})$ from two populations with observed values $(x_1, x_2, \ldots, x_{n_1})$ and $(y_1, y_2, \ldots, y_{n_2})$, respectively. In this case, the null hypothesis H_0 considering the location can be formulated as

$$H_0 : P(X > Y) = P(Y > X) = \frac{1}{2} .$$

The null hypothesis can be interpreted in the following way: the probability that a randomly drawn observation from the first population has a value x that is greater (or lower) than the value y of a randomly drawn subject from the second population is $\frac{1}{2}$. The alternative hypothesis H_1 is then

$$H_1 : P(X > Y) \neq P(Y > X) .$$

This means we are comparing the entire distribution of two variables. If there is a location shift in the sense that one distribution is shifted left (or right) compared with the other distribution, the null hypothesis will be rejected because this shift can be seen as part of the alternative hypothesis $P(X > Y) \neq P(Y > X)$. In fact, under some assumptions, the hypothesis can even be interpreted as comparing two medians, and this is what is often done in practice.

Observed test statistic.

To construct the test statistic, it is necessary to merge $(x_1, x_2, \ldots, x_{n_1})$ and $(y_1, y_2, \ldots, y_{n_2})$ into one sorted sample, usually in ascending order, while keeping the information which value belongs to which sample. For now, we assume that all values of the two samples are distinct; that is, no ties are present. Then, each observation has

a rank between 1 and $(n_1 + n_2)$. Let R_{1+} be the sum of ranks of the x-sample and let R_{2+} be the sum of ranks of the y-sample. The test statistic is defined as U, where U is the minimum of the two values U_1, U_2, $U = \min(U_1, U_2)$ with

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_{1+}, \tag{10.14}$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_{2+}. \tag{10.15}$$

Test decision.
H_0 is rejected if $U < u_{n_1,n_2;\alpha}$. Here, $u_{n_1,n_2;\alpha}$ is the critical value derived from the distribution of U under the null hypothesis. The exact (complex) distribution can, for example, be derived computationally (in R). We are presenting an approximate solution together with its implementation in R.

Since $U_1 + U_2 = n_1 \cdot n_2$, it is sufficient to compute only R_{i+} and $U = \min\{U_i, n_1 n_2 - U_i\}$ ($i = 1$ or $i = 2$ are chosen such that R_{i+} is calculated for the sample with the lower sample size). For $n_1, n_2 \geq 8$, one can use the approximation

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} \overset{approx.}{\sim} N(0,1) \tag{10.16}$$

as the test statistic. For two-sided hypotheses, H_0 is rejected if $|t(x, y)| > z_{1-\alpha/2}$; for one-sided hypotheses H_0 is rejected if $|t(x, y)| > z_{1-\alpha}$. In the case of ties, the denominator of the test statistic in (10.16) can be modified as

$$T(\mathbf{X}, \mathbf{Y}) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\left[\frac{n_1 \cdot n_2}{n(n-1)}\right]\left[\frac{n^3 - n}{12} - \sum_{j=1}^{G} \frac{t_j^3 - t_j}{12}\right]}} \overset{approx.}{\sim} N(0,1),$$

where G is the number of different (groups of) ties and t_j denotes the number of tied ranks in tie group j.

Example 10.6.1 In a study, the reaction times (in seconds) to a stimulus were measured for two groups. One group drank a strong coffee before the stimulus and the other group drank only the same amount of water. There were 9 study participants in the coffee group and 10 participants in the water group. The following reaction times were recorded:

Reaction time	1	2	3	4	5	6	7	8	9	10
Coffee group (C)	3.7	4.9	5.2	6.3	7.4	4.4	5.3	1.7	2.9	
Water group (W)	4.5	5.1	6.2	7.3	8.7	4.2	3.3	8.9	2.6	4.8

We test with the U-test whether there is a location difference between the two groups. First, the ranks of the combined sample are calculated as:

	1	2	3	4	5	6	7	8	9	10	Total
Value (C)	3.7	4.9	5.2	6.3	7.4	4.4	5.3	1.7	2.9		
Rank (C)	5	10	12	15	17	7	13	1	3		83
Value (W)	4.5	5.1	6.2	7.3	8.7	4.2	3.3	8.9	2.6	4.8	
Rank (W)	8	11	14	16	18	6	4	19	2	9	107

