# PROBABILITY & MEASURE THEORY

# Contents

# 1 Probability spaces

## 1.1. Random experiments

A random experiment is an experiment whose outcome cannot be predicted. In probability theory, a random experiment is described by three objects:

(1) **the state space:** the set of all possible outcomes of the random experiment, usually denoted by $\Omega$.
**Example 1.1.1.** Some examples are the following:
   (a) tossing a coin: $\Omega = \{h, t\}$, where $h$ denotes head and $t$ tail;
   (b) rolling a dice twice: $\Omega = \{(i, j) : i, j \in \{1, \dots, 6\}\}$;
   (c) number of companies going bankrupt today: $\Omega = \{0, 1, 2, 3, \dots\}$;
   (d) throwing a dice infinitely often:
   $\{1, 2, 3, 4, 5, 6\}^{\mathbb{N}} := \{(x_1, x_2, \dots,) : x_i \in \{1, 2, 3, 4, 5, 6\}\}$;
   (e) price of the share of company A tomorrow at noon: $\Omega = [0, \infty)$;
   (f) price evolution of the share of company A tomorrow during trading time: $\Omega = \{f : [9, 17] \to [0, \infty) : \text{f continuous function}\}$.

Example 1.1.1 shows various choices of sample spaces which we can categorise as:

- finite set: (a), (b),
- countable but not finite set: (c), (d)
- the real line: (e)
- finitely many replications: (b)
- infinitely many replications: (d)
- function space: (f)

(2) **the family of events:** an event is a set of outcomes of the random experiment, which can be observed to hold or not to hold after the experiment. Mathematically, an event is a subset $A$ of the state space $\Omega$, that is $A \subseteq \Omega$. Often, the family of all events is denoted by $\mathscr{A}$ in this lecture notes.

**Example 1.1.2.** (continues Example 1.1.1)

(a) the coin lands with heads up: $A = \{h\}$;

(b) the sum of the two throws is 5:

$$A = \{(i,j) \in \Omega : i + j = 5\} = \{(1,4),(2,3),(3,2),(4,1)\};$$

(c) strictly less than 10 companies go bankrupt: $A = \{0,1,2,\dots,9\}$;

(d) a five comes up before a six is thrown:
$A = \big\{x \in \Omega : \min\{j : x_j = 5\} < \min\{j : x_j = 6\}\big\}$;

(e) the share price exceeds £900 at noon tomorrow: $A = (900, \infty)$;

(f) the maximal share price tomorrow is larger than £4.000:
$\big\{f \in \Omega : \max_{t\in[9,17]} f(t) > 4.000\big\}$.

Verbal compositions of events correspond to set-theoretic compositions. If you do not remember the compositions of sets, you can find them in Appendix A.

- the event *not* $A$ is the contrary event and it is described by the complement $A^c$;
- the event $A$ *or* $B$ is described by the union $A \cup B$;
- the event $A$ *and* $B$ is described by the intersection $A \cap B$;
- the event $A$ *but not* $B$ is described by $A \backslash B := A \cap B^c$;

Binary set relations have also their interpretations:

- $A \subseteq B$ means that if $A$ occurs then also $B$ occurs;
- $A = B$ means that $A$ occurs if and only if $B$ occurs;
- $A$ and $B$ are called disjoint if they cannot occur simultaneously in the same random experiment, i.e. $A \cap B = \emptyset$.

Some special events are

- $A = \Omega$ is the *sure event;*
- $A = \emptyset$ is the *impossible event;*
- $A = \{\omega\}$ for some $\omega \in \Omega$ is an *elementary event.*

We also consider infinite sequences of sets. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets $A_n \subseteq \Omega$.

- the infinite union is defined as $\displaystyle\bigcup_{n=1}^{\infty} A_n := \{\omega \in \Omega : \omega \in A_n \text{ for some } n \in \mathbb{N}\}$. The infinite union is the event that at least one of the events $A_n$ occurs.
- the infinite intersection is defined as $\displaystyle\bigcap_{n=1}^{\infty} A_n := \{\omega \in \Omega : \omega \in A_n \text{ for all } n \in \mathbb{N}\}$. The infinite intersection is the event that all events $A_n$ occur.

(3) **the probability:** the probability is a number in $[0,1]$ which is assigned to each event. The more likely an event is the closer is the assigned number to 1. Mathematically, the probability is a function $P : \mathscr{A} \to [0,1]$, which is defined on a set $\mathscr{A}$ of subsets of $\Omega$ and which maps to the numbers in $[0,1]$.

**Example 1.1.3.** (continues Example 1.1.2)

(a) intuitively we would say $P(A) = \frac{\text{cardinality of } A}{\text{cardinality of } \Omega} = \frac{1}{2}$;

(b) intuitively we would say $P(A) = \frac{\text{cardinality of } A}{\text{cardinality of } \Omega} = \frac{4}{36}$;

(c) there is no obvious choice of a probability. A possible model assumption is that $P$ is the Poisson probability, say with parameter $\alpha = 11$. Then

$$P(A) = P(\{0\}) + P(\{1\}) + \cdots + P(\{9\}) = 0.341,$$

see Subsection 2.2.4.

(d) one could conclude $P(A) = \frac{1}{2}$, but we cannot formalise here because of the infinite sample space.

(e) there is no obvious choice of a probability. A possible model assumption is that $P$ is lognormally distributed, say with parameters $\mu = 0$ and $\sigma = 0.25$ with units in £$k$. Then we obtain $P(A) = 0.352$; see Subsection 2.3.4

(f) it depends how the share price is modelled. If the evaluation of the share price is modelled by a geometric Brownian motion, the probability of the event $A$ can be calculated.

Strictly speaking, a random experiment is described by the triplet $(\Omega, \mathscr{A}, P)$. However, often we are not interested in the single random outcome $\omega$ and its probability but on some other quantity which depends on $\omega$.

(4) **a random variable:** a random variable $X$ is a quantity which depends on the outcome of the random experiment. Mathematically, it is a function $X\colon \Omega \to \mathbb{R}$.

**Example 1.1.4.** (continues Example 1.1.3)

(b) One defines a random variable $X\colon \Omega \to \mathbb{R}$ by $X\big((i,j)\big) := i + j$. Obviously, we have

$$\{\omega \in \Omega :\ X(\omega) = 5\} = \text{“the sum of the two dice equals 5”}.$$

(e) For a fixed number $K > 0$ define a random variable $X(\omega) := \max\{\omega - K, 0\}$. Then $X$ describes the payoff of a European call option with strike price $K$ and maturity noon tomorrow, written on the share of Company A.

## 1.2. Kolmogorov's axioms

The set $\mathscr{A}$ of events is a collection of subsets of the state space $\Omega$, that is

$$\mathscr{A} \subseteq \mathscr{P}(\Omega) := \{A \subseteq \Omega\}.$$

Note that $\mathscr{A}$ is a set of sets. In contrast, a subset $A \subseteq \Omega$ is an element of $\mathscr{P}(\Omega)$, i.e. $A \in \mathscr{P}(\Omega)$.

**Example 1.2.1.**

(a) If $\Omega = \{1, 2, 3\}$ then

$$\mathscr{P}(\Omega) = \big\{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}, \emptyset\big\}.$$

(b) If $\Omega = [0, 2]$ then $\mathscr{P}(\Omega)$ consists of all subsets of $[0, 2]$, which cannot be described explicitly.

(c) For understanding the meaning of a set of sets: if $\mathscr{A} = \{\emptyset, \{1, 2, 3, 4\}, \{1, 2, 3\}, \{4\}\}$, then $\{1, 2, 3\}$ is an element in $\mathscr{A}$ but $\{1, 2\}$ is not in $\mathscr{A}$. Then $\{4\}$ is an element of $\mathscr{A}$ but not $\{3\}$. Also 4 is not an element of $\mathscr{A}$ as $\mathscr{A}$ is a set of sets here, but 4 is a number and not a set.

The considerations in part (b) of Section 1.1 indicates that $\mathscr{A}$ should be closed under finite or even countable unions and intersections. In order to achieve this, the following structure is sufficient:

**Definition 1.2.2.** *Let $\Omega$ be a non-empty set and $\mathscr{A}$ be a family of subsets of $\Omega$. Then $\mathscr{A}$ is called* a $\sigma$-algebra of $\Omega$ *if the following are satisfied:*

(1) $\Omega \in \mathscr{A}$;

(2) *if* $A \in \mathscr{A}$ *then* $A^c \in \mathscr{A}$;

(3) *if* $A_1, A_2, \ldots \in \mathscr{A}$ *then* $\displaystyle\bigcup_{k=1}^{\infty} A_k \in \mathscr{A}$.

**Proposition 1.2.3.** *Let $\mathscr{A}$ be a $\sigma$-algebra and $A$, $B \in \mathscr{A}$ and $A_1, A_2, \ldots \in \mathscr{A}$. Then we have*

(a) $\emptyset \in \mathscr{A}$ *as* $\emptyset = \Omega^c$.

(b) $A \cup B \in \mathscr{A}$ *as* $A \cup B = A \cup B \cup \emptyset \cup \emptyset \cup \cdots$.

(c) $A \cap B \in \mathscr{A}$ *as* $A \cap B = (A^c \cup B^c)^c$.

(d) $\bigcap_{n=1}^{\infty} A_n \in \mathscr{A}$ *as* $\bigcap_{n=1}^{\infty} A_n = \big(\bigcup_{n=1}^{\infty} A_n^c\big)^c$.

In the indicated proof of Proposition 1.2.3 we use the rules for set operations as presented in Appendix A and extended in Exercise 1.3.1. In conclusion, any countable repetition of any set-theoretic composition applied to sets in a $\sigma$-algebra results again in a set in the $\sigma$-algebra. One says, that a $\sigma$-algebra is closed under countable set-theoretic compositions.

**Example 1.2.4.** Some simple examples of a $\sigma$-algebra are the following:

(a) $\{\emptyset, \Omega\}$ (the trivial $\sigma$-algebra);

(b) $\mathscr{P}(\Omega) = \{A \subseteq \Omega\}$; (power set of $\Omega$)

(c) $\{\emptyset, \Omega, A, A^c\}$ for a subset $A \subseteq \Omega$.

In many cases it is not possible to describe explicitly all sets of a $\sigma$-algebra. However, for a given collection of sets one can define the *smallest $\sigma$-algebra* which contains the given collection of sets. For this purpose we need the following:

**Theorem 1.2.5.** *Let $\{\mathscr{A}_i\}_{i\in I}$ be a family of $\sigma$-algebras on $\Omega$ and $I$ an arbitrary (not necessarily countable) index set. Then*

$$\bigcap_{i\in I} \mathscr{A}_i := \big\{ A \subseteq \Omega : \; A \in \mathscr{A}_i \text{ for all } i \in I \big\}$$

*is a $\sigma$-algebra on $\Omega$.*

Theorem 1.2.5 enables us to define the concept of the *smallest $\sigma$-algebra which contains a given set of sets*: let $\mathscr{C}$ be a collection of subsets of $\Omega$, that is $\mathscr{C} \subseteq \mathscr{P}(\Omega)$, and assume that the smallest $\sigma$-algebra exists and is denoted by $\sigma(\mathscr{C})$. In order to make sense of the meaning *smallest $\sigma$-algebra which contains $\mathscr{C}$* one could postulate the following reasonable conditions:

(1) $\mathscr{C} \subseteq \sigma(\mathscr{C})$;

(2) if $\mathscr{A}$ is a $\sigma$-algebra of $\Omega$ with $\mathscr{C} \subseteq \mathscr{A}$ then $\sigma(\mathscr{C}) \subseteq \mathscr{A}$.

Obviously, (1) is the mathematical description of the phrase $\sigma(\mathscr{C})$ *contains* $\mathscr{C}$ and (2) of the phrase *smallest*. If we define $\sigma(\mathscr{C})$ as the intersection of all $\sigma$-algebras which contains $\mathscr{C}$, then Theorem 1.2.5 guarantees that $\sigma(\mathscr{C})$ is a $\sigma$-algebra. Its very definition as the intersection shows that it satisfies the Conditions (1) and (2) above, which leads to the following definition:

**Definition 1.2.6.** *For a fixed collection $\mathscr{C} \subseteq \mathscr{P}(\Omega)$ of sets, the* smallest $\sigma$-algebra which contains $\mathscr{C}$ *or the* $\sigma$-algebra generated by $\mathscr{C}$ *is defined by*

$$\sigma(\mathscr{C}) := \bigcap_{\substack{\mathscr{A}_i \; \sigma\text{-algebra} \\ \mathscr{C}\subseteq \mathscr{A}_i}} \mathscr{A}_i.$$

**Example 1.2.7.**

(a) If $\Omega = \{1,2,3,4\}$ and $\mathscr{C} = \{\{\{1,2\}\}$ then $\sigma(\mathscr{C}) = \{\emptyset, \{1,2\}, \{3,4\}, \Omega\}$.

(b) If $A$ is a subset of $\Omega$, then $\sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$.

(c) Let $\Omega$ be of the form $\Omega = C_1 \cup C_2 \cup \cdots$ for some disjoint sets $C_1, C_2, \dots$. If $\mathscr{C} = \{C_1, C_2, \dots\}$ then

$$\sigma(\mathscr{C}) = \Big\{ \bigcup_{i\in I} C_i : \; \text{for each } I \subseteq \mathbb{N} \Big\}.$$

Another example of a generated $\sigma$-algebra is the Borel-$\sigma$-algebra in $\mathbb{R}$:

**Definition 1.2.8.** *Let $\Omega = \mathbb{R}$. The $\sigma$-algebra generated by all open intervals,*

$$\mathscr{C} := \big\{(a,b) : -\infty < a < b < +\infty\big\},$$

*is called the* Borel $\sigma$-algebra in $\mathbb{R}$ *and is denoted by* $\mathfrak{B}(\mathbb{R})$.

It is not possible to write down explicitly all sets in $\mathfrak{B}(\mathbb{R})$. However, since it is generated by all intervals of the form $(a,b)$ one can deduce (see Exercise 1.3.8) that

- all elementary sets of the form $\{x\}$ for $x \in \mathbb{R}$ are in $\mathfrak{B}(\mathbb{R})$.
- all intervals of the form $[a,b)$ for $a \leqslant b$ are in $\mathfrak{B}(\mathbb{R})$;
- all intervals of the form $(a,b]$ for $a \leqslant b$ are in $\mathfrak{B}(\mathbb{R})$;
- all open sets are in $\mathfrak{B}(\mathbb{R})$;
- all closed sets are in $\mathfrak{B}(\mathbb{R})$;

In summary, each "reasonable" set is in the Borel $\sigma$-algebra. One can even replace the generator $\mathscr{C}$ in Definition 1.2.8 by other collection of sets and one still obtains the same generated $\sigma$-algebra (without proof):

$$\begin{aligned}\mathfrak{B}(\mathbb{R}) &= \sigma\big(\{(a,b] : -\infty < a < b < +\infty\}\big)\\ &= \sigma\big(\{(-\infty,b] : -\infty < b < +\infty\}\big)\\ &= \sigma\big(\{A \subseteq \mathbb{R} \text{ open}\}\big).\end{aligned}$$

Although many sets are included in the Borel $\sigma$-algebra $\mathfrak{B}(\mathbb{R})$, there are subsets of $\Omega$ which are *not* in $\mathfrak{B}(\mathbb{R})$, that is $\mathfrak{B}(\mathbb{R}) \subsetneq \mathscr{P}(\mathbb{R})$. An example of a set $A \subseteq \mathbb{R}$ which is not in $\mathfrak{B}(\mathbb{R})$ is the so-called Lusin's set of continued fractions (Wikipedia).

The Borel $\sigma$-algebra in higher dimensions is defined analogously by taking all Cartesian products of intervals. For example, $(0,1) \times (0,1) \times (0,1)$ is the cube with side length 1 in $\mathbb{R}^3$. Again, all reasonable sets are also in the Borel $\sigma$-algebra in higher dimensions.

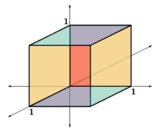


Figure 1.1.: cube

**Definition 1.2.9.** *Let $\Omega = \mathbb{R}^d$. The $\sigma$-algebra generated by*

$$\mathscr{C} := \big\{(a_1,b_1) \times \cdots \times (a_d,b_d) : -\infty < a_i < b_i < +\infty \textit{ for } i = 1,\dots,d\big\}$$

*is called the* Borel $\sigma$-algebra in $\mathbb{R}^d$ *and is denoted by* $\mathfrak{B}(\mathbb{R}^d)$.

**Definition 1.2.10.** *Let $\Omega$ be a non-empty set with a $\sigma$-algebra $\mathscr{A}$. A* probability measure *is a function $P\colon \mathscr{A} \to [0,1]$ which satisfies:*

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

(2) *each sequence* $A_1, A_2, \ldots \in \mathscr{A}$ *of pairwise disjoint sets satisfies*

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$

*The triplet* $(\Omega, \mathscr{A}, P)$ *is called a* probability space.

The sets $A_1, A_2, \ldots \in \mathscr{A}$ are said to be pairwise disjoint if $A_i \cap A_j = \emptyset$ for $i \neq j$. Property (2) in Definition 1.2.10 is called *countable additivity* or *$\sigma$-additivity*. From a probability measure we expect intuitively, that the probability for the union of two disjoint events $A$ and $B$ just adds up:

$$P(A \cup B) = P(A) + P(B).$$

Condition (2) is the generalisation to countably many events.

A probability measure $P$ is a function with the $\sigma$-algebra $\mathscr{A}$ as its domain and the interval $[0,1]$ as its codomain (target set). Hence, we say for example that $P$ is a measure on $\mathscr{A}$. However, for convenience (and this is very often the case in the literature), we also say that $P$ is a measure on $\Omega$, although this is formally incorrect.

**Example 1.2.11.** Let $\Omega$ be a set with a $\sigma$-algebra $\mathscr{A}$. For a fixed element $\omega_0 \in \Omega$ define

$$\delta_{\omega_0} \colon \mathscr{A} \to [0,1], \qquad \delta_{\omega_0}(A) = \begin{cases} 1, & \text{if } \omega_0 \in A, \\ 0, & \text{else.} \end{cases}$$

This probability measure is called the *Dirac measure in* $\omega_0$.

For fixed elements $\omega_0$, $\omega_1 \in \Omega$ and $\alpha \in [0,1]$ define

$$P \colon \mathscr{A} \to [0,1], \qquad P(A) = \alpha\delta_{\omega_0}(A) + (1-\alpha)\delta_{\omega_1}(A).$$

This probability measure describes the Bernoulli experiment; see Subsection 2.2.2.

