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 $A_1 = \{\omega_1\} = 1 = \text{head}$ and $A_2 = \{\omega_2\} = 0 = \text{tail}$. 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 7.2.1 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:

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

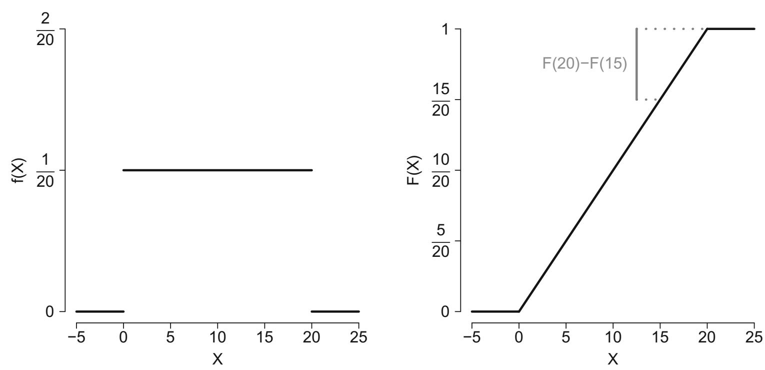
Suppose we are interested in calculating the probability of a waiting time between 15 and 20 min. This can be calculated using Theorem 7.2.2:

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

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:

```r
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 $P(x_1 \le X \le x_2)$. 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

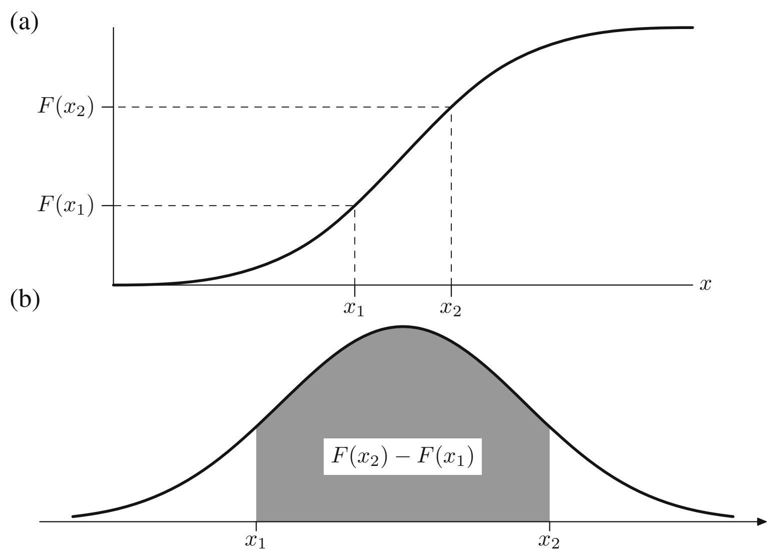


Fig. 7.2 Graphical representation of the probability $P(x_1 \leq X \leq x_2)$ **a** via the CDF and **b** via the PDF*

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

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

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

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

where I is an indicator function defined as

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

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

$$P(X \leq a) = F(a), \tag{7.7}$$

$$P(X < a) = P(X \leq a) - P(X = a) = F(a) - P(X = a), \tag{7.8}$$

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a), \tag{7.9}$$

$$P(X \geq a) = 1 - P(X < a) = 1 - F(a) + P(X = a), \tag{7.10}$$

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X < a) \\ &= F(b) - F(a) + P(X = a), \end{aligned} \tag{7.11}$$

$$P(a < X \leq b) = F(b) - F(a), \tag{7.12}$$

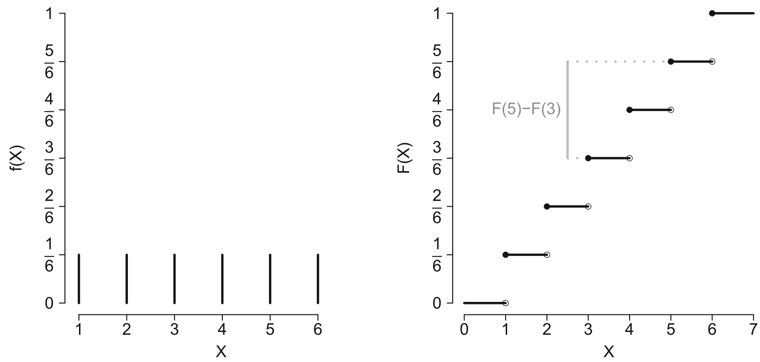
$$P(a < X < b) = F(b) - F(a) - P(X = b), \tag{7.13}$$

$$P(a \leq X < b) = F(b) - F(a) - P(X = b) + P(X = a). \tag{7.14}$$

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:

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



(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:

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

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

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

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

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

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:

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

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

7.3.3 Quantiles of a Distribution

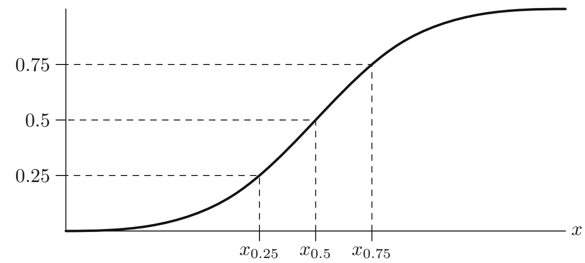
We introduced the concept of quantiles in Chap. 3, Sect. 3.1.2. Now, we define quantiles in terms of the distribution function.

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
	1	p_{11}	p_{12}	$\cdots$	p_{1J}	p_{1+}
	2	p_{21}	p_{22}	$\cdots$	p_{2J}	p_{2+}
X	$\vdots$	$\vdots$				$\vdots$
	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 $P(X = 2|Y = 3) = p_{2|3} = \frac{0.10}{0.41} = 0.24$.

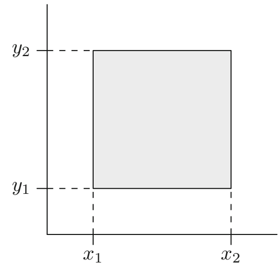


Fig. 7.5 Area covering all points of (X, Y) with $(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)^*$

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 $x_1 < x_2, y_1 < y_2$.

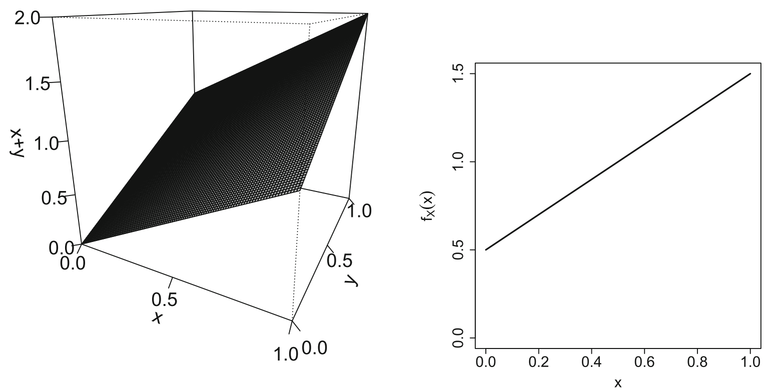
The last condition is sometimes referred to as the *rectangle inequality*. As in the univariate case, we can use the cumulative distribution function to calculate interval probabilities; similarly, we look at the rectangular area defined by (x_1, y_1), (x_1, y_2), (x_2, y_1), and (x_2, y_2) in the bivariate 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

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

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

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

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:

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

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 $\mathrm{E}(XY)$:

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

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:

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

✓ 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

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