With $R_{C+} = 83$ and $R_{W+} = 107$, we get

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - R_{C+} = 9 \cdot 10 + \frac{9 \cdot 10}{2} - 83 = 52,$$
$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_{W+} = 9 \cdot 10 + \frac{10 \cdot 11}{2} - 107 = 38.$$

With $n_1, n_2 \geq 8$ and $U = U_2 = 38$,

$$t(x, y) = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}} = \frac{38 - \frac{9 \cdot 10}{2}}{\sqrt{\frac{9 \cdot 10 \cdot (9 + 10 + 1)}{12}}} \approx -0.572.$$

Since $|t(x, y)| = 0.572 < z_{1-\alpha/2} = 1.96$, the null hypothesis cannot be rejected; that is, there is no statistical evidence that the two groups have different reaction times.

In *R*, one can use the `wilcox.test` command to obtain the results:

```
coffee <- c(3.7, 4.9, 5.2, 6.3, ..., 2.9)
water <- c(4.5, 5.1, 6.2, ..., 4.8)
wilcox.test(coffee, water)
```

The output is

```
    Wilcoxon rank sum test

data:  coffee.sample and water.sample
W = 38, p-value = 0.6038
alternative hypothesis: true location shift is not equal to 0
```

We can see that the null hypothesis is not rejected because $p = 0.6038 > \alpha = 0.05$. The displayed test statistic is W which equates to our statistic U_2. The alternative hypothesis in *R* is framed as location shift, an interpretation which has already been given earlier in the chapter. Note that the test also suggests that the medians of the two samples are not statistically different.

10.7 χ^2-Goodness-of-Fit Test

Test construction.
The χ^2-goodness-of-fit test is one of the most popular tests for testing the goodness of fit of the observed data to a distribution. The construction principle is very general and can be used for variables of any scale. The test statistic is derived such that the *observed* absolute frequencies are compared with the *expected* absolute frequencies *under the null hypothesis* H_0.

Example 10.7.1 Consider an experiment where a die is rolled $n = 60$ times. Under the null hypothesis H_0, we assume that the die is fair, i.e. $p_i = \frac{1}{6}, i = 1, 2, \ldots, 6$, where $p_i = P(X = i)$. We could have also said that H_0 is the hypothesis that the rolls are following a discrete uniform distribution. Thus, the expected absolute frequencies under H_0 are $np_i = 60 \cdot \frac{1}{6} = 10$, while the observed frequencies in the sample are $N_i, i = 1, 2, \ldots, 6$. The N_i generally deviate from np_i. The χ^2-statistic is based on the squared differences, $\sum_{i=1}^{6}(N_i - np_i)^2$, and becomes large as the differences between the observed and the expected frequencies become larger. The χ^2-test statistic is a modification of this sum by scaling each squared difference by the expected frequencies, np_i, and is explained below.

With a nominal variable, we can proceed as in Example 10.7.1. If the scale of the variable is ordinal or continuous, the number of different values can be large. Note that in the most extreme case, we can have as many different values as observations (n), leading to $N_i = 1$ for all $i = 1, 2, \ldots, n$. Then, it is necessary to group the data into k intervals before applying the χ^2-test. The reason is that the general theory of the χ^2-test assumes that the number k (which was 6 in Example 10.7.1 above) is fixed and does not grow with the number of observations n; that is, the theory says that the χ^2-test only works properly if k is fixed and n is large. For this reason, we group the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ into k classes as shown in Sect. 2.1.

Class	1	2	$\cdots$	k	Total
Number of observations	n_1	n_2	$\cdots$	n_k	n

The choice of the class intervals is somewhat arbitrary. As a rule of thumb $np_i > 5$ should hold for most class intervals. The general hypotheses can be formulated in the form of distribution functions:

$$H_0 : F(x) = F_0(x) \text{ versus } H_1 : F(x) \neq F_0(x).$$

Test statistic.
The test statistic is defined as

$$T(\mathbf{X}) = t(x) = \chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}. \tag{10.17}$$

Here,

- N_i $(i = 1, 2, \ldots, k)$ are the absolute frequencies of observations of the sample $\mathbf{X}$ in class i, N_i is a random variable with realization n_i in the observed sample;
- p_i $(i = 1, 2, \ldots, k)$ are calculated from the distribution under H_0, $F_0(x)$, and are the (hypothetical) probabilities that an observation of X falls in class i;
- np_i are the expected absolute frequencies in class i under H_0.