**Example 1.2.12.** Let $\Omega = \{1, \ldots, n\}$ and choose $\mathscr{A} = \mathscr{P}(\Omega)$. Then we can define a probability measure by

$$P \colon \mathscr{A} \to [0,1], \qquad P(A) := \tfrac{\text{cardinality of } A}{\text{cardinality of } \Omega}.$$

This probability measure is called the *uniform distribution on* $\Omega$; see Subsection 2.2.1

Figure 1.2.: Banach-Tarski ball magic

If the power set is always a $\sigma$-algebra, why do not we just take always the power set as the underlying set of possible events? The answer is given by the Banach-Tarski paradox. This mathematically correct theorem says that a solid ball in $\mathbb{R}^3$ can be decomposed into a finite number of disjoint subsets of $\mathbb{R}^3$, which can then be reassembled in a way that yields two copies of the original ball! The reassembly process only involves moving and rotating the subsets without changing their shape. (This theorem is true as the subsets are very strange objects, i.e. not solid in our sense).

As a consequence, one can not have a measure on all subsets of $\mathbb{R}^3$ which just assigns the volume to each subset. Because, on the one hand side, the union of the finite number of disjoint subsets is the given ball. On the other hand, by reassembling the subsets, which does not change their volume, we obtain 2 balls, which together have the double volume of the original ball. Thus, the assumed property of a probability measure cannot be satisfied! As a solution, we have to restrict the possible sets, which leads to the Borel-$\sigma$-algebra among others. In particular, the Borel $\sigma$-algebra is strictly smaller than the power set.

There are different interpretations of probabilities (Wikipedia: Probability interpretations). The most common one is the so-called *frequentist probability*, which interprets the probability $P(A)$ of an event $A$ as the limit of the relative frequency of the event $A$ in a large number of trials. This corresponds to the examples above of rolling a die and others. The mathematical approach, established by A. Kolmogorov in the 1930's, is purely axiomatic: within the mathematical framework we started here and will build up, one obtains many powerful results but all of them rely what the "user" actually models by the probability space $(\Omega, \mathscr{A}, P)$. In some cases, like rolling a dice, the model $(\Omega, \mathscr{A}, P)$ is quite obvious, but in other cases, e.g. modelling a share price, it is much less. Even if one has introduced a model, which seems to present the real world very well, one should not forget that in the end it is just a model (financial crisis 2007?).

**Theorem 1.2.13.** Properties of a probability measure
*Let $(\Omega, \mathscr{A}, P)$ be a probability space and $A, B, A_1, A_2, \ldots \in \mathscr{A}$. Then we have:*

(a) $P(\emptyset) = 0$;

(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;

(c) $P(A) = 1 - P(A^c)$;

(d) $A \subseteq B \Rightarrow P(A) \leqslant P(B)$; *(monotone);*

(e) $P\left(\bigcup_{k=1}^{\infty} A_k\right) \leqslant \sum_{k=1}^{\infty} P(A_k)$; *($\sigma$-subadditive);*

(f) *if* $A_1 \subseteq A_2 \subseteq \cdots$ *then* $\lim_{n\to\infty} P(A_n) = P\Big(\bigcup_{k=1}^{\infty} A_k\Big)$;

(g) *if* $A_1 \supseteq A_2 \supseteq \cdots$ *then* $\lim_{n\to\infty} P(A_n) = P\Big(\bigcap_{k=1}^{\infty} A_k\Big)$.

**Remark 1.2.14.** From a mathematical point of view, Property (d) is obvious and easy to proof. However, it does not always coincide with people's intuition; cf. Linda's problem originated by an article from A. Tversky and D. Kahneman (see Wikipedia: *conjunction fallacy*).

## 1.3. Exercises

1. (De Morgan's laws)
Let $\{A_k\}_{k\in\mathbb{N}}$ be a sequence of subsets of a set $\Omega$. Show that
   (a) $\big(A_1 \cap A_2\big)^c = A_1^c \cup A_2^c$ and $\big(A_1 \cup A_2\big)^c = A_1^c \cap A_2^c$.
   (b) $\left(\bigcap_{k=1}^{\infty} A_k\right)^c = \bigcup_{k=1}^{\infty} A_k^c$ and $\left(\bigcup_{k=1}^{\infty} A_k\right)^c = \bigcap_{k=1}^{\infty} A_k^c$.

2. A machine produces $n$ memory sticks. Let $A_k$ describe the event that the $k$-th memory stick is broken. Describe the following events by $A_1, \dots, A_n$ and by the usual set-theoretic operations:
   (a) at least one memory stick is broken;
   (b) all memory sticks work correctly;
   (c) exactly one memory stick is broken;
   (d) at most one memory stick is broken.

3. A student is randomly chosen in a lecture class. Let $A$ denote the event that this student is male, $B$ the event that this student does not smoke and $C$ that this student lives in a student hall.
   (a) Describe the event $A \cap (B \cap C)^c$ in words.
   (b) What is the interpretation of the equality $A \cap B \cap C = A$?
   (c) What is the interpretation the of the relation $C^c \neq B$?
   (d) What is the interpretation of $A^c = B$? Is $A = B^c$ true, if all male students are smokers?

4. A coin is flipped four times, where only head or tail can appear. We are interested in the events

$$\begin{aligned} A &= \text{“coin does not land with heads up before the third toss”} \\ B &= \text{“coin does not land with heads up in the first and the third toss”.} \end{aligned}$$

   (a) Define the state space $\Omega$.
   (b) Describe the events $A$ and $B$ as subsets of $\Omega$ in mathematical terms.
   (c) State explicitly the events $A \cup B$ and $A \cap B$ and give their verbal interpretation.

5. Let $f\colon \Omega \to \mathbb{R}$ be a function. Define the pre-images of $f$ by

$$f^{-1}(B) := \{\omega \in \Omega : \, f(\omega) \in B\} \qquad \text{for all } B \subseteq \mathbb{R}\,.$$

   a) Show that for each $B \subseteq \mathbb{R}$ we have $f^{-1}(B^c) = f^{-1}(B)^c$.

b) Let $B_k \subseteq \mathbb{R}$ for $k \in I$ and an index set $I$. Show that

$$f^{-1}\left(\bigcup_{k\in I} B_k\right) = \bigcup_{k\in I} f^{-1}(B_k) \quad \text{and} \quad f^{-1}\left(\bigcap_{k\in I} B_k\right) = \bigcap_{k\in I} f^{-1}(B_k).$$

c) Let $\mathscr{A}$ be a $\sigma$-algebra in $\Omega$ and $\mathscr{E}$ a $\sigma$-algebra in $\mathbb{R}$. Show that

$$\mathscr{C} := \{B \in \mathscr{E} : f^{-1}(B) \in \mathscr{A}\}$$

defines a $\sigma$-algebra in $\mathbb{R}$.

6. Let $\mathscr{A}$ be a $\sigma$-algebra on $\Omega$ and let $B \subseteq \Omega$. Define

$$\mathscr{A} \cap B := \{A \cap B : A \in \mathscr{A}\}.$$

Show that $\mathscr{A} \cap B$ defines a $\sigma$-algebra of $B$.

7. Let $E = \{a, b, c, d, e, f\}$ and $\mathscr{C} = \big\{\{a, b\}, \{c, d, e\}\big\}$. Determine the $\sigma$-algebra $\sigma(\mathscr{C})$ of $E$ which is generated by $\mathscr{C}$.

8. Show by only using the Definition of the Borel-$\sigma$-algebra that the following sets are in the Borel $\sigma$-algebra $\mathfrak{B}(\mathbb{R})$:
   (a) $(-\infty, b)$, $[b, \infty)$, $[a, b)$ and $[a, b]$ for $a < b$.
   (b) $\{x\}$ for $x \in \mathbb{R}$;
   (c) $\mathbb{N}$.
   (d) $\mathbb{Q}$.

9. Let $f : \Omega \to \mathbb{R}$ be a function on a set $\Omega$. Show that

$$\mathscr{A} := \big\{f^{-1}(B) : B \in \mathfrak{B}(\mathbb{R})\big\}$$

defines a $\sigma$-algebra of $\Omega$, where $f^{-1}(B) := \{\omega \in \Omega : f(\omega) \in B\}$.

10. (1) Let $A$ and $B$ be subsets of a non-empty set $\Omega$.
    (a) Determine $\sigma(\{A, B\})$ if $A \cap B = \emptyset$;
    (b) Determine $\sigma(\{A, B\})$ if $A \cap B \neq \emptyset$;
    (2) Proof claim in Example 1.2.7.b.

11. $(*)$ Let $\Omega$ be a non-empty set and define

$$\mathscr{A} := \{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\}.$$

   (a) Show that $\mathscr{A}$ defines a $\sigma$-algebra.
   (b) Define the collection $\mathscr{C} := \{\{\omega\} : \omega \in \Omega\}$. Show that $\mathscr{A} = \sigma(\mathscr{C})$.

(c) Assume that $\Omega$ is not countable, e.g. $\Omega = \mathbb{R}$. Show that

$$P(A) := \begin{cases} 0, & \text{if } A \text{ is countable,} \\ 1, & \text{if } A^c \text{ is countable} \end{cases}$$

defines a probability measure.

12. (∗∗)(completion of a probability space)
Let $(\Omega, \mathscr{A}, P)$ be a probability space and define

$$\mathfrak{N} := \{F \subseteq \Omega : \text{ there exists } N \in \mathscr{A} \text{ such that } P(N) = 0 \text{ and } F \subseteq N\}.$$

(a) Show that $\mathscr{A}' := \{A \cup F : A \in \mathscr{A}, F \in \mathfrak{N}\}$ defines a $\sigma$-algebra of $\Omega$.

(b) Show that

$$P' : \mathscr{A}' \to [0, \infty], \qquad P'(A \cup F) := P(A)$$

defines a probability measure on $\mathscr{A}'$.

# 2

# Probability measures

In this chapter we assume $\Omega = \mathbb{R}$ and $\mathscr{A} = \mathfrak{B}(\mathbb{R})$ and and introduce various examples of probability measures on $\mathfrak{B}(\mathbb{R})$. Since we only consider probability measures on the Borel $\sigma$-algebra on $\mathbb{R}$ and not on an arbitrary $\sigma$-algebra, it is possible to completely characterise probability measures by their values for intervals of the form $(-\infty, x]$.

## 2.1. Cumulative distribution function

**Definition 2.1.1.** *Let $P\colon \mathscr{A} \to [0,1]$ be a probability measure on $\mathfrak{B}(\mathbb{R})$. The* cumulative distribution function *of $P$ is defined by*

$$F_P\colon \mathbb{R} \to [0,1], \qquad F_P(x) = P\big((-\infty, x]\big).$$

One can show (we do not) that the cumulative distribution function determines completely a probability measure: if $F_P$ and $F_Q$ are distribution functions of two probability measures $P$ and $Q$ on $\mathfrak{B}(\mathbb{R})$ one has the following equivalence:

$$F_P(x) = F_Q(x) \quad \text{for all } x \in \mathbb{R} \iff P(A) = Q(A) \quad \text{for all } A \in \mathfrak{B}(\mathbb{R}).$$

Since it is often easier to consider all values $x \in \mathbb{R}$ instead of all sets in $\mathfrak{B}(\mathbb{R})$, it is sometimes preferable to work with the cumulative distribution function of a probability measure.

**Lemma 2.1.2.** *Let $F_P$ be the cumulative distribution function of a probability measure $P$ on $\mathfrak{B}(\mathbb{R})$. Then we obtain:*

(a) *$F_P$ is non-decreasing: $F_P(x) \leqslant F_P(y)$ for all $x \leqslant y$;*

(b) *$F_P$ is right-continuous: $F_P(x) = \lim\limits_{y \searrow x} F_P(y)$ for all $x \in \mathbb{R}$;*

(c) $\lim_{x\to-\infty} F_P(x) = 0$ *and* $\lim_{x\to\infty} F_P(x) = 1$.

The notation $\lim_{y\searrow x} F_P(y)$ denotes the limit of $y \to x$ with $y \geqslant x$ and analogously $\lim_{y\nearrow x} F_P(y)$ the limit of $y \to x$ with $y \leqslant x$ for fixed $x \in \mathbb{R}$. Since $F_P$ is non-decreasing, also the left-hand limit $\lim_{y\nearrow x} F(x)$ always exists and we use the notation

$$F_P(x-) := \lim_{y\nearrow x} F_P(y) \quad \text{and} \quad F_P(x+) := \lim_{y\searrow x} F_P(y)$$

But a cumulative distribution function is not necessarily continuous, and thus we could have $F_P(x-) \neq F_P(x+)$.

From the very definition of a cumulative distribution function, one obtains the following relations between the cumulative distribution function $F$ and the corresponding probability measure $P$ (see Exercise 2.5.1):

$$\begin{aligned} F_P(b) &= P\big((-\infty, b]\big) & 1 - F_P(b-) &= P\big([b, \infty)\big) \\ F_P(b-) &= P\big((-\infty, b)\big) & 1 - F_P(b) &= P\big((b, \infty)\big) \\ F_P(b) - F_P(a) &= P\big((a, b]\big) & F_P(b) - F_P(a-) &= P\big([a, b]\big) \\ F_P(b) - F(b-) &= P\big(\{b\}\big) & F_P(b-) - F_P(a-) &= P\big([a, b)\big) \end{aligned}$$

for all $a < b$.

## 2.2. Discrete probability measures

**Definition 2.2.1.** *A probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ is called* discrete *if there exists a countable set $\Gamma \subseteq \mathbb{R}$ such that $P(\Gamma) = 1$. The set $\Gamma$ is called the* support of $P$.

A countable set $\Gamma$ is always of the form $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ for some fixed $n \in \mathbb{N}$, in which case it is finite, or $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$, in which case it is infinitely countable. In most cases, $\Gamma = \{1, \ldots, n\}$ for a number $n \in \mathbb{N}$ or $\Gamma = \mathbb{N}$. We always write $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$ for both the finite and infinitely countable case.

**Theorem 2.2.2.** *Let $P\colon \mathfrak{B}(\mathbb{R}) \to [0, 1]$ be a function. Then the following is equivalent:*

(a) *$P$ is a discrete probability measure;*

(b) *there are numbers $\gamma_1, \gamma_2, \ldots \in \mathbb{R}$ and $p_1, p_2, \ldots \geqslant 0$ with $\sum_{k=1}^{\infty} p_k = 1$ and*

$$P(A) = \sum_{k=1}^{\infty} p_k \delta_{\gamma_k}(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}), \tag{2.2.1}$$

*where $\delta_{\gamma_k}$ denotes the Dirac measure in $\gamma_k$; see Example 1.2.11.*

*In this case, it follows that $P(\{\gamma_k\}) = p_k$ for all $k \in \mathbb{N}$.*

According to Theorem 2.2.2, a discrete probability measure $P$ with support $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$ is determined by the probabilities $P(\{\gamma_k\})$ of the elementary events $\{\gamma_k\}$. In particular, to define a discrete probability measure with support $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$ it is sufficient to assign the probability $p_k$ to each elementary event $\{\gamma_k\}$. One only has to make sure that $p_1 + p_2 + \cdots = 1$. This recipe will be used in the following examples.

**Corollary 2.2.3.** *Let $P\colon \mathfrak{B}(\mathbb{R}) \to [0,1]$ be a discrete probability measure with support $\{\gamma_1, \gamma_2, \ldots\}$. Then its cumulative distribution function $F_P$ is of the form*

$$F_P\colon \mathbb{R} \to [0,1], \qquad F_P(x) = \sum_{k:\gamma_k \leqslant x} P(\{\gamma_k\}).$$

*In particular, $F_P$ is a piece-wise constant functions with jumps of magnitude $P(\{\gamma_k\})$ in $\gamma_k$.*

**Remark 2.2.4.** If $P$ is a discrete probability measure with support $\Gamma$ we have $P(A) = 0$ for all sets $A \in \mathfrak{B}(\mathbb{R})$ with $A \cap \Gamma = \emptyset$ due to (5.1.2). Consequently, we sometimes consider discrete probability measures only defined on the $\sigma$-algebra $\mathscr{P}(\Gamma)$. Since $\mathscr{P}(\Gamma) = \mathfrak{B}(\mathbb{R}) \cap \Gamma$ this is indeed a restriction of the mapping $P\colon \mathfrak{B}(\mathbb{R}) \to [0,1]$.

### 2.2.1. Uniform distribution

Assume that $\Gamma = \{\gamma_1 \ldots, \gamma_n\}$ is a finite set for a fixed number $n \in \mathbb{N}$ and one wants to assign the same probability to each elementary event $\{\gamma_k\}$. If we set $p_k := P(\{\gamma_k\}) := \frac{1}{n}$ for all $k = 1, \ldots, n$, we obtain

$$\sum_{k=1}^{n} p_k = n\frac{1}{n} = 1.$$

Consequently, Theorem 2.2.2 implies that

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \quad P(A) = \sum_{k=1}^{n} \frac{1}{n}\delta_{\gamma_k}(A),$$

defines a discrete probability measure with support $\Gamma$, which is called the *uniform distribution on* $\Gamma$.

The probability measure can be presented in a different form:

$$\sum_{k=1}^{n} \frac{1}{n}\delta_{\gamma_k}(A) = \frac{1}{n}\sum_{k=1}^{n} \delta_{\gamma_k}(A) = \frac{1}{n}\,|A \cap \Gamma| = \frac{|A \cap \Gamma|}{|\Gamma|},$$

where $|A \cap \Gamma|$ denotes the cardinality of the set $A \cap \Gamma$ and analogously $|\Gamma|$.

In many random experiments the underlying probability measure is a uniform distribution. In the following we adapt Example 1.1.3.a to our present setting.

**Example 2.2.5.** Continues Example 1.1.3.a.
Instead of $h$ and $t$ we denote head and tail by 1 and 0. If we set $\Gamma = \{0, 1\}$ and $p_k = \frac{1}{2}$ for $k = 0, 1$ then

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P(A) = \tfrac{1}{2}\delta_0(A) + \tfrac{1}{2}\delta_1(A) = \frac{|A \cap \{0,1\}|}{2}$$

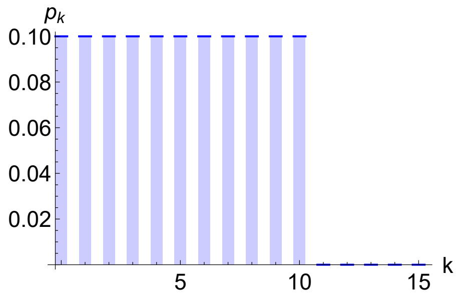



Figure 2.1.: uniform distribution on $\{1, \dots, 10\}$

defines the uniform distribution with support $\Gamma = \{0, 1\}$. Obviously, as not many sets $A$ have a non-empty intersection $A \cap \{0, 1\}$ this looks more complicated than it is.