Test decision.
For a significance level α, H_0 is rejected if $t(x)$ is greater than the $(1 - \alpha)$-quantile of the χ^2-distribution with $k - 1 - r$ degrees of freedom, i.e. if

$$t(x) = \chi^2 > c_{k-1-r,1-\alpha}.$$

Note that r is the number of parameters of $F_0(x)$, if these parameters are estimated from the sample. The χ^2-test statistic is only asymptotically χ^2-distributed under H_0.

Example 10.7.2 Let $F_0(x)$ be the distribution function of the test distribution. If one specifies a normal distribution such as $F_0(x) = N(3, 10)$, or a discrete uniform distribution with $p_i = 0.25$ $(i = 1, 2, 3, 4)$, then $r = 0$, since no parameters have to be estimated from the data. Otherwise, if we simply want to test whether the data is generated from a normal distribution $N(\mu, \sigma^2)$ or the data follows a normal distribution $N(\mu, \sigma^2)$, then μ and σ^2 may be estimated from the sample by $\bar{x}$ and s^2. Then, $r = 2$ and the number of degrees of freedom is reduced.

Example 10.7.3 Gregor Mendel (1822–1884) conducted crossing experiments with pea plants of different shape and colour. Let us look at the outcome of a pea crossing experiment with the following results:

Crossing result	Round Yellow	Round Green	Edged Yellow	Edged Green
Observations	315	108	101	32

Mendel had the hypothesis that the four different types occur in proportions of 9:3:3:1, that is

$$p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3 = \frac{3}{16},\ p_4 = \frac{1}{16}.$$

The hypotheses are

$$H_0 : P(X = i) = p_i \quad \text{versus} \quad H_1 : P(X = i) \neq p_i, \qquad i = 1, 2, 3, 4.$$

With $n = 556$ observations, the test statistic can be calculated from the following observed and expected frequencies:

i	N_i	p_i	np_i
1	315	$\frac{9}{16}$	312.75
2	108	$\frac{3}{16}$	104.25
3	101	$\frac{3}{16}$	104.25
4	32	$\frac{1}{16}$	34.75

The χ^2-test statistic is calculated as

$$t(x) = \chi^2 = \frac{(315 - 312.75)^2}{312.75} + \cdots + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$

Since $\chi^2 = 0.47 < 7.815 = \chi^2_{0.95,3} = c_{0.95,3}$, the null hypothesis is not rejected. Statistically, there is no evidence that Mendel was wrong with his 9:3:3:1 assumption. In *R*, the test can be conducted by applying the `chisq.test` command:

```
chisq.test(c(315, 108, 101, 32),
p=c(9/16,3/16,3/16,1/16))
qchisq(df=3, p=0.95)
```

which leads to the following output

```
        Chi-squared test for given probabilities

data:  c(315, 108, 101, 32)
X-squared = 0.47, df = 3, p-value = 0.9254
```

and the critical value is

```
[1] 7.814728
```

Remark 10.7.1 In this example, the data was already summarized in a frequency table. For raw data, the `table` command can be used to preprocess the data, i.e. we can use `chisq.test(table(var1,var2))`.

Another popular goodness-of-fit test is the test of Kolmogorov–Smirnov. There are two different versions of this test, one for the one-sample scenario and one for the two-sample scenario. The null hypothesis for the latter is that the two independent samples come from the same distribution. In *R*, the command `ks.test()` can be used to perform Kolmogorov–Smirnov tests.

10.8 χ^2-Independence Test and Other χ^2-Tests

In Chap. 4, we introduced different methods to describe the association between two variables. Several association measures are possibly suitable if the variables are categorical, for example Cramer's V, Goodman's and Kruskal's γ, Spearman's rank correlation coefficient, and the odds ratio. If we are not interested in the strength of association but rather in finding out whether there is an association at all, one can use the χ^2-independence test.

Test construction.
In the following we assume that we observe a sample from a bivariate discrete distribution of two variables X and Y which can be summarized in a contingency table with absolute frequencies n_{ij}, $(i = 1, 2, \ldots, I;\ j = 1, 2 \ldots, J)$:

		Y				
		1	2	$\cdots$	J	
X	1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
	2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

Remember that

n_{i+} is the ith row sum,
n_{+j} is the jth column sum, and
n is the total number of observations.

The hypotheses are H_0: X and Y are independent versus $H_1 : X$ and Y are not independent. If X and Y are independent, then the expected frequencies m_{ij} are

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = \frac{n_{i+}n_{+j}}{n}. \tag{10.18}$$

Test statistic.
Pearson's χ^2-test statistic was introduced in Chap. 4, Eq. (4.6). It is

$$T(\mathbf{X}, \mathbf{Y}) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}},$$

where $m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ (expected absolute cell frequencies under H_0). Strictly speaking, m_{ij} are the true, unknown expected frequencies under H_0 and are estimated by $\hat{m}_{ij} = n\pi_{i+}\pi_{+j}$, such that the realized test statistic equates to

$$t(x, y) = \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \tag{10.19}$$

Test decision.
The number of degrees of freedom under H_0 is $(I-1)(J-1)$, where $I-1$ are the parameters which have to be estimated for the marginal distribution of X, and $J-1$ are the number of parameters for the marginal distribution of Y. The test decision is:

$$\text{Reject } H_0, \text{ if } t(x,y) = \chi^2 > c_{(I-1)(J-1);1-\alpha}.$$

Note that the alternative hypothesis H_1 is very general. If H_0 is rejected, nothing can be said about the structure of the dependence of X and Y from the χ^2-value itself.

Example 10.8.1 Consider the following contingency table. Here, X describes the educational level (1: primary, 2: secondary, 3: tertiary) and Y the preference for a specific political party (1: Party A, 2: Party B, 3: Party C). Our null hypothesis is that the two variables are independent, and we want to show the alternative hypothesis which says that there is a relationship between them.

		Y			Total
		1	2	3	
X	1	100	200	300	600
	2	100	100	100	300
	3	80	10	10	100
	Total	280	310	410	1000

For the (estimated) expected frequencies $\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$, we get

		Y		
		1	2	3
X	1	168	186	246
	2	84	93	123
	3	28	31	41

For example: $\hat{m}_{11} = 600 \cdot 280/1000 = 168$. The test statistic is

$$\begin{aligned} t(x,y) &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}} \\ &= \frac{(100-168)^2}{168} + \cdots + \frac{(10-41)^2}{41} \approx 182.54. \end{aligned}$$

Since $\chi^2_{4;0.95} = 9.49 < t(x,y) = 182.54$, H_0 is rejected.

In *R*, either the summarized data (as shown below) can be used to calculate the test statistic or the raw data (summarized in a contingency table via `table(var1,var2)`):

```
ct <- matrix(nrow=3,ncol=3,byrow=T,
data=c(100,200,300,100,100,100,80,10,10))
chisq.test(ct)
qchisq(df=(3-1)*(3-1), p=0.95)
```

The output is

```
    Pearson's Chi-squared test

data:  contingency.table
X-squared = 182.5428, df = 4, p-value < 2.2e-16
```

with the critical value

```
[1] 9.487729
```

which confirms our earlier manual calculations. The p-value is smaller than $\alpha = 0.05$ which further confirms that the null hypothesis has to be rejected.

For a binary outcome, the χ^2-test of independence can be formulated as a test for the null hypothesis that the proportions of the binary variable are equal in several (≥ 2) groups, i.e. for a $K \times 2$ (or $2 \times K$) table. This test is called the **χ^2-test of homogeneity**.