## 2.2.2. Bernoulli distribution

Let $\Gamma = \{0, 1\}$ and $p$ be a paramter in $[0, 1]$. Define

$$P(\{1\}) = p, \quad P(\{0\}) = 1 - p.$$

Define $p_0 := 1 - p$ and $p_1 = p$. Since $p_0 + p_1 = 1$, Theorem 2.2.2 implies that

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0, 1], \quad P(A) = p_0 \delta_0(A) + p_1 \delta_1(A)$$

defines a discrete probability measure with support $\Gamma$, which is called *Bernoulli distribution*. The Bernoulli distribution often models the success of an experiment, such as in Example 1.1.1.(a), when we consider heads as success, which appears with probability $p = 0.5$. In this case, $p$ is often called the success probability and $1 - p$ the failure probability. Example 2.2.5 is the special case of the Bernoulli distribution for $p = \frac{1}{2}$.

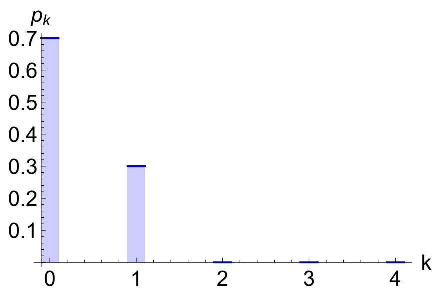



Figure 2.2.: Bernoulli distribution $p = 0.3$

## 2.2.3. Binomial distribution

Let $n \in \mathbb{N}$ and $p \in [0, 1]$ be fixed parameters. Set $\Gamma = \{0, \dots, n\}$ and define

$$p_k := \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for all } k \in \Gamma.$$

Since the binomial formula yields

$$\sum_{k=0}^{n} p_k = \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1,$$

it follows from Theorem 2.2.2 that

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P(A) = \sum_{k=0}^{n} p_k \delta_k(A),$$

defines a probability measure with support $\Gamma$ which is called the *Binomial distribution*. We use the notation $P = \operatorname{Bin}(p,n)$.

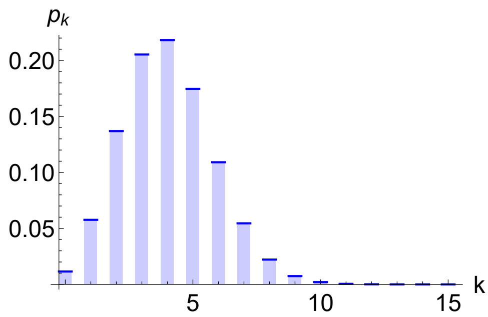



Figure 2.3.: $B(0.2, 20)$

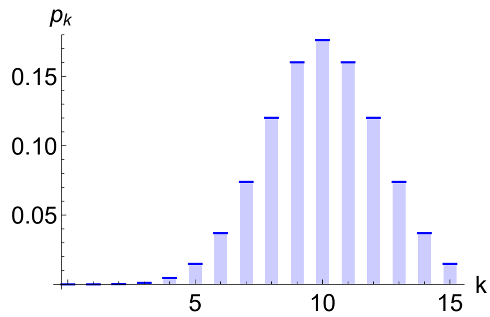


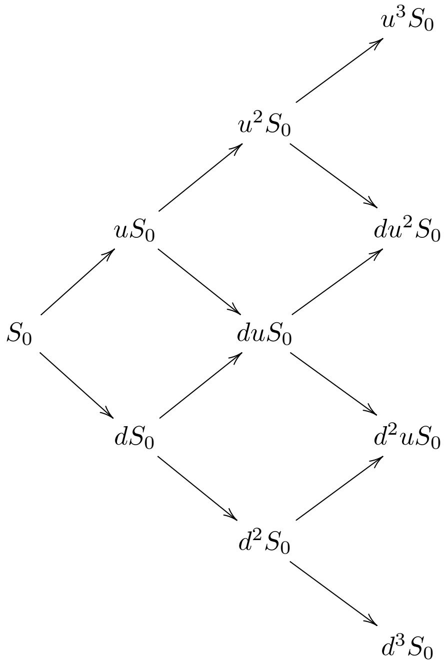

Figure 2.4.: $B(0.5, 20)$

We will see later in Example 7.1.4 that the Binomial distribution arises from the Bernoulli distribution in the following way: one repeats a random experiment $n$ times in such a way that each repetition does not influence each other. If one only counts in each repetition whether success does or does not occur then the total number of successes are binomially distributed.

**Example 2.2.6.** Assume that a die is rolled 10 times and we are only interested how often the number 6 is rolled in total but not in which toss. Define $\Gamma = \{0, \dots, 10\}$, $n = 10$ and $p = \frac{1}{6}$. Then the Binomial distribution $\operatorname{Bin}(p,n)$ models the distribution how often the number 6 is rolled in 10 tosses.

**Example 2.2.7.** Binomial model in financial mathematics

The following model in financial mathematics was introduced by Cox, Ross and Rubinstein in 1979. Although the model is rather simple it is widely used and has a few advantages to other more complicated models (see Wikipedia). Assume that the share price for the next 3 days is modeled by the lattice on the right hand side. The share price can only move up by a fixed factor $u > 1$ or move down by a fixed factor $d < 1$ at all times $t = 1, 2, 3$. With probability $p \in (0, 1)$ the price moves up on each day, and it moves down with probability $q = 1 - p$. The possible share prices $d^3S_0, d^2uS_0, du^2S_0, u^3S_0$ on day $t = 3$ are denoted by $0, 1, 2, 3$. It then follows that the share prices on day $t = 3$ are binomially distributed with parameters $p$ and $n = 3$. share value $u^3S_0$ then $X$ has a Binomial distribution $\mathrm{Bin}(p, 3)$. However, for some options like barrier options it is not enough to consider only the final value on day $t = 3$ but one has to keep track of the paths the share price went.




## 2.2.4. Poisson distribution

Let $\Gamma = \{0, 1, 2, \ldots\}$ and $\alpha > 0$ be a fixed paramter. Define

$$p_k := e^{-\alpha}\frac{\alpha^k}{k!} \quad \text{for all } k \in \Gamma.$$

Since the series representation of the exponential function yields

$$\sum_{k=0}^{\infty} p_k = e^{-\alpha}\sum_{k=0}^{\infty}\frac{\alpha^k}{k!} = e^{-\alpha}e^{\alpha} = 1,$$

it follows from Theorem 2.2.2 that

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P(A) = \sum_{k=0}^{\infty} p_k\delta_k(A),$$

defines a probability measure with support $\Gamma$, which is called the *Poisson distribution*. We use the notation $P = \mathrm{Pois}(\alpha)$.

The Poisson distribution often models the probability of the number of events occurring in a fixed time interval if these events happen with a known average rate and independently. It also can be used to approximate the Binomial distribution, see Exercise **??**.3.

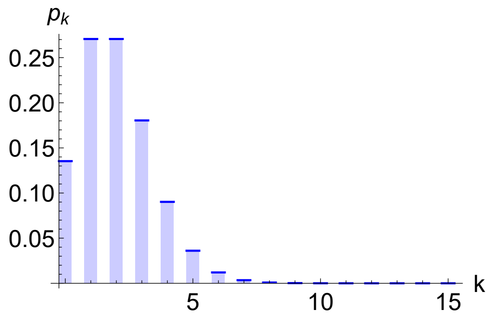



Figure 2.5.: Poisson distribution $\alpha = 2$

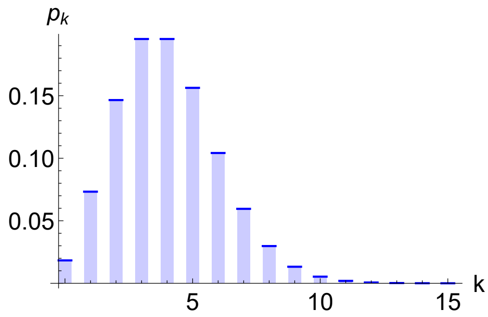



Figure 2.6.: Poisson distribution $\alpha = 4$

**Example 2.2.8.** (Continues Example 1.1.3.c)
We can model the number of bankruptcy today by a Poisson measure, say with a parameter $\alpha = 11$. We will see later that this mean we expect an average of 11 bankrupt companies today. Denote by $A_k$ the event that exactly $k$ companies go bankrupt. Then we obtain

$$P(\text{“stricly less than 10 companies bankrupt”}) = P(A_0 \cup A_1 \cup \cdots \cup A_9) = 0.341$$

### 2.2.5. Geometric distribution

Set $\Gamma = \{0, 1, 2, \ldots\}$ and let $\alpha \in (0, 1)$ be a fixed parameter. Define

$$p_k := (1-\alpha)\alpha^k \quad \text{for all } k \in \Gamma.$$

Since the convergence of the geometric series yields

$$\sum_{k=0}^{\infty} p_k = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k = (1-\alpha)\frac{1}{1-\alpha} = 1,$$

it follows from Theorem 2.2.2 that

$$P\colon\ \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P(A) = \sum_{k=0}^{\infty} p_k \delta_k(A),$$

defines a probability measure with support $\Gamma$, which is called the *geometric distribution.*

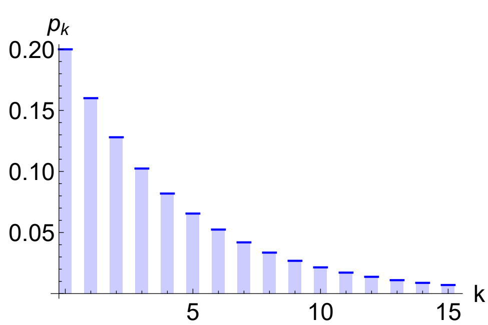



Figure 2.7.: Geometric distribution $\alpha = 0.8$

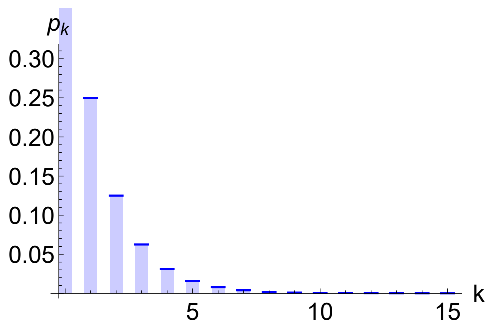



Figure 2.8.: Geometric distribution $\alpha = 0.5$

We use the notation $P = \mathrm{Geo}(\alpha)$. Note, that one can also find an alternative definition of the geometric distribution which only assigns non-zero probabilities to numbers in $\mathbb{N}$.

**Remark 2.2.9.** There is an alternative definition for the geometric distribution by

$$p_k := \alpha(1-\alpha)^{k-1} \qquad \text{for all } k \in \mathbb{N}\,.$$

Both distributions are used to describe the same random models, but the results have to be interpreted differently.

## 2.3. Absolutely continuous probability measures

Absolutely continuous probability measures are not the opposite of discrete probability measures, as there is a third type of probability measure; see Section 2.3.4. However, discrete and absolutely continuous probability measures differ in various ways.

**Definition 2.3.1.** *A probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ is called* absolutely continuous*, if there exists a non-negative (Riemann-integrable) function $f\colon \mathbb{R} \to \mathbb{R}_+$ such that*

$$P((a,b]) = \int_a^b f(u)\,du \qquad \textit{for all } -\infty < a < b < \infty.$$

*The function $f$ is called* density *of $P$.*

**Remark 2.3.2.** Definition 2.3.1 does not give the full picture, as it is based on the Riemann-integral, which is not good enough here. In fact, since the probability measure $P$ is defined on $\mathfrak{B}(\mathbb{R})$ we expect

$$P(A) = \int_A f(u)\,du \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}) \text{ and not only } A = (a,b].$$

However, if we take $f(u) = 1$ for $u \in (0,1]$ and otherwise $f(u) = 0$, which is a possible choice according to Subsection 2.3.1, and the set $A = Q \cap (0,1]$ we obtain

$$P(A) = \int_{Q\cap(0,1]} du = \int_0^1 \mathbb{1}_Q(u)\,du.$$

However, the function $u \mapsto \mathbb{1}_{\mathbb{Q}}(u)$ is not Riemann-integrable. The solution to this dilemna is the use of a more stable kind of integration, the so-called *Lebesgue integral*.

On the other hand, if the function $f$ is nice enough and we restrict ourselves only on integrals over intervals of the form $(a,b]$ we can use the Riemman-integral, which makes our life much easer. As in all of the following examples the density $f$ is “nice” enough, we will just use here the Riemann-integral.

**Theorem 2.3.3.** *If $f\colon \mathbb{R} \to \mathbb{R}_+$ is a (Riemann-integrable) function with*

$$\int_{-\infty}^{\infty} f(u)\,du = 1,$$

*then there exists an absolutely continuous probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ satisfying*

$$P\big((a,b]\big) = \int_a^b f(u)\,du \qquad \text{for all } -\infty < a < b < \infty.$$

*That is, $f$ is the density of $P$.*

We mention two very obvious properties of an absolutely continuous probability measure:

**Corollary 2.3.4.** *Let $P\colon \mathfrak{B}(\mathbb{R}) \to [0,1]$ be an absolutely continuous probability measure with density $f$. Then we have:*

(a) *The cumulative distribution function $F_P$ is of the form*

$$F_P\colon \mathbb{R} \to [0,1], \qquad F_P(x) = \int_{-\infty}^x f(u)\,du.$$

(b) $P(\{x\}) = 0$ *for all* $x \in \mathbb{R}$.

Part $(a)$ in Corollary 2.3.4 shows that the cumulative distribution function of an absolutely continuous probability measure is always continuous. One can even show that it is *absolutely continuous*, which is a stronger form of continuity. However, we do not need this property here, although it gives the name of this kind of probability measures.

### 2.3.1. Uniform distribution

Many random experiments for a finite state space are described by the uniform distribution introduced in Section 2.2.1. Analogously, we can define the uniform distribution on a finite interval.

For fixed parameters $a, b \in \mathbb{R}$ with $a < b$ define

$$f\colon \mathbb{R} \to \mathbb{R}_+, \qquad f(u) := \frac{1}{b-a}\mathbb{1}_{(a,b]}(u).$$

Since

$$\int_{-\infty}^{\infty} f(u)\,du = \int_a^b \frac{1}{b-a}\,du = 1,$$

Theorem 2.3.3 implies that there exists an absolutely continuous probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ with density $f$. The probability measure $P$ is called the *uniform distribution on* $[a,b]$.

The cumulative distribution function of the uniform distribution $P$ is given by

$$F_P(x) = \int_{-\infty}^x \frac{1}{b-a}\mathbb{1}_{[a,b]}(u)\,du = \begin{cases} 0, & \text{for } x < a, \\ \frac{x-a}{b-a}, & \text{for } x \in [a,b], \\ 1, & \text{for } x > b. \end{cases}$$

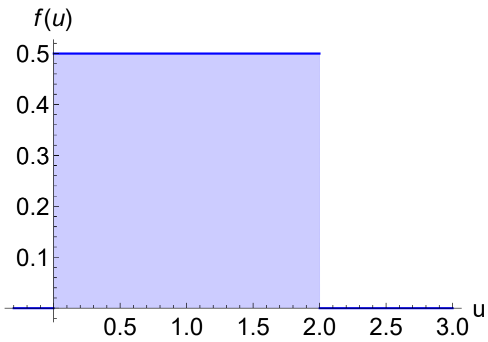



Figure 2.9.: Uniform distribution on the interval $(0, 2]$

**Example 2.3.5.** tearing a rope
A force acts on a rope of length $\ell$ such that it tears at $R \in [0, \ell]$. One assumes that for each interval of the same length it is equally likely that the rope breaks within the interval. Consequently, the probability that the rope tears in an interval $(s, t]$ with $0 \leqslant s < t \leqslant \ell$ is modelled by

$$P((s,t]) = \frac{\text{length of } (s,t]}{\text{length of rope}} = \frac{t-s}{\ell} = \int_s^t \frac{1}{\ell} \mathbb{1}_{(0,\ell]}(u)\, du.$$

For example, the set $A_c = \{x \in \mathbb{R} : x \leqslant c \text{ or } \ell - x \leqslant c\}$ describes the event that the shorter part of the torn rope is smaller than a constant $c > 0$. If $c \in (0, \ell/2]$, one obtains that

$$P(A_c) = P\big((0,c] \cup (\ell - c, \ell]\big) = P\big((0,c]\big) + P\big((\ell - c, \ell]\big) = \frac{2c}{\ell}.$$

If $c \leqslant 0$ one obtains $P(A_c) = 0$ and if $c > \ell/2$ then $P(A_c) = 1$.

### 2.3.2. Exponential distribution

For a parameter $\alpha > 0$ define

$$f\colon \mathbb{R} \to \mathbb{R}, \qquad f(u) = \alpha e^{-\alpha u} \mathbb{1}_{[0,\infty)}(u).$$

Since

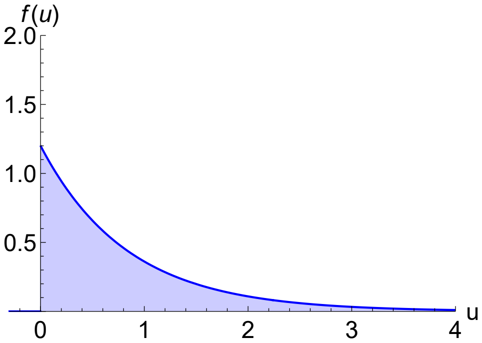
$$\int_{-\infty}^{\infty} f(u)\, du = \int_0^{\infty} \alpha e^{-\alpha u}\, du = 1,$$

Theorem 2.3.3 implies that there exists an absolutely continuous probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ with density $f$. The probability measure $P$ is called the *exponential distribution with parameter* $\alpha$ and is denoted by $\operatorname{Exp}(\alpha)$.

The cummulative distribution function of $P$ is given by

$$F_P(x) = \int_{-\infty}^{x} f(u)\, du = \begin{cases} 0, & \text{for } x \leqslant 0, \\ 1 - e^{-\alpha x}, & \text{for } x > 0. \end{cases}$$

The exponential distribution is often used to model waiting times, for example in a queue.