Example 10.8.2 Consider two variables X and Y, where X is describing the rating of a coffee brand with the categories "bad taste" and "good taste" and Y denotes three age subgroups, e.g. "18–25", "25–35", and "35–45". The observed data is

		Y			
		18–25	25–35	35–45	Total
X	Bad	10	30	65	105
	Good	90	70	35	195
	Total	100	100	100	300

Assume H_0 is the hypothesis that the probabilities $P(X = \text{'good'}|Y = \text{'18–25'})$, $P(X = \text{'good'}|Y = \text{'25–35'})$, and $P(X = \text{'good'}|Y = \text{'35–45'})$ are all equal. Then, we can use the function either `prop.test` or `chisq.test` in *R* to test this hypothesis:

```
prop.test(x=rbind(c(10,30,65), c(90,70,35) ))
chisq.test(x=rbind(c(10,30,65), c(90,70,35) ))
```

This produces the following outputs:

```
    3-sample test for equality of proportions

data:  cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3
  0.10   0.30   0.65
```

and

```
    Pearson's Chi-squared test

data:  cbind(c(10, 30, 65), c(90, 70, 35))
X-squared = 68.1319, df = 2, p-value = 1.605e-15
```

The results (test statistic, p-value) are identical and H_0 is rejected. Note that `prop.test` strictly expects a $K \times 2$ table (i.e. exactly 2 columns).

Remark 10.8.1 For 2×2-tables with small sample sizes and therefore small cell frequencies, it is recommended to use the exact test of Fisher as described in Appendix C.5.

Remark 10.8.2 The test described in Example 10.8.2 is a special case (since one variable is binary) of the general χ^2-test of homogeneity. The χ^2-test of homogeneity is valid for any $K \times C$ table, where K is the number of subgroups of a variable Y and C is the number of values of the outcome X of interest. The null hypothesis H_0 assumes that the conditional distributions of X given Y are identical in all subgroups, i.e.

$$P(X = x_c|Y = y_k) = P(X = x_c|Y = y_{k'})$$

forall $c = 1, 2, \ldots, C; k, k' = 1, 2, \ldots, K, k \neq k'$. Again, the usual χ^2-test statistic can be used.

10.9 Key Points and Further Issues

Note:

- ✓ A graphical summary on when to use the tests introduced in this chapter is given in Appendices D.2 and D.3.
- ✓ To arrive at a test decision, i.e. accept H_0 or reject it, it does not matter whether one compares the test statistic to the critical region, one uses the p-value obtained from statistical software, or one evaluates the appropriate confidence interval. However, it is important not to misinterpret the p-value (see Sect. 10.2.6) and to choose the correct confidence interval.
- ✓ There is a difference between relevance and significance. A test might be significant, but the point estimate of the quantity of interest may not be relevant from a substantive point of view. Similarly, a test might not be significant, but the point and interval estimates may still yield relevant conclusions.
- ✓ The test statistic of the t-test (one-sample, two-sample, paired) is *asymptotically* normally distributed. This means that for relatively large n (as a rule of thumb >30 per group) the sample does not need to come from a normal distribution. However, the application of the t-test makes sense only when the expectation μ can be interpreted meaningfully; this may not be the case for skewed distributions or distributions with outliers.

10.10 Exercises

Exercise 10.1 Two people, A and B, are suspects for having committed a crime together. Both of them are interrogated in separate rooms. The jail sentence depends on who confesses to have committed the crime, and who does not:

	B does not confess	B does confess
A does not confess	Each serves 1 year	A: 3 years; B: goes free
A does confess	A: goes free; B: 3 years	Each serves 2 years

A has two hypotheses:

$$H_0: \text{ B does not confess} \quad \text{versus} \quad H_1: \text{ B does confess.}$$

Given the possible sentences he decides to not confess if H_0 is true and to confess otherwise. Explain the concepts of type I error and type II error for this situation. Comment on the consequences if these errors are made.

Exercise 10.2 A producer of chocolate bars hypothesizes that his production does not adhere to the weight standard of 100 g. As a measure of quality control, he weighs 15 bars and obtains the following results in grams:

96.40, 97.64, 98.48, 97.67, 100.11, 95.29, 99.80, 98.80, 100.53, 99.41, 97.64, 101.11, 93.43, 96.99, 97.92

It is assumed that the production process is standardized in the sense that the variation is controlled to be $\sigma = 2$.