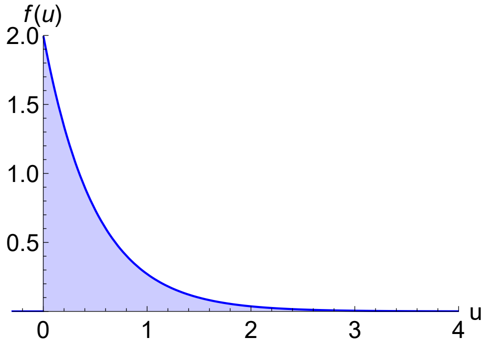




Figure 2.10.: Exponential distribution $\alpha = 1.2$ Figure 2.11.: Exponential distribution $\alpha = 2$

### 2.3.3. Normal distribution

The most important distribution with a density is considered in this subsection:

For fixed parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ define

$$\psi_{\mu,\sigma^2} \colon \mathbb{R} \to \mathbb{R}, \qquad \psi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We show that $\psi_{\mu,\sigma^2}$ is a density for $\mu = 0$ und $\sigma^2 = 1$. The general case follows from part (b) in Lemma 2.3.6. By changing to polar coordinates we obtain

$$\begin{aligned}\left(\int_{\mathbb{R}} e^{-\frac{u^2}{2}}\,du\right)^2 &= \left(\int_{\mathbb{R}} e^{-\frac{u_1^2}{2}}\,du_1\right)\left(\int_{\mathbb{R}} e^{-\frac{u_2^2}{2}}\,du_2\right)\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-\frac{u_1^2+u_2^2}{2}}\,du_1\,du_2 = \int_0^{2\pi}\int_0^{\infty} re^{-\frac{r^2}{2}}\,dr\,d\vartheta = 2\pi.\end{aligned}$$

Theorem 2.3.3 implies that there exists a probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ with density $\psi_{\mu,\sigma^2}$. The probability measure $P$ is called the *normal distribution with parameter $\mu$ and $\sigma^2$*. and is denoted by $N(\mu,\sigma^2)$.

The cumulative distribution function $\Psi_{\mu,\sigma^2}$ of the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is

$$\Psi_{\mu,\sigma^2} : \mathbb{R} \to [0,1], \qquad \Psi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du.$$

It is not possible to present $\Psi_{\mu,\sigma^2}$ in an explicit form. One has to use either tables or appropriate software in order to obtain the values for $\Psi_{\mu,\sigma^2}$.

**Lemma 2.3.6.**

(a) $\Psi_{0,1}(x) = 1 - \Psi_{0,1}(-x)$ *for all* $x \in \mathbb{R}$.

(b) $\Psi_{\mu,\sigma^2}(x) = \Psi_{0,1}\left(\frac{x-\mu}{\sigma}\right)$ *for all* $x \in \mathbb{R}$.

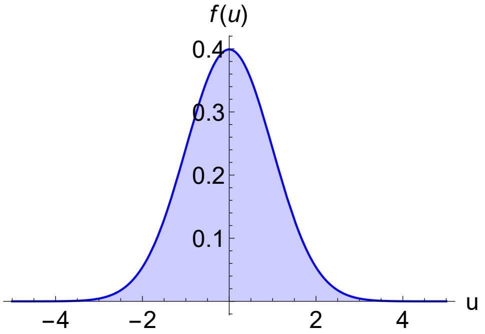



Figure 2.12.: $N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$

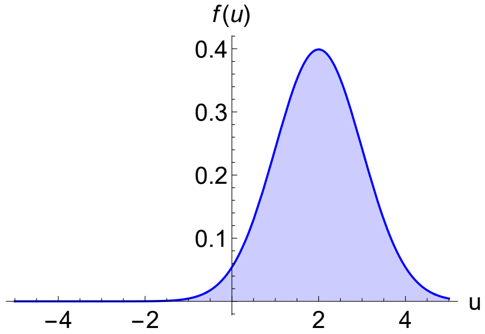



Figure 2.13.: $N(\mu, \sigma^2)$ with $\mu = 2$ and $\sigma^2 = 1$

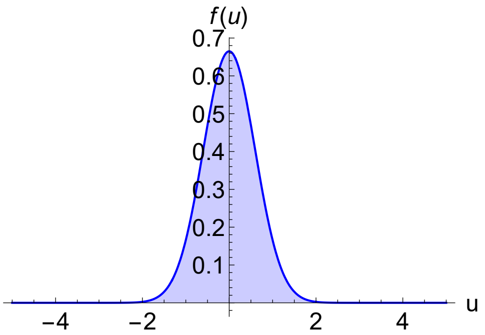



Figure 2.14.: $N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 0.6$

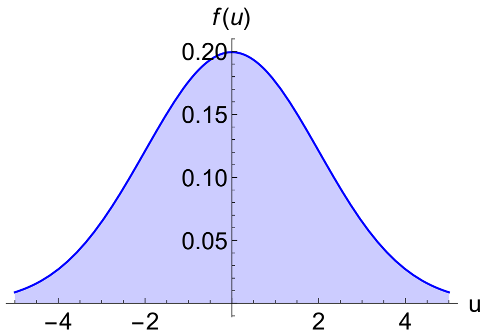



Figure 2.15.: $N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 2$

### 2.3.4. Log-normal distribution

For fixed parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ define

$$f\colon \mathbb{R} \to \mathbb{R}, \qquad f(u) = \frac{1}{u\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln u - \mu)^2}{2\sigma^2}\right) \mathbb{1}_{(0,\infty)}(u).$$

Note, that the density is only non-zero on the interval $(0, \infty)$. Integration by substitution $u \mapsto \exp(u)$ transfers the density $f$ to the density of the normal distribution $\psi_{\mu,\sigma^2}$:

$$\int_0^\infty \frac{1}{u\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln u - \mu)^2}{2\sigma^2}\right) du = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^\infty \exp\left(-\frac{(u - \mu)^2}{2\sigma^2}\right) du = 1.$$

Consequently, Theorem 2.3.3 implies that there exists an absolutely continuous probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ with density $f$. The probability measure $P$ is called the *log-normal distribution with parameter $\mu$ and $\sigma^2$*. We will later show how the log-normal distribution

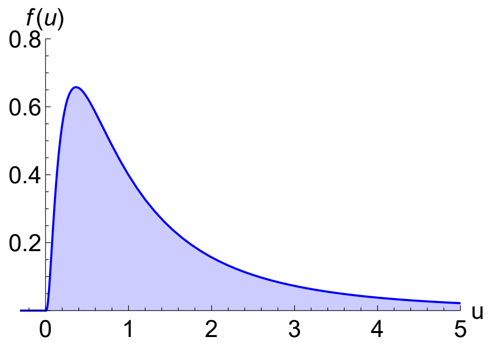



Figure 2.16.: Log-normal $\mu = 0$, $\sigma_2 = 1$

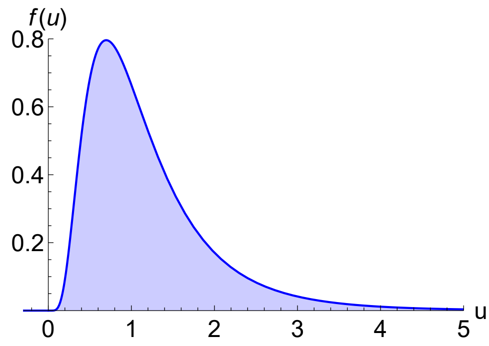



Figure 2.17.: Log-normal $\mu = 0$, $\sigma_2 = 0.6$

is related to the normal distribution.

The log-normal distribution appears the first time already in 1870 in works by Francis Galton. It is in particular used to model the share prices and other models in financial mathematics.

## 2.4. Mixed probability distributions

One can easily obtain probability measures which are mixture of discrete and absolutely continuous probability measures. For example, if $P_1$ is a discrete probability measure and $P_2$ is an absolutely continuous probability measure, then for each fixed $c \in [0, 1]$

$$P(A) := cP_1(A) + (1 - c)P_2(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R})$$

defines a probability measure on $\mathfrak{B}(\mathbb{R})$; see Exercise 2.5.6

**Example 2.4.1.** We take the mixture of the Dirac measure $P_1 = \delta_\beta$ in a point $\beta > 0$ and the exponential distribution $P_2 = \text{Exp}(\alpha)$ with parameter $\alpha > 0$ and define:

$$P(A) := c\delta_\beta(A) + (1 - c)\, P_2(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}),$$

for a constant $c \in [0, 1]$. One obtains as the cumulative distribution function of $P$:

$$F_P(x) = P\big((-\infty, x]\big) = \begin{cases} 0, & \text{if } x < 0, \\ (1 - c)(1 - e^{-\alpha x}), & \text{if } 0 \leqslant x \leqslant \beta, \\ 1 - (1 - c)e^{-\alpha x}, & \text{if } x > \beta. \end{cases}$$

In fact, one can define a third kind of probability measure, the so-called *singular probability measure with continuous cumulative function*. Then it follows that each probability measure $P$ on $\mathfrak{B}(\mathbb{R})$ is of the form

$$P(A) = c_1 P_s(A) + c_2 P_d(A) + c_3 P_c(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}),$$

where $c_1, c_2, c_3 \in [0, 1]$ with $c_1 + c_2 + c_3 = 1$ and $P_s$ is a singular probability measure with continuous cumulative function, $P_d$ is a discrete probability measure and $P_c$ is an absolutely continuous probability measure.

## 2.5. Exercises

1. Let $P$ be a probability measure on $\mathfrak{B}(\mathbb{R})$ with cumulative distribution function $F$. Show that
   (a) $F(b) - F(a) = P((a, b])$ for all $a \leqslant b$.
   (b) $F(b) - F(b-) = P(\{b\})$ for all $b \in \mathbb{R}$.
   (c) $F(b) - F(a-) = P([a, b])$ for all $a \leqslant b$.

2. (a) Let $P$ be the Poisson distribution $\mathrm{Pois}(\alpha)$. For which value of $k \in \mathbb{N}_0$ is the probability $P(\{k\})$ the greatest? (Hint: consider $P(\{k\})/P(\{k-1\})$.
   (b) Let $P$ be the Poisson distribution $\mathrm{Pois}(\alpha)$. For fixed $k \in \mathbb{N}_0$ which value of $\alpha$ maximises the probability $P(\{k\})$?
   (c) Let $P$ be the binomial distribution $\mathrm{Bin}(p, n)$. For which value of $k \in \{0, 1, \dots, n\}$ is the probability $P(\{k\})$ the greatest?

3. ($*$) (Poisson approximation) For each $n \in \mathbb{N}$ let $P_n$ be the binomial distribution with parameter $n \in \mathbb{N}$ and $p_n \in (0, 1)$, that is

$$P(\{k\}) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} \qquad \text{for } k = 0, 1, \dots, n.$$

   Show that if there exists a constant $\alpha > 0$ such that $np_n \to \alpha$ for $n \to \infty$ then it follows that

$$\lim_{n\to\infty} P(\{k\}) = \frac{\alpha^k}{k!} e^{-\alpha} \quad \text{for all } k \in \mathbb{N}_0 \,.$$

4. Derive that the cumulative distribution function of the geometric distribution $\mathrm{Geo}(\alpha)$ with parameter $\alpha > 0$ is given by

$$F \colon \mathbb{R} \to [0, 1], \qquad F(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - \alpha^{[x]+1}, & \text{else,} \end{cases}$$

   where $[x]$ denotes the largest integer smaller than $x$.

5. Define a function

$$f \colon \mathbb{R} \to \mathbb{R}_+, \qquad f(u) = \frac{1}{\pi(1 + x^2)}.$$

   (a) Show that there exists a absolutely continuous probability measure $P$ with density $f$. (This probability measure is called *Cauchy distributuion*).
   (b) Show that the distribution function $F_P$ of the Cauchy distribution $P$ is given by

$$F_P(x) = \frac{1}{\pi}\left(\frac{\pi}{2} + \arctan x\right) \qquad \text{for all } x \in \mathbb{R}\,.$$

6. Let $P_1$ and $P_2$ be two probability measures on $\mathfrak{B}(\mathbb{R})$ and $c$ a constant in $[0, 1]$. Show that

$$P\colon \mathfrak{B}(\mathbb{R}) \to [0, 1], \qquad P(A) = cP_1(A) + (1 - c)P_2(A)$$

defines a probability measure on $\mathfrak{B}(\mathbb{R})$.

# 3 Random Variables

The core object in probability theory are random variables, which we introduce in this chapter. A typical example of a random variable is given in Example 1.1.4.

## 3.1. Measurable functions

**Definition 3.1.1.** *Let $\mathscr{A}$ be a $\sigma$-algebra $\mathscr{A}$ on a set $\Omega$.*

(a) *A function $f\colon \Omega \to \mathbb{R}$ is called* measurable *if*

$$f^{-1}(B) := \{\omega \in \Omega :\ f(\omega) \in B\} \in \mathscr{A} \qquad \text{for all sets } B \in \mathfrak{B}(\mathbb{R}). \tag{3.1.1}$$

(b) *A function $X\colon \Omega \to \mathbb{R}$ is a* random variable *if it is measurable.*

*If we want to mention the underlying $\sigma$-algebra we say* $\mathscr{A}$-measurable *or* measurable with respect to $\mathscr{A}$.

The set $f^{-1}(B)$ is called the pre-image or inverse image of $B$ under $f$. It completely differs from the inverse function (unfortunately the notions are the same). In Appendix C, we collect some properties of pre-images.

**Example 3.1.2.** Let $\mathscr{A}$ be a $\sigma$-algebra $\mathscr{A}$ on a set $\Omega$. For an arbitrary subset $A \subseteq \Omega$ define the function

$$\mathbb{1}_A\colon \Omega \to \mathbb{R}, \qquad \mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{else.} \end{cases}$$

For each $B \in \mathfrak{B}(\mathbb{R})$ it follows that

$$\mathbb{1}_A^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B \\ A, & \text{if } 0 \notin B, 1 \in B \\ A^c, & \text{if } 0 \in B, 1 \notin B \\ \Omega, & \text{if } 0 \in B, 1 \in B \end{cases}$$

Consequently, $\mathbb{1}_A^{-1}(B) \in \mathscr{A}$ for all $B \in \mathfrak{B}(\mathbb{R})$ if and only if $A \in \mathscr{A}$, and it follows that the function $\mathbb{1}_A$ is measurable if and only if $A \in \mathscr{A}$.

The function $\mathbb{1}_A$ is called the *indicator function of the set* $A$.

A random variable $X\colon \Omega \to \mathbb{R}$ is required to be measurable in order to guarantee that the set $f^{-1}(B)$ is an element in the $\sigma$-algebra $\mathscr{A}$ on $\Omega$. Because only if the latter is satisfied, we can consider the probability $P\big(f^{-1}(B)\big)$, since the probability measure $P$ is only defined on $\mathscr{A}$.

It is useful to simplify some of our notations:

$$\begin{aligned} \{X \in B\} &:= X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \\ P(X \in B) &:= P(\{X \in B\}) = P\left(f^{-1}(B)\right). \end{aligned}$$

If $X\colon \Omega \to \mathbb{R}$ is a random variable we can define a new function

$$P_X\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P_X(B) := P(X \in B).$$

The function $P_X$ is well defined since $X\colon \Omega \to \mathbb{R}$ is measurable and thus $\{X \in B\} \in \mathscr{A}$. It turns out that $P_X$ is a new probability measure on $\mathfrak{B}(\mathbb{R})$, as it satisfies the conditions in Definition 1.2.10:

- $P_X(\mathbb{R}) = P(X \in \mathbb{R}) = P(\Omega) = 1$;
- For disjoint sets $B_1, B_2, \ldots \in \mathfrak{B}(\mathbb{R})$ one obtains by Lemma C.0.2:

$$P_X\left(\bigcup B_k\right) = P\left(X^{-1}\left(\bigcup B_k\right)\right) = P\left(\bigcup X^{-1}(B_k)\right) = \sum P\big(X^{-1}(B_k)\big) = \sum P_X(B_k).$$

Consequently, we can define:

**Definition 3.1.3.** *For a random variable $X\colon \Omega \to \mathbb{R}$ the probability measure*

$$P_X\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P_X(B) := P(X \in B)$$

*is called the* probability distribution of $X$ (under $P$).

Notations we introduced in Chapter 2 for probability measures on $\mathfrak{B}(\mathbb{R})$ are equally for random variables:

**Definition 3.1.4.** *Let $X\colon \Omega \to \mathbb{R}$ be a random variable.*

(a) *the random variable $X$ is called* discrete *if its probability distribution $P_X$ is discrete. If $\Gamma$ is the support of $P_X$ then we say that $X$ takes values in $\Gamma$.*

(b) *the random variable $X$ is called* absolutely continuously *if its probability distribution $P_X$ is absolutely continuous. We call the density $f$ of $P_X$ also the density of $X$.*

(c) *the* cumulative distribution function of $X$ *is defined by*

$$F_X\colon \mathbb{R} \to [0,1], \qquad F_X(x) = P_X\big((-\infty, x]\big) = P\big(X \leqslant x\big).$$

**Remark 3.1.5.** The examples of probability measures introduced in Chapter 2 are typcidally examples of probability distributions of random variables.

In fact, the core object in random experiments are always random variables. The underlying (original) probability space $(\Omega, \mathscr{A}, P)$ is often not known and one can only reasonably model the distribution of some random variables $X$. Nevertheless, the concept of probability spaces is essential, as $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), P_X)$ is a new probability space.

We can illustrate this by the following: assume we want to model the share price of Company A at noon tomorrow. Then this price depends on various things and in the end, we can conclude that the configuration of the whole world will influence the price. Thus, the share price is given by a figure $X(\omega)$, where $X$ is a mapping from the set $\Omega$ of all possible world's configurations, i.e. $X$ is a function from $\Omega$ to the reals. Now, it is impossible to model a reasonable probability measure $P$ of the world's configurations. However, we can come up with some reasonable models for the share price, i.e. a probability distribution of the random variable $X$ without actually knowing $P$.

Figure 3.1.: the whole world

**Remark 3.1.6.** Remark 3.1.5 typically refers to real-world complex application. In contrast, we often consider simple models assuming ideal conditions, e.g. throwing some dice, where we can explicitly define the underlying probability space $(\Omega, \mathscr{A}, P)$; see Section 1.1.

The following result gives an important reduction of the problem for checking measurability of a function.

**Theorem 3.1.7.** *A function $X\colon \Omega \to \mathbb{R}$ is measurable if and only if*

$$\{X \leqslant x\} \in \mathscr{A} \qquad \textit{for all } x \in \mathbb{R}\,.$$

Since $\{X \leqslant x\} = X^{-1}((-\infty, x])$, Theorem 3.1.7 reduces the problem of checking $f^{-1}(B) \in \mathscr{A}$ to only sets of the form $B = (-\infty, x]$ instead of for all sets in $\mathfrak{B}(\mathbb{R})$.

**Remark 3.1.8.** Theorem 3.1.7 is a special part of a general theorem. It says, that if $\mathfrak{B}(\mathbb{R}) = \sigma(\mathscr{C})$ for a class $\mathscr{C}$ of sets, then $X\colon \Omega \to \mathbb{R}$ is measurable if and only if

$$f^{-1}(C) \in \mathscr{A} \qquad \text{for all } C \in \mathscr{C},$$

Thus, in Theorem 3.1.7 the class $\mathscr{C} = \{(-\infty, x] : x \in \mathbb{R}\}$ is taken. Other possibilities for $\mathscr{C}$ are mentioned after Definition 1.2.8.

Random variables can be transformed to new functions by algebraic operations applied pointwise. By this we mean, if $X$, $Y\colon \Omega \to \mathbb{R}$ are random variables, i.e. in particular functions, then for example $Z(\omega) := X(\omega) + Y(\omega)$ defines a new function $Z\colon \Omega \to \mathbb{R}$. Analogously, we define $\max\{X, Y\}$ and so on. The following theorem guarantees that such algebraic operations results in random variables again.

**Corollary 3.1.9.** *Let $X$, $Y\colon \Omega \to \mathbb{R}$ be random variables and $\alpha \in \mathbb{R}$ a constant. Then the following are random variables:*

(a) $X + Y$;

(b) $\alpha X$;

(c) $YX$;

(d) $\max\{X, Y\}$;

(e) $\min\{X, Y\}$.

Let $X_n\colon \Omega \to \mathbb{R}$ be a random variable for each $n \in \mathbb{N}$. Then, as before, new functions are defined by for example

$$R\colon \Omega \to \mathbb{R}, \qquad R(\omega) := \sup\{X_n(\omega) : n \in \mathbb{N}\},$$

and we use the short hand notation $\sup_{n\in\mathbb{N}} X_n = R$. Analogously, one defines

$$S\colon \Omega \to \mathbb{R}, \qquad S(\omega) := \limsup_{n\to\infty} X_n(\omega).$$

Again, it turns out that these limiting operations result in measurable functions:

**Corollary 3.1.10.** *Let $X_n\colon \Omega \to \mathbb{R}$ be a random variable for each $n \in \mathbb{N}$.*

(a) *The following mappings are also random variables:*

$$\sup_{n\in\mathbb{N}} X_n, \qquad \inf_{n\in\mathbb{N}} X_n, \qquad \liminf_{n\to\infty} X_n, \qquad \limsup_{n\to\infty} X_n.$$

(b) *if there exists a function $X\colon \Omega \to \mathbb{R}$ such that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$, then $X$ is a random variable.*

We come to the special case that we consider $\Omega = \mathbb{R}$ and $\mathscr{A} = \mathfrak{B}(\mathbb{R})$, i.e. we consider functions $f\colon \mathbb{R} \to \mathbb{R}$ and their measurability. Note, that the following theorem does not make sense for an arbitrary $\sigma$-algebra $\mathscr{A}$ on a set $\Omega$, as one does not have necessarily a meaning of continuity.

**Corollary 3.1.11.** *A continuous function $f\colon \mathbb{R} \to \mathbb{R}$ is measurable.*

Also the composition of measurable functions turns out to be measurable.

**Corollary 3.1.12.** *If $X : \Omega \to \mathbb{R}$ is measurable and $f\colon \mathbb{R} \to \mathbb{R}$ is measurable, then*

$$f \circ X : \Omega \to \mathbb{R} \text{ is measurable.}$$

In the application of Corollary 3.1.12 one has to note the different $\sigma$-algebras:

$$X : \Omega \to \mathbb{R} \text{ is measureable } \iff X^{-1}(B) \in \mathscr{A} \text{ for all } B \in \mathfrak{B}(\mathbb{R})$$
$$g\colon \mathbb{R} \to \mathbb{R} \text{ is measureable } \iff g^{-1}(C) \in \mathfrak{B}(\mathbb{R}) \text{ for all } C \in \mathfrak{B}(\mathbb{R}).$$

## 3.2. $\sigma$-algebra generated by a random variable

An essential part of the definition of a random variable $X : \Omega \to \mathbb{R}$ is the $\sigma$-algebra $\mathscr{A}$ of $\Omega$. In this subsection, we reverse the approach and start with a set $\Omega$ and a function $X : \Omega \to \mathbb{R}$, without requiring that $X$ is measurable (we do not have any $\sigma$-algebra of $\Omega$). Can we always find a $\sigma$-algebra such that $X$ is measurable? The answer is yes: Exercise 1.1.9 shows that

$$\sigma(X) := \{X^{-1}(B) : B \in \mathfrak{B}(\mathbb{R})\}$$

is a $\sigma$-algebra. If we assume $\mathscr{A} = \sigma(X)$ then $X$ is measurable by the very definition of $\sigma(X)$.

**Definition 3.2.1.** *For a function $X : \Omega \to \mathbb{R}$ the collection of sets*

$$\sigma(X) := \{A \subseteq \Omega : X^{-1}(B) = A \text{ for some } B \in \mathfrak{B}(\mathbb{R})\},$$

*is called the* $\sigma$-algebra generated by $X$.

**Remark 3.2.2.** Similarly as in Definition 1.2.6, one can think of $\sigma(X)$ as the smallest $\sigma$-algebra of $\Omega$ which guarantees that the given function $X : \Omega \to \mathbb{R}$ is measurable.

**Example 3.2.3.** Let $A$ be a subset of $\Omega$ and consider the indicator function $\mathbb{1}_A : \Omega \to \mathbb{R}$ introduced in Example 3.1.2. In this case we obtain $\sigma(\mathbb{1}_A) = \{\emptyset, \Omega, A, A^c\}$.

**Example 3.2.4.** Let $\Omega$ be a set and assume that $\Omega = \cup_{k=1}^n A_k$ for some sets $A_1, \dots, A_n \subseteq \Omega$. Define a function by

$$X : \Omega \to \mathbb{R}, \qquad X(\omega) = \sum_{k=1}^n \alpha_k \mathbb{1}_{A_k}(\omega),$$

for some fixed numbers $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. In this case, we obtain

$$\sigma(X) = \left\{ \bigcup_{k\in I} A_k : I \subseteq \{1, \dots, n\} \right\}$$

since we have for each $B \in \mathfrak{B}(\mathbb{R})$ that

$$X^{-1}(B) = \bigcup_{k=1}^n \{A_k : \alpha_k \in B\}.$$

## 3.3. Almost sure (a.s.) properties

Recall that random variables $X$ and $Y$ are measurable functions on $\Omega$ and thus we would say that $X$ equals $Y$ if $X(\omega) = Y(\omega)$ for all $\omega \in \Omega$. However, given the probabilistic context, this definition is often too stringent.

**Definition 3.3.1.** *Let $(\Omega, \mathscr{A}, P)$ be a probability space and $X$ and $Y$ random variables. We say that $X$* equals $Y$ $P$-almost surely (a.s.)*, if*

$$P\big(\{\omega \in \Omega : X(\omega) = Y(\omega)\}\big) = 1.$$

*We often use the notation $X = Y$ $P$-a.s.*

A much weaker correspondence between random variables is the following:

**Definition 3.3.2.** *Let $(\Omega, \mathscr{A}, P)$ be a probability space and $X$ and $Y$ random variables. We say $X$ and $Y$ have the same distribution if*

$$P(X \in B) = P(Y \in B) \quad \textit{for all } B \in \mathfrak{B}(\mathbb{R}).$$

*We sometimes use the notation $X \overset{\mathfrak{D}}{=} Y$.*

**Example 3.3.3.** Let $X$ be a standard normally distributed random variable. Define $Y(\omega) := -X(\omega)$ for all $\omega \in \Omega$. Then $X$ and $Y$ have the same distribution, but $X(\omega) \neq Y(\omega)$ for all $\omega \in \Omega$ where $X(\omega) \neq 0$.

**Definition 3.3.4.** *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables $X_n\colon \Omega \to \mathbb{R}$* converges $P$-a.s. *to a random variable $X\colon \Omega \to \mathbb{R}$ if*

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

We will consider $P$-a.s. convergence of random variables in more details in Section 9.

## 3.4. Exercises

1. Let $\Omega$ be a set and $\mathscr{A} = \{\emptyset, \Omega\}$. Show that if $X\colon \Omega \to \mathbb{R}$ is a random variable then $X$ is constant.

2. Let $\Omega = \{1, 3, 5, 7, 9\}$ be equipped with the $\sigma$-algebra

$$\mathscr{A} = \{\emptyset, E, \{1,3\}, \{5\}\{7,9\}, \{1,3,5\}, \{5,7,9\}, \{1,3,7,9\}\}.$$

Define a function by

$$X\colon \Omega \to \mathbb{R}, \qquad X(x) = \begin{cases} 2, & \text{if } x \in \{1,3,5\}, \\ 0, & \text{if } x \in \{7,9\}. \end{cases}$$

(a) Show that $X$ is measurable.

(b) Determine the $\sigma$-algebra $\sigma(X)$ generated by $X$.

3. Let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables. Show that
   (a) $\{X > Y\} \in \mathscr{A}$.
   (b) $\{X = Y\} \in \mathscr{A}$.

4. Let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables. Assume that $Y = g(X)$ for a measurable function $g\colon \mathbb{R} \to \mathbb{R}$. Show that $\sigma(Y) \subseteq \sigma(X)$.

5. Let $X$ be a random variable with cumulative distribution function $F_X$. Show that the cumulative distribution function $F_Y$ of the random variable $Y := |X|$ is given by
$$F_Y\colon \mathbb{R} \to [0,1], \qquad F_Y(y) = F_X(y) - F_X(-y-)$$

6. Let $X$ be uniformly distributed on the interval $(-1,1)$ and define $Y := \max\{X, 0\}$. Show that the distribution function of $Y$ is given by
$$F_Y\colon \mathbb{R} \to [0,1], \qquad F_Y(y) = \begin{cases} 0, & \text{if } y < 0, \\ \frac{1+y}{2}, & \text{if } y \in [0,1), \\ 1, & \text{else.} \end{cases}$$

7. Let $(A_k)_{k=1}^n$ be a sequence of disjoint sets $A_k \in \mathscr{A}$ with $\Omega = \cup A_k$. Show that the following is equivalent for a function $X\colon \Omega \to \mathbb{R}$:
   (i) $X$ is measurable with respect to $\mathscr{A}$;
   (ii) $X\colon A_k \to \mathbb{R}$ is measurable with respect to $(\mathscr{A} \cap A_k)$ for all $k = 1, \dots, n$.

   Recall the definition of $\mathscr{A} \cap A_k$ in Exercise 1.3.6.

8. Does there exists a function $X\colon \Omega \to \mathbb{R}$ such that $X$ is not measurable but $X^2$ is measurable?

9. Let $\Omega = [0,1)$ and let $\mathscr{D} = \big\{[0, \frac{1}{2}), [\frac{1}{4}, 1)\big\}$.
   a) Determine the $\sigma$-algebra $\sigma(\mathscr{D})$ on $\Omega$ which is generated by $\mathscr{D}$.
   b) Define a function
$$X\colon \Omega \to \mathbb{R}, \qquad X(\omega) := \begin{cases} 2, & \text{if } \omega \in [0, \frac{1}{4}), \\ 3, & \text{if } \omega \in [\frac{1}{4}, \frac{1}{2}), \\ -1.1, & \text{else.} \end{cases}$$
   Is $X$ a $\sigma(\mathscr{D})$–measurable function? Justify your answer.
   c) Define a function
$$Y\colon \Omega \to \mathbb{R}, \qquad Y(\omega) = \omega^2.$$
   Is $Y$ a $\sigma(\mathscr{D})$–measurable function? Justify your answer.

d) $(*)$ Determine the smallest $\sigma$-algebra $\mathscr{F}$ of $\Omega$ such that the function $Y$ defined in part (c) is $\mathscr{F}$–measurable. Justify your answer.

10. $(*)$ For a given probability space $(\Omega, \mathscr{A}, P)$ let $(\Omega, \mathscr{A}', P')$ be the probability space defined in Exercise 1.3.12. Let $X$, $Y\colon \Omega \to \mathbb{R}$ be two functions with $X = Y$ $P$-a.s. Show that $X$ is $\mathscr{A}'$-measurable if and only if $Y$ is $\mathscr{A}'$-measurable.

    This result is the motivation for introducing the augmented probability space $(\Omega, \mathscr{A}', P')$ for the following reason. If $X = Y$ $P$-almost surely, it is naturally to expect that measurabilty of $X$ implies measurability of $Y$. Although this is true for measurability with respect to $\mathscr{A}'$ it is not true with respect to $\mathscr{A}$.

# 4 Integration

In this section, $(\Omega, \mathscr{A}, P)$ is always the underlying probability space. We will introduce a integral with respect to a probability measure for measurable functions.

## 4.1. The integral

In this section we define an integral with respect to probability measures. We pursue this in 3 steps: firstly, we define the integral for simple random variables, which are those random variables attaining only finitely many values. In the second step, we extend this integral for non-negative random variables, and the last step introduces the integral for arbitrary random variables.

**Definition 4.1.1.** *A function $X \colon \Omega \to \mathbb{R}_+$ is called* simple *if it is of the form*

$$X(\omega) = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}(\omega) \qquad \text{for all } \omega \in \Omega, \tag{4.1.1}$$

*where $\alpha_k \in \mathbb{R}_+$ and $A_k \in \mathscr{A}$ are pairwise disjoint sets for $k = 1, \dots, n$ and $\Omega = \cup_{k=1}^{n} A_k$. The set of all simple functions is denoted by $\mathscr{S}(\Omega, \mathscr{A})$.*

A simple function $X \colon \Omega \to \mathbb{R}_+$ is measurable and thus a random variable, according to Example 3.1.2 and Corollary 3.1.9. Definition 4.1.1 says that a random variable $X$ is simple if it attains only finitely many values in $\mathbb{R}$, since a random variable $X$ of the form (4.1.1) can be rewritten as

$$X(\omega) = \alpha_k \qquad \text{if } \omega \in A_k.$$

Motivated by the Riemann integral, there is only one reasonable definition of the integral for simple functions, if we interpret the probabilities as a "weight":

**Definition 4.1.2.** *For a simple random variable $X \in \mathscr{S}(\Omega, \mathscr{A})$ of the form (4.1.1) we define the $P$*-integral of $X$ *by*

$$\int_\Omega X(\omega)\, P(d\omega) := \sum_{k=1}^{n} \alpha_k P(A_k).$$

The representation of a simple function is not unique. However, one can show that two different representations of a simple function yields the same value of the $P$-integral of $f$; see Exercise 4.3.3. For $X$, $Y \in \mathscr{S}(\Omega, \mathscr{A})$ and $\alpha$, $\beta \in \mathbb{R}_+$ it is easy to check that

(1) $\alpha X(\cdot) + \beta Y(\cdot) \in \mathscr{S}(\Omega, \mathscr{A})$ and

$$\int_\Omega \big(\alpha X(\omega) + \beta Y(\omega)\big)\, P(d\omega) = \alpha \int_\Omega X(\omega)\, P(d\omega) + \beta \int_\Omega Y(\omega)\, P(d\omega). \tag{4.1.2}$$

(2) if $X(\omega) \leqslant Y(\omega)$ for all $\omega \in \Omega$ then

$$\int_\Omega X(\omega)\, P(d\omega) \leqslant \int_\Omega Y(\omega)\, P(d\omega). \tag{4.1.3}$$

(3) If $X = \mathbb{1}_A$ for some $A \in \mathscr{A}$ we have

$$\int_\Omega \mathbb{1}_A(\omega)\, P(d\omega) = P(A). \tag{4.1.4}$$

The main ingredient for our next step is the following fact, that each measurable real valued function can be approximated by a sequence of simple functions.

**Theorem 4.1.3.** *For every random variable $X\colon \Omega \to \mathbb{R}_+$ there exists an increasing sequence $(X_n)_{n\in\mathbb{N}}$ of simple random variables $X_n \in \mathscr{S}(\Omega, \mathscr{A})$ converging pointwise to $X$, that is*

$$X_n(\omega) \leqslant X_{n+1}(\omega) \quad \text{and} \quad X(\omega) = \lim_{n\to\infty} X_n(\omega) \quad \text{for all } \omega \in \Omega.$$

*Proof.* Define for each $\omega \in \Omega$ and $n \in \mathbb{N}$ the function

$$X_n(\omega) = \begin{cases} \frac{k}{2^n}, & \text{if } X(\omega) \in \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right), \ k = 0, \dots, n2^n - 1, \\ n, & \text{if } X(\omega) \geqslant n. \end{cases}$$

For each $n$ the function $X_n$ is simple and satisfies for all $\omega \in \Omega$:

$$\lim_{n\to\infty} X_n(\omega) = \sup_{n\in\mathbb{N}} X_n(\omega) = f(\omega),$$

which completes the proof. □

In order to define the $P$-integral for a larger class of functions we take limits. Here, we use the following fact: if $(\alpha_n)_{n\in\mathbb{N}}$ is an increasing sequence of numbers $\alpha_n \in \mathbb{R}_+$ then there exists an element $\alpha \in \mathbb{R}\cup\{\infty\}$ such that

$$\alpha = \lim_{n\to\infty} \alpha_n.$$

Note, that the limit might have the value $+\infty$ but nevertheless it is well defined. Let a non-negative random variable $X$ be approximated by a sequence $(X_n)$ of simple random variables according to Theorem 4.1.3. Thus, it follows from (4.1.3) that

$$\alpha_n := \int_\Omega X_n(\omega)\, P(d\omega)$$

defines an increasing sequence $(\alpha_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$, which must have a limit, possibly $= \infty$. This limit is now defined as the integral of $X$ in the following definition:

**Definition 4.1.4.** *For a random variable $X\colon \Omega \to \bar{\mathbb{R}}_+$ we define the $P$-integral by*

$$\int_\Omega X(\omega)\, P(d\omega) := \lim_{n\to\infty} \int_\Omega X_n(\omega)\, P(d\omega),$$

*where $(X_n)_{n\in\mathbb{N}} \subseteq \mathscr{S}(\Omega,\mathscr{A})$ is the sequence approximating $X$ according to Theorem 4.1.3.*

It remains to check that this definition of the $P$-integral does not depend on the approximating sequence of $X$, see Exercise 4.3.4. It follows from the construction that the $P$-integral is also linear as in (4.1.2) and monotone as in (4.1.3), that is if $X\colon \Omega \to \mathbb{R}_+$ and $Y\colon \Omega \to \mathbb{R}_+$ are non-negative random variables then we have

(1) for each $\alpha$, $\beta \in \mathbb{R}_+$ that

$$\int_\Omega \big(\alpha X(\omega) + \beta Y(\omega)\big)\, P(d\omega) = \alpha \int_\Omega X(\omega)\, P(d\omega) + \beta \int_\Omega Y(\omega)\, P(d\omega). \tag{4.1.5}$$

(2) if $X(\omega) \leqslant Y(\omega)$ for all $\omega \in \Omega$ then

$$\int_\Omega X(\omega)\, P(d\omega) \leqslant \int_\Omega Y(\omega)\, P(d\omega). \tag{4.1.6}$$

The final step is now to define the integral for arbitrary random variable $X\colon \Omega \to \mathbb{R}$. Define

$$X^+(\omega) := \max\{X(\omega), 0\}, \qquad X^-(\omega) := -\min\{X(\omega), 0\} \qquad \text{for all } \omega \in \Omega.$$

It follows from Corollary 3.1.9 that the functions $X^+$ and $X^-$ are again measurable and thus, their $P$-integrals are defined according to Definition 4.1.4. However, it may happen that both integrals are infinite, i.e.

$$\int_\Omega X^+(\omega)\, P(d\omega) = \infty \quad \text{and} \quad \int_\Omega X^-(\omega)\, P(d\omega) = \infty,$$

which is something we have to exclude in the following definition, as otherwise we could have $-\infty + \infty$ which is not defined. If we only consider a non-negative (or non-positive) random variable, when this is not a concern if we remember that the integral might be infinite.