(a) What are the hypotheses regarding the expected weight μ for a two-sided test?
(b) Which test should be used to test these hypotheses?
(c) Conduct the test that was suggested to be used in (b). Use $\alpha = 0.05$.
(d) The producer wants to show that the expected weight is smaller than 100 g. What are the appropriate hypotheses to use?
(e) Conduct the test for the hypothesis in (d). Again use $\alpha = 0.05$.

Exercise 10.3 Christian decides to purchase the new CD by Bruce Springsteen. His first thought is to buy it online, via an online auction. He discovers that he can also buy the CD immediately, without bidding at an auction, from the same online store. He also looks at the price at an internet book store which was recommended to him by a friend. He notes down the following prices (in €):

Internet book store 16.95

Online store, no auction 18.19, 16.98, 19.97, 16.98, 18.19, 15.99, 13.79, 15.90, 15.90, 15.90, 15.90, 15.90, 19.97, 17.72

Online store, auction 10.50, 12.00, 9.54, 10.55, 11.99, 9.30, 10.59, 10.50, 10.01, 11.89, 11.03, 9.52, 15.49, 11.02

(a) Calculate and interpret the arithmetic mean, variance, standard deviation, and coefficient of variation for the online store, both for the auction and non-auction offers.
(b) Test the hypothesis that the mean price at the online store (no auction) is unequal to €16.95 ($\alpha = 0.05$).
(c) Calculate a confidence interval for the mean price at the online store (no auction) and interpret your findings in the light of the hypothesis in (b).
(d) Test the hypothesis that the mean price at the online store (auction) is less than €16.95 ($\alpha = 0.05$).
(e) Test the hypothesis that the mean non-auction price is higher than the mean auction price. Assume that (i) the variances are equal in both samples and (ii) the variances are unequal ($\alpha = 0.05$).
(f) Test the hypothesis that the variance of the non-auction price is unequal to the variance of the auction price ($\alpha = 0.05$).

(g) Use the U-test to compare the location of the auction and non-auction prices. Compare the results with those of (e).
(h) Calculate the results of (a)–(g) with R.

Exercise 10.4 Ten of Leonard's best friends try a new diet: the "Banting" diet. Each of them weighs him/herself before and after the diet. The data is as follows:

Person (i)	1	2	3	4	5	6	7	8	9	10
Before diet (x_i)	80	95	70	82	71	70	120	105	111	90
After diet (y_i)	78	94	69	83	65	69	118	103	112	88

Choose a test and a confidence interval to test whether there is a difference between the mean weight before and after the diet ($\alpha = 0.05$).

Exercise 10.5 A company producing clothing often finds deficient T-shirts among its production.

(a) The company's controlling department decides that the production is no longer profitable when there are more than 10 % deficient shirts. A sample of 230 shirts yields 30 shirts which contain deficiencies. Use the approximate binomial test to decide whether the T-shirt production is profitable or not ($\alpha = 0.05$).
(b) Test the same hypothesis as in (a) using the exact binomial test. You can use R to determine the quantiles needed for the calculation.
(c) The company is offered a new cutting machine. To test whether the change of machine helps to improve the production quality, 115 sample shirts are evaluated, 7 of which have deficiencies. Use the two-sample binomial test to decide whether the new machine yields improvement or not ($\alpha = 0.05$).
(d) Test the same hypothesis as in (c) using the test of Fisher in R.

Exercise 10.6 Two friends play a computer game and each of them repeats the same level 10 times. The scores obtained are:

	1	2	3	4	5	6	7	8	9	10
Player 1	91	101	112	99	108	88	99	105	111	104
Player 2	261	47	40	29	64	6	87	47	98	351

(a) Player 2 insists that he is the better player and suggests to compare their mean performance. Use an appropriate test ($\alpha = 0.05$) to test this hypothesis.
(b) Player 1 insists that he is the better player. He proposes to not focus on the mean and to use the U-test for comparison. What are the advantages and disadvantages of using this test compared with (a)? What are the results ($\alpha = 0.05$)?

Exercise 10.7 Otto loves gummy bears and buys 10 packets at a factory store. He opens all packets and sorts them by their colour. He counts 222 white gummy bears, 279 red gummy bears, 251 orange gummy bears, 232 yellow gummy bears, and 266 green ones. He is disappointed since white (pineapple flavour) is his favourite flavour. He hypothesizes that the producer of the bears does not uniformly distribute the bears into the packets. Choose an appropriate test to find out whether Otto's speculation could be true.