**Definition 4.1.5.** *A random variable $X : \Omega \to \mathbb{R}$ is $P$*-integrable *if*

$$\int_\Omega X^+(\omega)\, P(d\omega) < \infty, \quad \int_\Omega X^-(\omega)\, P(d\omega) < \infty. \tag{4.1.7}$$

*In this case, the $P$-integral of $f$ is given by*

$$\int_\Omega X(\omega)\, P(d\omega) := \int_\Omega X^+(\omega)\, P(d\omega) - \int_\Omega X^-(\omega)\, P(d\omega).$$

*The space of all $P$-integrable random variables is denoted by $\mathscr{L}^1(\Omega, P)$.*

Note, for a non-negative random variable $X\colon \Omega \to \mathbb{R}_+$ the $P$-integral always exists according to Definition 4.1.4 but the random variable is only $P$-integrable if the $P$-integral is finite. A random variable $X : \Omega \to \bar{\mathbb{R}}$ satisfies the conditions in (4.1.7) if and only if

$$\int_\Omega |X(\omega)|\; P(d\omega) < \infty.$$

Thus,we have that $X \in \mathscr{L}^1(\Omega, P)$ if and only if $|X| \in \mathscr{L}^1(\Omega, P)$. In particular, we obtain

$$\mathscr{L}^1(\Omega, P) = \Big\{ X\colon \Omega \to \mathbb{R} \text{ random variable with } \int_\Omega |X(\omega)|\; dP(\omega) < \infty \Big\}. \tag{4.1.8}$$

If $X \in \mathscr{L}^1(\Omega, P)$ and $A \in \mathscr{A}$ then $X(\cdot)\mathbb{1}_A(\cdot)\colon \Omega \to \mathbb{R}$ is measurable according to Corollary 3.1.9, where $\mathbb{1}_A$ is the indicator function defined in Example 3.1.2. Since

$$\int_\Omega \mathbb{1}_A(\omega)\, |X(\omega)|\; P(d\omega) \leqslant \int_\Omega |X(\omega)|\; P(d\omega) < \infty,$$

it follows that $X(\cdot)\mathbb{1}_A(\cdot) \in \mathscr{L}^1(\Omega, P)$ for every $A \in \mathscr{A}$. Thus, we can define

$$\int_A X(\omega)\, P(d\omega) := \int_\Omega \mathbb{1}_A(\omega) X(\omega)\, P(d\omega) \qquad \text{for each } A \in \mathscr{A}.$$

We use several equivalent notations for the $P$-integral:

$$\int_\Omega X(\omega)\, P(d\omega), \quad \int X(\omega)\, dP(\omega), \quad \int X\, dP.$$

Recall Definition 3.3.1 for random variables to equal $P$-a.s.

**Theorem 4.1.6.** Properties of the $P$-integral
*For random variable $X$, $Y \in \mathscr{L}^1_P(\Omega)$ and $\alpha$, $\beta \in \mathbb{R}$ we have:*

(a) $\alpha X + \beta Y \in \mathscr{L}^1(\Omega, P)$ *and* $\int (\alpha X + \beta Y)\, dP = \alpha \int X\, dP + \beta \int Y\, dP$.

(b) *if* $0 \leqslant X \leqslant Y$ *then* $\int X\, dP \leqslant \int Y\, dP$.

(c) $\left| \int X\, dP \right| \leqslant \int |X|\; dP$.

(d) *$P\big(\{\omega \in \Omega : |X(\omega)| = 0\}\big) = 1$ if and only if $\int |X|\, dP = 0$.*

(e) *if $P(N) = 0$ for a set $N \in \mathscr{A}$ then $\int_N X\, dP = 0$.*

Part (d) shows that the integrand can be changed on a set $N \in \mathscr{A}$ with $P(N) = 0$ without affecting the value of the integral. More precisely, if $X, Y : \Omega \to \mathbb{R}$ are $P$-integrable and $Y$ obeys $Y(\omega) = 0$ for all $\omega \in N^c$ where $N$ is a $P$-null set, then it follows that

$$\int X\, dP = \int X\, dP + \int Y\, dP = \int (X + Y)\, dP.$$

**Example 4.1.7.** For some $\omega_0 \in \Omega$ let $P = \delta_{\omega_0}$, where $\delta_{\omega_0}$ denotes the Dirac measure introduced in Example 1.2.11. In this case, we obtain for every measurable function $X : \Omega \to \mathbb{R}$ that

$$\int X\, dP = X(\omega_0).$$

**Example 4.1.8.** Let $P$ be a discrete probability measure with support $\Gamma = \{\gamma_1, \gamma_2, \dots\}$. Theorem 2.2.2 guarantees that

$$P(A) = \sum_{k=1}^{\infty} p_k \delta_{\gamma_k}(A) \qquad \text{for all } A \in \mathscr{A}.$$

For a $P$-integrable random variable $X : \Omega \to \mathbb{R}$ we obtain

$$\int X\, dP = \sum_{k=0}^{\infty} p_k X(\gamma_k).$$

One of the most important advantages of the $P$-integral is that it enables to interchange the order of integration and taking a limit, whereas the classical Riemann integral is not amenable to such kind of results. The proofs of the following theorems are beyond the scope of these notes, but they and other versions of the results can be found in the pertinent literature.

**Theorem 4.1.9.** Monotone convergence theorem
*Assume that $(X_n)_{n\in\mathbb{N}}$ is an increasing sequence of non-negative random variables $X_n : \Omega \to \mathbb{R}_+$ and define $X(\omega) := \lim_{n\to\infty} X_n(\omega)$ for all $\omega$. It follows that*

$$\int X\, dP = \lim_{n\to\infty} \int X_n\, dP.$$

**Theorem 4.1.10.** Lebesgue's theorem of dominated convergence
*Assume that $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables $X_n : \Omega \to \mathbb{R}$ converging pointwise to a function $X : \Omega \to \mathbb{R}$, that is $X(\omega) = \lim X_n(\omega)$ for all $\omega \in \Omega$. If there exists an $P$-integrable random variable $Y : \Omega \to \mathbb{R}_+$ satisfying*

$$|X_n(\omega)| \leqslant Y(\omega) \qquad \textit{for all } \omega \in \Omega \textit{ and all } n \in \mathbb{N},$$

*then $X$ is $P$-integrable and satisfies*

$$\int X\, dP = \lim_{n\to\infty} \int X_n\, dP.$$

## 4.2. Expectation, variance and covariance

In this section we define the expectation of a random variable $X$ as the $P$-integral of $X$.

**Definition 4.2.1.** *Let $X\colon \Omega \to \mathbb{R}$ be a random variable. If $X \in \mathscr{L}^1(\Omega, P)$ then*

$$E[X] := \int_\Omega X(\omega)\, P(d\omega) \tag{4.2.9}$$

*is called the* (finite) expectation of $X$.

Note that Theorem 4.1.6 implies that the expectation is linear and monotone: if $X$ and $Y$ have finite expectations and $\alpha$, $\beta \in \mathbb{R}$ then:

(a) $\alpha X + \beta Y$ has finite expectation and $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$;

(b) if $X \leqslant Y$ then $E[X] \leqslant E[Y]$;

(c) $|E[X]| \leqslant E[|X|]$.

From (4.1) we obtain

(d) $E[\mathbb{1}_A] = P(A)$ for all $A \in \mathscr{A}$.

To calculate the expectation of a random variable, formula (4.2.9) would require knowing the probability measure $P$. However, as mentioned earlier, this is often not known, but only the probability distribution of the random variable; see Remark 3.1.5.

**Theorem 4.2.2.** Expectation rule
*Let $X\colon \Omega \to \mathbb{R}$ be a random variable with probability distribution $P_X$ and $h\colon \mathbb{R} \to \mathbb{R}$ be a measurable function. Then the following are equivalent:*

(a) *$h(X)$ has finite expectation;*

(b) *$h$ is $P_X$-integrable.*

*In this case, we have*

$$E[h(X)] = \int_\mathbb{R} h(x)\, P_X(dx).$$

**Corollary 4.2.3.** *Let $X\colon \Omega \to \mathbb{R}$ be a random variable in $\mathscr{L}^1(\Omega, P)$ with probability distribution $P_X$. Then we have*

$$E[X] = \int_\mathbb{R} x\, P_X(dx).$$

*Proof.* The claim follows from Theorem 4.2.2 by taking $h = \mathrm{Id}$. □

Corollary 4.2.3 in particular highlights that the expectation of a random variable $X\colon \Omega \to \mathbb{R}$ only depends on its probability distribution $P_X$ and not on its actual value $X(\omega)$ as a function; cf. Example 3.3.3.

**Corollary 4.2.4.** *Let $X\colon \Omega \to \mathbb{R}$ be a discrete random variable with $\Gamma = \{\gamma_1, \gamma_2, \dots\}$ support of $P_X$. For a measurable function $h\colon \mathbb{R} \to \mathbb{R}$ the following are equivalent:*

(a) *the random variable $h(X)\colon \Omega \to \mathbb{R}$ has finite expectation;*

(b) $\displaystyle\sum_{k=1}^{\infty} |h(\gamma_k)|\, P(X = \gamma_k) < \infty.$

*In this case, we have*

$$E\big[h(X)\big] = \sum_{k=0}^{\infty} h(\gamma_k) P(X = \gamma_k).$$

**Example 4.2.5.**

(a) If $X$ is uniform distributed on $E = \{1, \dots, n\}$ then $E[X] = \dfrac{n+1}{2}$.

(b) If $X$ is Bernoulli distributed with parameter $p$ then $E[X] = p$.

(c) If $X$ has a $\mathrm{Bin}(p, n)$ distribution then $E[X] = np$. This example of an expectation is calculated later.

(d) If $X$ has a $\mathrm{Pois}(\alpha)$ distribution then $E[X] = \alpha$.

(e) If $X$ has a $\mathrm{Geo}(\alpha)$ distribution then $E[X] = \dfrac{\alpha}{1-\alpha}$.

**Corollary 4.2.6.** *Let $X\colon \Omega \to \mathbb{R}$ be an absolutely continuous random variable with density $f\colon \mathbb{R} \to \mathbb{R}_+$ and let $h\colon \mathbb{R} \to \mathbb{R}$ be a measurable function. Then the following are equivalent:*

(a) *$h(X)$ has finite expectation;*

(b) $\displaystyle\int_{\mathbb{R}} |h(u)|\, f(u)\, \lambda(du) < \infty.$

*In this case, we have*

$$E[h(X)] = \int_{-\infty}^{\infty} h(u) f(u)\, du.$$

**Example 4.2.7.**

(a) If $X$ is uniformly distributed on $[a, b]$ then $E[X] = \frac{1}{2}(b - a)$.

(b) If $X$ is exponentially distributed with parameter $\alpha > 0$ then $E[X] = \frac{1}{\alpha}$.

(c) If $X$ is normally distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ then $E[X] = \mu$.

(d) If $X$ is log-normally distributed with parameters , then $E[X] = \exp(\mu + \frac{1}{2}\sigma^2)$.

**Theorem 4.2.8.** *(Cauchy-Schwarz inequality)*
*Let $X$ and $Y$ be random variables with $E[|X|^2] < \infty$ and $E[|Y|^2] < \infty$. Then we have*

$$E[|XY|] \leqslant \left(E[|X|^2]\right)^{1/2} \left(E[|Y|^2]\right)^{1/2}.$$

Cauchy-Schwarz inequality implies in particular, that if $E[|X|^2] < \infty$ then $E[|X|] < \infty$. For, take $Y = 1$ and $p = q = 2$, then it follows that

$$E[|X|] \leqslant \left(E[|X|^2]\right)^{1/2} \left(E[|1|^2]\right)^{1/2} = \left(E[|X|^2]\right)^{1/2} < \infty. \tag{4.2.10}$$

Random variables $X$ with $E[|X|^2] < \infty$ are often considered. In analogy to (4.1.8) we define

$$\mathscr{L}^2(\Omega, P) := \Big\{X : \Omega \to \mathbb{R} \text{ random variable with } \int_\Omega |X(\omega)|^2 \, P(d\omega) < \infty\}.$$

Note: the notation of $|X(\omega)|$ is redundant and one could just write $X^2(\omega)$. In particular, it follows from (4.2.10) that

$$\mathscr{L}^2(\Omega, P) \subseteq \mathscr{L}^1(\Omega, P). \tag{4.2.11}$$

**Definition 4.2.9.** *Let $X : \Omega \to \mathbb{R}$ be a random variable. If $E[X^2] < \infty$ then*

$$\operatorname{Var}[X] := \int_\Omega \big(X(\omega) - \mu\big)^2 \, P(d\omega)$$

*is called the* variance of $X$, *where $\mu := E[X]$.*

**Theorem 4.2.10.** Properties of the variance
*Let $X : \Omega \to \mathbb{R}$ be a random variable with $E[X^2] < \infty$. Then we have:*

(a) $\operatorname{Var}[\alpha X] = \alpha^2 \operatorname{Var}[X]$ *for all $\alpha \in \mathbb{R}$.*

(b) $\operatorname{Var}[X] = E[X^2] - (E[X])^2$.

(c) *if* $\operatorname{Var}[X] = 0$ *then there exists a constant $c \in \mathbb{R}$ with $P(X = c) = 1$.*

The proof of Theorem 4.2.10 shows that the constant $c$ in Part (c) can be chosen as $c = E[X]$.

For calculating the variance, we can apply Theorem 4.2.2 to the function $h(x) = (x - E[X])^2$ to obtain

$$\operatorname{Var}[X] = E[h(X)] = \int_{\mathbb{R}} \big(x - E[X]\big)^2 \, P_X(dx).$$

In particular in the cases of discrete or absolutely continuous random variables, we can apply Corollaries 4.2.4 and 4.2.6 to the same function $h$ to obtain the corresponding formula.

If $X$ is a discrete random variable we can use Corollary 4.2.4, possibly together with the equality $\operatorname{Var}[X] = E[X^2] - (E[X])^2$ to calculate the variances:

**Example 4.2.11.**

(a) If $X$ is uniform distributed on $\Gamma = \{1, 2, \dots, n\}$ then $\mathrm{Var}[X] = \dfrac{n^2-1}{12}$.

(b) If $X$ is Bernoulli distributed with parameter $p$ then $Var[X] = p(1-p)$.

(c) If $X$ has a $\mathrm{Bin}(p, n)$ distribution then $\mathrm{Var}[X] = np(1-p)$. We derive this later with a result in Chapter 6.

(d) If $X$ has a $\mathrm{Pois}(\alpha)$ distribution then $\mathrm{Var}[X] = \alpha$.

(e) If $X$ has a $\mathrm{Geo}(\alpha)$ distribution then $\mathrm{Var}[X] = \dfrac{\alpha}{(1-\alpha)^2}$.

If $X$ is an absolutely continuous distributed random variable, we can use Corollary 4.2.6, possibly together with the equality $\mathrm{Var}[X] = E[X^2] - (E[X])^2$ to calculate the variances:

**Example 4.2.12.**

(a) If $X$ is uniformly distributed on $[a, b]$ then $\mathrm{Var}[X] = \frac{1}{12}(b-a)^2$.

(b) If $X$ is exponentially distributed with parameter $\alpha > 0$ then $\mathrm{Var}[X] = \frac{1}{\alpha^2}$.

(c) If $X$ is normally distributed with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ then $\mathrm{Var}[X] = \sigma^2$.

(d) If $X$ is log-normally distributed with parameters , then $\mathrm{Var}[X] = \left(e^{\sigma^2} - 1\right) \exp^{2\mu+\sigma^2}$.

**Definition 4.2.13.** *Let $X$ and $Y$ be random variables with $E[X^2] < \infty$ and $E[Y^2] < \infty$.*

(a) *The* covariance of $X$ and $Y$ *is defined as*

$$\mathrm{Cov}(X, Y) := E\big[(X - E[X])(Y - E[Y])\big]$$

(b) *If* $\mathrm{Var}[X] > 0$ *and* $\mathrm{Var}[Y] > 0$*, then the* correlation coefficient of $X$ and $Y$ *is defined as*

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}.$$

Often, it is more convenient to work with the correlation as Cauchy-Schwarz inequality in Theorem 4.2.8 implies that $\mathrm{Corr}(X, Y) \in [-1, 1]$, and thus correlation is scaled (and free of a unit). The correlation measures the so-called *linear dependence* of the random variables $X$ and $Y$ due to the result below. Note, that if $\mathrm{Corr}(X, Y) = 0$, then $X$ and $Y$ are not necessarily independent (although we consider independence later in Chapter 6 I mention this popular mistake here).

**Lemma 4.2.14.** *Let $X$ and $Y$ be random variables with $E[X^2] < \infty$ and $E[Y^2] < \infty$. If $Corr(X, Y) = \pm 1$ then there exist two values $a, b \in \mathbb{R}$, where $a$ has the same sign as $Corr(X, Y) = \pm 1$, such that $P(X = aY + b) = 1$.*

Lemma 4.2.14 indicates that the correlation measures the linearity between two random variables with the two extreme cases of $\operatorname{Corr}(X, Y) = -1$ and $\operatorname{Corr}(X, Y) = 1$. Which values of the correlation are considered to be close to linearity heavily depends on the application. Although the correlation coefficient indicates the strength of linearity between two random variables, it does not completely characterise the relationship between the two random variables. The correlation is also called the *Pearson product-moment correlation coefficient.*

**Example 4.2.15.** The correlation is often used to analyse the relationship between two data sets $x_1, \dots, x_n$ and $y_1, \dots, y_n$, in which case the correlation is estimated by

$$r_{X,Y} := \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2}\,\sqrt{\sum_{k=1}^n (y_k - \bar{x})^2}},$$

where

$$\bar{x} = \frac{1}{n}\sum_{k=1}^n x_k, \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^n y_k.$$

Francis Anscombe:

## 4.3. Exercises

1. Let $X, Y : \Omega \to \mathbb{R}$ be random variables with $E[X^2] < \infty$ and $E[Y^2] < \infty$. Show that
   (a) $\operatorname{Cov}(X, Y) = E[XY] - E[X]E[Y]$.
   (b) $\operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X, Y)$.

2. *Standardised random variable*
   For a random variable $X : \Omega \to \mathbb{R}$ with $E[X^2] < \infty$ define

   $$X^* := \frac{X - E[X]}{\sqrt{\operatorname{Var} X}}.$$

   Show that $E[X^*] = 0$ and $\operatorname{Var}[X^*] = 1$.

3. Assume a simple random variable $X \in \mathscr{S}(\Omega, \mathscr{A})$ has different presentations:

   $$X(\omega) = \sum_{k=1}^n \alpha_k \mathbb{1}_{A_k}(\omega) = \sum_{k=1}^m \beta_k \mathbb{1}_{B_k}(\omega) \qquad \text{for all } \omega \in \Omega.$$

   Show that the $P$-integral does not depend on the presentation, i.e.

   $$\sum_{k=1}^n \alpha_k P(A_k) = \sum_{k=1}^m \beta_k P(B_k).$$

4. (**) Let $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$ be two increasing sequences of simple functions in $\mathscr{S}(\Omega, \mathscr{A})$.
   (a) Show that each $Z \in \mathscr{S}(\Omega, \mathscr{A})$ with $Z(\omega) \leqslant \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$ obeys
   $$\int_\Omega Z(\omega)\, P(d\omega) \leqslant \lim_{n\to\infty} \int_\Omega X_n(\omega)\, P(d\omega).$$
   (b) Show that if $\lim\limits_{n\to\infty} X_n(\omega) = \lim\limits_{n\to\infty} Y_n(\omega)$ for all $\omega \in \Omega$ then
   $$\lim_{n\to\infty} \int_\Omega X_n(\omega)\, P(d\omega) = \lim_{n\to\infty} \int_\Omega Y_n(\omega)\, P(d\omega).$$

5. Show that the sequence $(X_n)_{n\in\mathbb{N}}$ of simple functions defined in the proof of Theorem 4.1.3 satisfies for each $\omega \in \Omega$ that $X_n(\omega) \leqslant X_{n+1}(\omega)$ for all $n \in \mathbb{N}$ and $X_n(\omega) \to X(\omega)$ for $n \to \infty$.

6. (Important later for the risk neutral measure)
   Assume that $X : \Omega \to \mathbb{R}$ is a random variable with $P(X > 0) = 1$ and $E[X] = 1$. Define a mapping by
   $$Q : \mathscr{A} \to \mathbb{R}, \qquad Q(A) = E[\mathbb{1}_A X].$$
   (a) Show that $Q$ is a probability measure on $(\Omega, \mathscr{A})$.
   (b) Show that for each $A \in \mathscr{A}$ we have the equivalence:
   $$P(A) = 0 \;\Leftrightarrow\; Q(A) = 0.$$
   (c) Show that a random variable $Y$ is $Q$-integrable if and only if $YX$ is $P$-integrable. In this case, it follows $E_Q[Y] = E[YX]$, where $E_Q[Y]$ denotes the expectation w.r.t. $Q$.
   (d) Show that the random variable $1/X$ is $Q$-integrable.
   (e) Define a mapping by
   $$R : \mathscr{A} \to \mathbb{R}, \qquad R(A) = \int_\Omega \mathbb{1}_A(\omega) \frac{1}{X(\omega)}\, Q(d\omega).$$
   Conclude that $P = R$.
   (f) Is the equivalence in (b) still true, if we have only $P(X \geqslant 0) = 1$?

   In fact, the answer to $f$ is negative. This is the motivation to require that a risk-neutral measure is equivalent: negligible events under the "real-world measure $P$" should also be negligible under the risk neutral measure and vice-versa.

7. Let $X : \Omega \to \mathbb{R}_+$ be a non-negative random variable. Show that $E[X] < \infty$ implies $P(X < \infty) = 1$.

8. a) Let $X_k\colon \Omega \to \mathbb{R}_+$ be a non-negative random variable for each $k \in \mathbb{N}$. Show that

$$E\left[\sum_{k=1}^{\infty} X_k\right] = \sum_{k=1}^{\infty} E[X_k]. \tag{4.3.12}$$

   b) Let $X_k\colon \Omega \to \mathbb{R}$, $k \in \mathbb{N}$ be random variables satisfying

$$\sum_{k=1}^{\infty} E\left[|X_k|\right] < \infty.$$

   (i) Show that $\sum_{k=1}^{\infty} X_k$ converges $P$-a.s.

   (ii) Show that the sum $\sum_{k=1}^{\infty} X_k$ has finite expectation.

   (iii) Show that in this case the sum also obeys (4.3.12).

9. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with finite expectation. Show that if $P(A_k) \to 0$ for a sequence $\{A_k\}_{k\in\mathbb{N}} \subseteq \mathscr{A}$ then $E[\mathbb{1}_{A_k} X] \to 0$.

10. $(*)$ Let $(\Omega, \mathscr{A}, P)$ be a probability space and $Y\colon \Omega \to \mathbb{R}$ a random variable. Assume that $Z\colon \Omega \to \mathbb{R}$ is $\sigma(Y)$-measurable. Then there exists a measurable function $g\colon \mathbb{R} \to \mathbb{R}$ such that $Z = g(Y)$.

    (Hint: follow the classical 3 cases: simple, non-negative, arbitrary $Z$.)

# 5 Random vectors

In this chapter we extend some of the previously introduced concepts to random vectors. Most definitions are a straightforward extension of the one dimensional case and the corresponding results can be proved analogously as before.

## 5.1. Random vectors

To cover the multi-dimensional case, we first adapt Definition 3.1.1. This is achieved by just replacing the Borel $\sigma$-algebra $\mathfrak{B}(\mathbb{R})$ of $\mathbb{R}$ by its multi-dimensional analog $\mathfrak{B}(\mathbb{R}^d)$ introduced in Definition 1.2.9.

**Definition 5.1.1.** *Let $\mathscr{A}$ be a $\sigma$-algebra on a set $\Omega$.*

(a) *A function $f\colon \Omega \to \mathbb{R}^d$ is called* -measurable *if*

$$f^{-1}(B) := \{\omega \in \Omega : f(\omega) \in B\} \in \mathscr{A} \qquad \text{for all sets } B \in \mathfrak{B}(\mathbb{R}^d). \tag{5.1.1}$$

(b) *A function $X\colon \Omega \to \mathbb{R}^d$ is a* random vector *if it is measurable.*

A multivariate function $X\colon \Omega \to \mathbb{R}^d$ always consists of $d$ entries, i.e. $X = (X_1, \dots, X_d)$ with $X_i\colon \Omega \to \mathbb{R}$ for $i = 1, \dots, d$. Checking measurability reduces to the one-dimensional case:

**Theorem 5.1.2.** *A function $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ is a random vector if and only if $X_i\colon \Omega \to \mathbb{R}$ is a random variable for all $i = 1, \dots, d$.*

As before, the definition of the probability distribution can be extended to the multi-dimensional case by just replacing the Borel $\sigma$-algebra $\mathfrak{B}(\mathbb{R})$ by its multi-dimensional analog $\mathfrak{B}(\mathbb{R}^d)$. The multi-dimensionality let us consider the distribution of the $i$-th entry.

**Definition 5.1.3.** *Let $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ be a random vector.*

(a) *The* probability distribution of $X$ (under $P$) *is defined by*

$$P_X\colon \mathfrak{B}(\mathbb{R}^d) \to [0,1], \qquad P_X(B) := P(X \in B).$$

(b) *The $i$-th marginal probability distribution of $X$ is defined by*

$$P_{X_i}\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P_{X_i}(B) := P(X_i \in B).$$

Theorem 5.1.2 guarantees that if $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ is a random vector then $X_i\colon \Omega \to \mathbb{R}$ is a random variable. Thus, the $i$-th marginal distribution $P_{X_i}$ is nothing else than the probability distribution of the random variable $X_i$. From the probability distribution $P_X$ of the random vector $X\colon \Omega \to \mathbb{R}^d$ we always obtain the $i$-th marginal distribution $P_{X_i}$ in the following way:

$$\begin{aligned} P_{X_i}(B) &= P(X_i \in B) \\ &= P(X_1, \in \mathbb{R}, \dots, X_{i-1} \in \mathbb{R}, X_i \in B, X_{i+1} \in \mathbb{R}, \dots, X_d \in \mathbb{R}) \\ &= P_X(\mathbb{R} \times \mathbb{R} \times B \times \mathbb{R} \times \mathbb{R}) \end{aligned}$$

for all $B \in \mathfrak{B}(\mathbb{R})$. The converse direction is not possible: one cannot construct in general the probability distribution $P_X$ of a random vector $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ by its marginal distributions $P_{X_1}, \dots, P_{X_d}$, see Example 5.1.6.

Discrete and absolutely continuous probability measures on $\mathfrak{B}(\mathbb{R}^d)$ are defined analogously as in Definition 2.2.1 and 2.3.1. The notations are extended to random vectors by relating these properties to the probability distribution as in Definition 3.1.4 before. We do not extend the notion of the cumulative distribution function to the multi-dimensional setting. Although this is possible, the multi-dimensional analogue is usually less easy to work with than in the one-dimensional setting.

**Definition 5.1.4.**

(a) *A probability measure $P$ on $\mathfrak{B}(\mathbb{R}^d)$ is called* discrete *if there exists a countable set $\Gamma \subseteq \mathbb{R}^d$ such that $P(\Gamma) = 1$. The set $\Gamma$ is called support of $P$.*

(b) *A probability measure $P$ on $\mathfrak{B}(\mathbb{R}^d)$ is called* absolutely continuous *if there exists a (Riemann) integrable function $f\colon \mathbb{R}^d \to \mathbb{R}_+$ such that*

$$P\big((a_1, b_1], \dots, (a_d, b_d]\big) = \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(u_1, \dots, u_d)\, du_1 \cdots du_d,$$

*for all $-\infty < a_i < b_i < \infty$ and $i = 1, \dots, d$. In this case, $f$ is called the density of $P$.*

(c) *A random vector $X\colon \Omega \to \mathbb{R}^d$ is called discrete if its probability distribution $P_X$ is discrete. If $\Gamma$ is the support of $P_X$ then we say that $X$ takes values in $\Gamma$.*

(d) *A random vector $X\colon \Omega \to \mathbb{R}^d$ is called absolutely continuous if $f$ of $P_X$ is also called density of $X$.*

**Theorem 5.1.5.** *Let $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ be a random vector.*

(a) *the random vector $X$ is discrete if and only if each $X_i$ is discrete for $i = 1, \dots, d$.*

(b) *if the random vector $X$ is absolutely continuous with density $f\colon \mathbb{R}^d \to \mathbb{R}_+$ then each $X_i$ is absolutely continuously distributed with density $f_{X_i}\colon \mathbb{R} \to \mathbb{R}_+$ defined by*

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(u_1, \dots, u_{i-1}, x, u_{i+1}, \dots, u_d)\, du_1 \cdots du_{i-1} du_{i+1} \cdots du_d.$$

In Part (b) the important assumption is that the random vector $X$ is absolutely continuous. There are examples of absolutely random variables $X$ and $Y$ such that the random vector $(X, Y)$ is not absolutely continuous. We show this case in Example 7.1.12.

The following example shows that the probability distribution $P_X$ of a random vector $X = (X_1, X_2)\colon \Omega \to \mathbb{R}^2$ cannot be derived from the marginal distributions $P_{X_1}$ and $P_{X_2}$.

**Example 5.1.6.** (*Urn with red and black balls*)
One draws two balls from an urn with $r$ red and $b$ black balls without replacing them. The state space can be described by

$$\Omega := \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \dots, r+b\}, \omega_1 \neq \omega_2\},$$

where the numbers $1, \dots, r$ corresponds to the red balls and the numbers $r+1, \dots, r+b$ to the black balls. We assume that each outcome is equally likely, that is we assume a uniform distribution on $\Omega$ (see Subsection 5.2.1):

$$P(\{(\omega_1, \omega_2)\}) = \frac{1}{|\Omega|} = \frac{1}{(r+b)(r+b-1)} \qquad \text{for all } (\omega_1, \omega_2) \in \Omega.$$

Here, we obtain the cardinality of $\Omega$ by counting the possibilities of the first ball $(= r+b)$ multiplied by the possibilities of the second ball $(= r+b-1)$. Define the random variables

$$U : \Omega \to \{1, 0\}, \qquad U((\omega_1, \omega_2)) = \begin{cases} 1, & \text{if } \omega_1 \leqslant r, \\ 0, & \text{else.} \end{cases},$$

$$V : \Omega \to \{1, 0\}, \qquad V((\omega_1, \omega_2)) = \begin{cases} 1, & \text{if } \omega_2 \leqslant r, \\ 0, & \text{else.} \end{cases}.$$

Thus, the random variable $U$ indicates the colour of the first ball and $V$ indicates the colour of the second ball. It follows from Theorem 5.1.5 that $X := (U, V)\colon \Omega \to \mathbb{R}^2$ defines a discrete random vector. The probability distribution of $X$ and the marginal distributions are given in the following table:

| $P_X(\quad)$ | $\{V=1\}$ | $\{V=0\}$ | |
|---|---|---|---|
| $\{U=1\}$ | $\frac{r(r-1)}{(r+b)(r+b-1)}$ | $\frac{rb}{(r+b)(r+b-1)}$ | $\frac{r(r+b-1)}{(r+b)(r+b-1)}$ |
| $\{U=0\}$ | $\frac{rb}{(r+b)(r+b-1)}$ | $\frac{b(b-1)}{(r+b)(r+b-1)}$ | $\frac{b(r+b-1)}{(r+b)(r+b-1)}$ |
| | $\frac{r(r+b-1)}{(r+b)(r+b-1)}$ | $\frac{b(r+b-1)}{(r+b)(r+b-1)}$ | 1 |

The cardinality can be derived as above for $\Omega$. For example, the event $\{U=1,\, V=0\}$ has cardinality $rb$ since the first ball has $r$ possibilities and the second ball has $b$ possibilities. The last column gives the marginal distributions of $U$, i.e. $P(U=1)$ and $P(U=0)$, and the last row gives the marginal distribution of $V$,i.e. $P(V=1)$ and $P(V=0)$. These entries are just the sum of the corresponding row and column, respectively.

For constructing examples of multivariate distributions, we extend Theorem 2.2.2 and Theorem 2.3.3 to the multi-dimensional setting. We restrict ourselves to the two-dimensional case for easier presentation.

**Theorem 5.1.7.** *Let $P\colon \mathfrak{B}(\mathbb{R}^2) \to [0,1]$ be a function. Then the following is equivalent:*

(a) *$P$ is a discrete probability measure;*

(b) *there are numbers $\gamma_1, \gamma_2, \ldots \in \mathbb{R}$ and $\delta_1, \delta_2, \ldots \in \mathbb{R}$ and $p_{k,\ell} \geqslant 0$ with $\sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} p_{k,\ell} = 1$ and*

$$P(A) = \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} p_{k,\ell}\delta_{\{(\gamma_k,\delta_\ell)\}}(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}^2). \tag{5.1.2}$$

*In this case, it follows that $P(\{(\gamma_k,\delta_\ell)\}) = p_{k,\ell}$ for all $k \in \mathbb{N}$.*

*Proof.* Analogue to Theorem 2.2.2. □

The Dirac measure $\delta_{\{(\gamma_k,\delta_\ell)\}}$ is defined according to Example 1.2.11 by seting $\Omega = \mathbb{R}^2$ and $\omega_0 = (\gamma_k, \delta_\ell)$.

**Theorem 5.1.8.** *If $f\colon \mathbb{R}^2 \to \mathbb{R}_+$ is a (Riemann-integrable) function with*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u_1,u_2)\,du_1du_2 = 1,$$

*then there exists an absolutely continuous probability measure $P$ on $\mathfrak{B}(\mathbb{R}^2)$ satisfying*

$$P\big((a_1,b_1]\times(a_2,b_2]\big) = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f(u)\,du_1du_2$$

*for all $-\infty < a_1 < b_1 < \infty$ and $-\infty < a_2 < b_2 < \infty$. That is, $f$ is the density of $P$.*

*Proof.* Analogue to Theorem 2.3.3. □

## 5.2. Examples of multivariate distributions

### 5.2.1. Uniform distribution (discrete)

Let $\Gamma = \{\gamma_1, \dots, \gamma_m\} \times \{\delta_1, \dots, \delta_n\}$ be a finite subset of $\mathbb{R}^2$. Each element $(\gamma_k, \delta_\ell)$ of $\Gamma$ is assigned the same probability. If $\Gamma$ is the support of a discrete probability measure on $\mathfrak{B}(\mathbb{R}^2)$ and each $P(\{(\gamma_k, \delta_l)\})$ is required to have the same probability we obtain

$$
\begin{aligned}
P(\Gamma) = 1 \iff & P\left(\bigcup_{k=1}^{m} \bigcup_{\ell=1}^{n} \{(\gamma_k, \delta_\ell)\}\right) = 1 \\
\iff & \sum_{k=1}^{m} \sum_{\ell=1}^{n} P(\{(\gamma_k, \delta_\ell)\}) = 1 \\
\iff & P(\{(\gamma_k, \delta_\ell)\}) = \frac{1}{mn} \quad \text{for all } k \in \{1, \dots, m\}, \ell \in \{1, \dots, n\}.
\end{aligned}
$$

Setting $p_{k,\ell} = \frac{1}{mn}$ for all $k \in \{1, \dots, n\}$ and $\ell \in \{1, \dots, n\}$, Theorem 5.1.7 implies that

$$
P\colon \mathfrak{B}(\mathbb{R}^2) \to [0,1], \quad P(A) = \sum_{k=1}^{m} \sum_{\ell=1}^{n} \frac{1}{mn} \delta_{\{(\gamma_k, \delta_\ell)\}}(A), \tag{5.2.3}
$$

defines a discrete probability measure with support $\Gamma$, which is called the *uniform distribution on* $\Gamma$. The probability measure can be simplified as

$$
\sum_{k=1}^{m} \sum_{\ell=1}^{n} \frac{1}{mn} \delta_{\{(\gamma_k, \delta_\ell)\}}(A) = \frac{|A \cap \Gamma|}{|\Gamma|},
$$

where $|A \cap \Gamma|$ and $|\Gamma|$ denotes the cardinality of the sets $A \cap \Gamma$ and $\Gamma$.