Exercise 10.8 We consider Exercise 4.4 where we evaluated which of the passengers from the *Titanic* were rescued. The data was summarized as follows:

	1. Class	2. Class	3. Class	Staff	Total
Rescued	202	125	180	211	718
Not rescued	135	160	541	674	1510

(a) The hypothesis derived from the descriptive analysis was that travel class and rescue status are not independent. Test this hypothesis.

(b) Interpret the following *R* output:

```
4-sample test for equality of proportions
data:  titanic
X-squared = 182.06, df = 3, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.5994065 0.4385965 0.2496533 0.2384181
```

(c) Summarize the data in a 2×2 table: passengers from the first and second class should be grouped together, and third class passengers and staff should be grouped together as well. Is the probability of being rescued higher in the first and second class? Provide an answer using the following three tests: exact test of Fisher, χ^2-independence test, and χ^2-homogeneity test. You can use *R* to conduct the test of Fisher.

Exercise 10.9 We are interested in understanding how well the t-test can detect differences with respect to the mean. We use *R* to draw 3 samples each of 20 observations from three different normal distributions: $X \sim N(5, 2^2)$, $Y_1 \sim N(4, 2^2)$, and $Y_2 \sim N(3.5, 2^2)$. The summary statistics of this experiment are as follows:

- $\bar{x} = 4.97$, $s_x^2 = 2.94$,
- $\bar{y}_1 = 4.55$, $s_{y_1}^2 = 2.46$,
- $\bar{y}_2 = 3.27$, $s_{y_2}^2 = 3.44$.

(a) Use the t-test to compare the means of X and Y_1.
(b) Use the t-test to compare the means of X and Y_2.
(c) Interpret the results from (a) and (b).

Exercise 10.10 Access the theatre data described in Appendix A.4. The data summarizes a survey conducted on visitors of a local Swiss theatre in terms of age, sex, annual income, general expenditure on cultural activities, expenditure on theatre visits, and the estimated expenditure on theatre visits in the year before the survey was done.

(a) Compare the mean expenditure on cultural activities for men and women using the Welch test ($\alpha = 0.05$).
(b) Would the conclusions change if the two-sample t-test or the U-test were used for comparison?
(c) Test the hypothesis that women spend on average more money on theatre visits than men ($\alpha = 0.05$).
(d) Compare the mean expenditure on theatre visits in the year of the survey and the preceding year ($\alpha = 0.05$).

Exercise 10.11 Use *R* to read in and analyse the pizza data described in Appendix A.4 (assume $\alpha = 0.05$).

(a) The manager's aim is to deliver pizzas in less than 30 min and with a temperature of greater than 65 °C. Use an appropriate test to evaluate whether these aims have been reached on average.
(b) If it takes longer than 40 min to deliver the pizza, then the customers are promised a free bottle of wine. This offer is only profitable if less than 15 % of deliveries are too late. Test the hypothesis $p < 0.15$.
(c) The manager wonders whether there is any relationship between the operator taking the phone call and the pizza temperature. Assume that a hot pizza is defined to be one with a temperature greater 65 °C. Use the test of Fisher, the χ^2-independence test, and the χ^2-test of homogeneity to test his hypothesis.
(d) Each branch employs the same number of staff. It would thus be desirable if each branch receives the same number of orders. Use an appropriate test to investigate this hypothesis.
(e) Is the proportion of calls taken by each operator the same in each branch?
(f) Test whether there is a relationship between drivers and branches.

Exercise 10.12 The authors of this book went to visit historical sites in India. None of them has a particularly strong interest in photography, and they speculated that each of them would take about the same number of pictures on their trip. After returning home, they counted 110, 118, and 105 pictures, respectively. Use an appropriate test to find out whether their speculation was correct ($\alpha = 0.01$).

→ Solutions to all exercises in this chapter can be found on p. 393

*Toutenburg, H., Heumann, C., *Induktive Statistik*, 4th edition, 2007, Springer, Heidelberg