**Example 5.2.1.** (Continues Example 1.1.3.b)

The outcome of rolling a dice twice can be described by $\Gamma := \{1, \dots, 6\} \times \{1, \dots, 6\}$. Assuming equal probabilities for all elements in $\Gamma$ we obtain $P(\{(k, \ell)\}) = \frac{1}{36}$ for all $(k, \ell) \in \Gamma$. Consequently, for each $A \in \mathfrak{B}(\mathbb{R}^2)$ we have

$$
P(A) = \frac{1}{36} \sum_{k=1}^{6} \sum_{\ell=1}^{6} \delta_{\{(k,\ell)\}}(A) = \frac{|A \cap \Gamma|}{36}.
$$

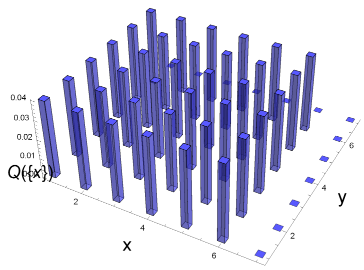



Figure 5.1.: uniform distribution on $\Gamma$

For the event $A$ that the sum of the two dice equals 5 (cf. Example 1.1.3.b), we obtain accordingly that

$$
P(A) = \frac{|(1,4),(2,3),(3,2),4,1)|}{36} = \frac{4}{36}
$$

Formally, we have defined the probability measure $P$ on $\mathfrak{B}(\mathbb{R})$.

As before, one can define the uniform distribution on a $d$-dimensional, finite subset $\Gamma$ of $\mathbb{R}^d$. This results in $d$ sums in the expression corresponding to (5.2.3), which again can be simplified to

$$P\colon \mathfrak{B}(\mathbb{R}^d) \to [0,1], \quad P(A) = \frac{|A \cap \Gamma|}{|\Gamma|}. \tag{5.2.4}$$

**Remark 5.2.2.** If $P$ is the uniform distribution on a set $\Gamma$, then we have $P(A) = 0$ for each set $A \in \mathfrak{B}(\mathbb{R}^d)$ with $A \cap \Gamma = \emptyset$ due to (5.2.4). For this reson, we also say that $P$ is the uniform distribution on $\Gamma$ and consider $P$ as a mapping on the $\sigma$-algebra $\mathfrak{B}(\mathbb{R}^d) \cap \Gamma$.

**Example 5.2.3.** Birthdays
What is the probability that among $n = 25$ students two of them have birthday on the same day? Define $I := \{1, \dots, 365\}$ and the state space by

$$\begin{aligned} \Gamma : &= I \times I \times \cdots \times I \\ &= \{(\gamma_1, \dots, \gamma_{25}) : \gamma_i \in I \text{ for } i = 1, \dots, 25\}. \end{aligned}$$

We can assume a uniform distribution on $\Gamma$, thus we have

$$P\colon \mathfrak{B}(\mathbb{R}^{25}) \to [0,1], \qquad P(A) = \frac{|A \cap \Gamma|}{|\Gamma|}.$$

In order to calculate the required probability it is enough to determine the cardinality of the set $A$, where

$$A = \{(\gamma_1, \dots, \gamma_{25}) \in \Gamma : \gamma_i = \gamma_j \text{ for some } i \neq j\}.$$

It is easier to obtain the cardinality of the complement $A^c$, which can be accomplished by some combinatorics based on urn models (not part of this course):

$$P(A) = 1 - P(A^c) = 1 - \frac{365!}{340!} \frac{1}{365^{25}} \approx 0.568.$$

### 5.2.2. Uniform distribution (absolutely continuous)

Let $S \subseteq \mathbb{R}^2$ be a set with finite area $a(S)$, e.g. the rectangle $S = [1,2] \times [0,2]$ with $a(S) = 2$ or the circle $S = \{u = (u_1, u_2) \in \mathbb{R}^2 : u_1^2 + u_2^2 \leqslant 1\}$ with $a(S) = 2\pi$. Define a function

$$f\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f(u_1, u_2) = \frac{1}{a(S)} \mathbb{1}_S(u_1, u_2).$$

Since

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u_1, u_2)\, du_1 du_2 = \frac{1}{a(S)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}_S(u_1, u_2)\, du_1 du_2 = 1,$$

Theorem 5.1.8 implies that $f$ defines a probability measure with

$$P\big((a_1, b_2] \times (a_2, b_2]\big) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(u_1, u_2)\, du_1 du_2.$$

for all $a_1 \leqslant b_2, a_2 \leqslant b_2$. Obviously, we have

$$P\big((a_1, b_2] \times (a_2, b_2]\big) = \frac{1}{a(S)} \int_{a_1}^{b_1} \int_{a_2}^{b_2} du_1 du_2 = \frac{\text{area of } \Big(\big((a_1, b_2] \times (a_2, b_2]\big) \cap S\Big)}{a(S)}.$$

With a little more mathematics, this formula can be generalised to all sets $A \in \mathfrak{B}(\mathbb{R})$ and not only rectangles that is $P(A)$ =area of $A/a(S)$ for all $A \in \mathfrak{B}(\mathbb{R})$ and $A \subseteq S$.

### 5.2.3. Multivariate normal distribution

In Subsection 2.3.3 we introduce the normal distribution in $\mathbb{R}$. The following extend this definition to the multidimensional setting. Definition D.0.5 recalls the notation of symmetric and positive-definite matrices.

**Definition 5.2.4.** *For a vector $\mu \in \mathbb{R}^d$ and a positive definite, symmetric matrix $\Sigma$ in $\mathbb{R}^{d\times d}$ define*

$$\psi^{(d)}_{\mu,\Sigma}\colon \mathbb{R}^d \to \mathbb{R}_+, \qquad \psi^{(d)}_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{det\,\Sigma}} \exp\left(-\frac{1}{2}\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle\right)$$

*where $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbb{R}^d$. Then the probability measure corresponding to the density $\psi^{(d)}_{\mu,\Sigma}$ is called the* normal distribution with expectation $\mu$ and covariance matrix $\Sigma$. *We use the notation $N(\mu, \Sigma)$.*

A random vector $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ which is distributed according a $N(\mu, \Sigma)$-distribution for some $\mu = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$ and $\Sigma = (s_{ij})^d_{i,j=1} \in \mathbb{R}^{d\times d}$ satisfies

$$E[X_i] = \mu_i, \qquad \operatorname{Cov}(X_i, X_j) = s_{ij} \qquad \text{for all } i, j = 1, \dots, d. \tag{5.2.5}$$

For this reason the parameter $\mu$ is called the *expectation vector* and $\Sigma$ is called the *covariance matrix of the distribution* $N(\mu, \Sigma)$. We show this fact in the following example for the case $d = 2$:

**Example 5.2.5.** Assume that $(X, Y)\colon \Omega \to \mathbb{R}^2$ has a normal distribution $N(\mu, \Sigma)$ with $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$ of the form

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} s_{1,1} & s_{1,2} \\ s_{2,1} & s_{2,2} \end{pmatrix}.$$

That is, the probability distribution $P_X$ of $X$ is a normal distribution with parameters $\mu$ and $\Sigma$. It is convenient to represent the matrix $\Sigma$ in the form

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\varrho \\ \sigma_1\sigma_2\varrho & \sigma_2^2 \end{pmatrix}.$$

for some $\sigma_1, \sigma_2 > 0$. Since $\det[\Sigma] = \sigma_1^2\sigma_2^2(1-\varrho^2)$ and $\det[\Sigma] > 0$ it follows that we can assume that $\sigma_1, \sigma_2$ are positive and that $\varrho \in (-1, 1)$. The inverse of $\Sigma$ is given by

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2(1-\varrho^2)} & \frac{-\varrho}{\sigma_1\sigma_2(1-\varrho^2)} \\ \frac{-\varrho}{\sigma_1\sigma_2(1-\varrho^2)} & \frac{1}{\sigma_2^2(1-\varrho^2)} \end{pmatrix}.$$

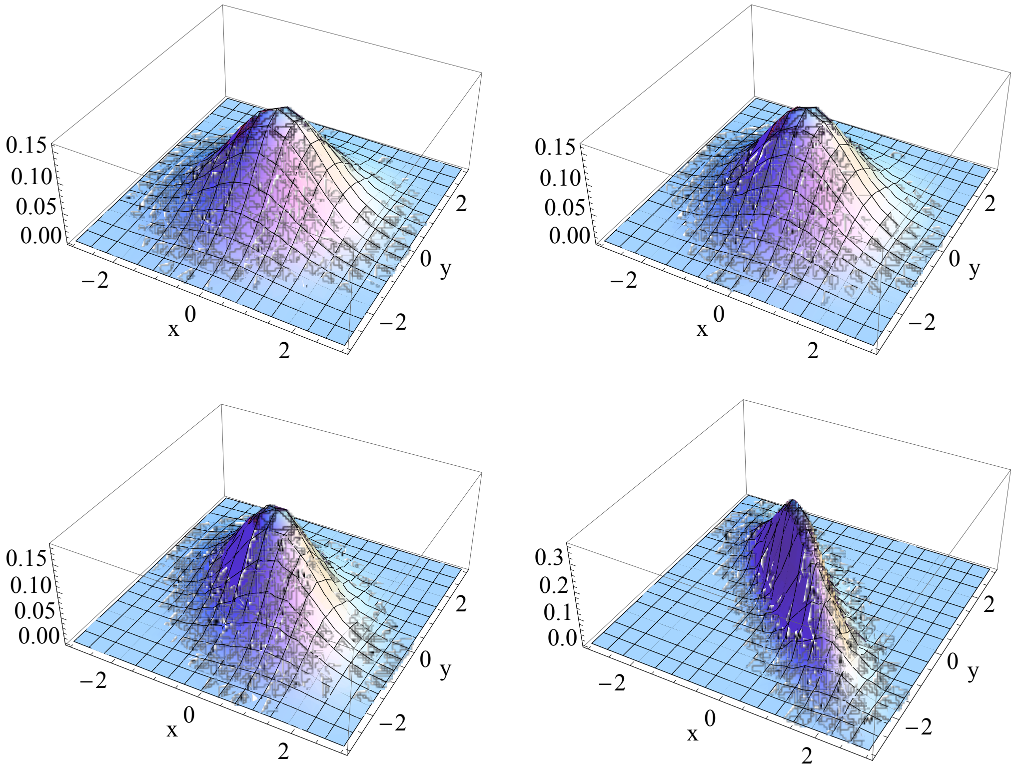



Figure 5.2.: $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and $s_{1,2} = s_{2,1} = 0, -0.2, -0.5, -0.9$

Substituting this in the density given in Definition 5.2.4 we obtain for each $(x, y) \in \mathbb{R}^2$ that

$$\begin{aligned}
&\psi_{\mu,\Sigma}(x, y)\\
&\quad = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\varrho^2)}} \exp\left(-\frac{1}{2(1-\varrho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\varrho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right).
\end{aligned}$$

By completing the square one establishes that

$$\int_{\mathbb{R}} e^{-au^2+2bu}\, du = \sqrt{\frac{\pi}{a}}\, e^{b^2/a} \qquad \text{for all } a > 0,\ b \in \mathbb{R}\,.$$

By using this identity, Theorem 5.1.5 implies that the density $f_X$ of $X$ is given by

$$\begin{aligned}
&f_X(x)\\
&= \int_{\mathbb{R}} \psi_{\mu,\Sigma}(x, y)\, dy\\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2(1-\varrho^2)}\right) \int_{\mathbb{R}} \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2(1-\varrho^2)} + \varrho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2(1-\varrho^2)}\right) dy
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2(1-\varrho^2)}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma_2^2(1-\varrho^2)}\left(y^2 + 2\varrho\frac{\sigma_2}{\sigma_1}(x-\mu_1)y\right)\right) dy \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2(1-\varrho^2)}\right) \exp\left(\frac{1}{2\sigma_2(1-\varrho^2)}\frac{\sigma_2^2}{\sigma_1^2}\varrho^2(x-\mu_1)^2\right) \\
&\quad \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma_2^2(1-\varrho^2)}\left(y^2 - 2\varrho\frac{\sigma_2}{\sigma_1}(x-\mu_1)y + \frac{\sigma_2^2}{\sigma_1^2}\varrho^2(x-\mu_1)^2\right)\right) dy \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma_2^2(1-\varrho^2)}\left(y - \frac{\sigma_2}{\sigma_1}\varrho(x-\mu_1)\right)^2\right) dy \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right).
\end{aligned}
$$

Thus, $X$ is normally distributed with

$$E[X] = \mu_1, \qquad \operatorname{Var}[X] = s_{1,1} = \sigma_1^2,$$

and it follows analogously that $Y$ is normally distributed with

$$E[Y] = \mu_2, \qquad \operatorname{Var}[Y] = s_{2,2} = \sigma_2^2.$$

For calculating the covariance we need to know how to calculate integrals with respect to the probability distribution of a random vector. We will obtain in Example 5.3.4:

$$\operatorname{Cov}[X,Y] = s_{2,1} = s_{1,2} = \varrho\,\sigma_1\sigma_2,$$

and thus $\varrho = \operatorname{Cov}[X,Y]/\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}$ is the correlation coefficient.

## 5.3. Moments of functions of random vectors

At the end of Chapter 4 we introduce the covariance of two random variables $X$ and $Y$. In order to give a formula for the covariance similar to Corollary 4.2.3, one needs the probability distribution of the random vector $(X,Y)\colon \Omega \to \mathbb{R}^2$. More general, one can derive a multivariate version of Theorem 4.2.2. We skip this step and only formulate the analogue of Corollaries 4.2.4 and 4.2.6 in the multi-dimensional setting and restrict ourselves to the 2-dimensional case.

**Corollary 5.3.1.** *Let $X = (X_1, X_2)\colon \Omega \to \mathbb{R}^2$ be a discrete random vector with values in $\Gamma = \{\gamma_1, \gamma_2, \ldots\} \times \{\delta_1, \delta_2, \ldots\} \subseteq \mathbb{R}^2$. For a measurable function $h\colon R^2 \to \mathbb{R}$ the following are equivalent:*

(a) *the random variable $h(X)\colon \Omega \to \mathbb{R}$ has finite expectation;*

(b) $\displaystyle\sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} |h(\gamma_k,\delta_\ell)|\, P(X_1 = \gamma_k, X_2 = \delta_\ell) < \infty.$

*In this case, we have*

$$E\big[h(X_1, X_2)\big] = \sum_{k=1}^{\infty}\sum_{\ell=1}^{\infty} h(\gamma_k, \delta_\ell) P(X_1 = \gamma_k, X_2 = \delta_\ell).$$

**Example 5.3.2.** (continues Example 5.1.6)
In this example we obtain

$E[U] = 0P(U = 0) + 1P(U = 1) = \frac{r}{r+b}$,
$E[V] = 0P(V = 0) + 1P(V = 1) = \frac{r}{r+b}$,

and for the mixed moments

$$\begin{aligned}
&E[UV] \\
&= 0 \cdot 0P(U = 0, V = 0) + 0 \cdot 1P(U = 0, V = 1) + 1 \cdot 0P(U = 1, V = 0) + 1 \cdot 1P(u = 1, V = 1) \\
&= \frac{r(r-1)}{(r+b)(r+b-1)}.
\end{aligned}$$

Consequently, we obtain

$$\mathrm{Cov}(U, V) = E\big[(U - E[U])(V - E[V])\big] = E[UV] - E[U]E[V] = \tfrac{-rb}{(r+b)^2(r+b-1)}.$$

**Corollary 5.3.3.** *Let $X = (X_1, X_2)\colon \Omega \to \mathbb{R}^2$ be an absolutely continuous random vector with density $f\colon \mathbb{R}^2 \to \mathbb{R}_+$ and let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be a measurable function. Then the following are equivalent:*

(a) *$h(X)$ has finite expectation;*

(b) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(u_1, u_2)|\, f(u_1, u_2)\, du_1 du_2 < \infty.$

*In this case, we have*

$$E[h(X_1, X_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(u_1, u_2) f(u_1, u_2)\, du_1 du_2.$$

**Example 5.3.4.** (continues Example 5.2.5 )

Now we are in the position to calculate the covariance of a 2-dimensional normal distribution:

$$\begin{aligned}
\operatorname{Cov}(X,Y) &= E\big[(X-\mu_1)(Y-\mu_2)\big] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} (x-\mu_1)(y-\mu_2)\psi_{\mu,\Sigma}(x,y)\,dy\,dx \\
&= \int_{\mathbb{R}} (x-\mu_1)\left(\int_{\mathbb{R}} (y-\mu_2)\psi_{\mu,\Sigma}(x,y)\,dy\right)dx \\
&= \int_{\mathbb{R}} (x-\mu_1) f_X(x)\frac{\sigma_2}{\sigma_1}\varrho(x-\mu_1)dx \\
&= \frac{\sigma_2}{\sigma_1}\varrho\int_{\mathbb{R}} (x-\mu_1)^2 f_X(x)\,dx \\
&= \frac{\sigma_2}{\sigma_1}\varrho\sigma_1^2 \\
&= \sigma_1\sigma_2\varrho.
\end{aligned}$$

Here, we used with a similar calculation as in Example 5.2.5 that

$$\begin{aligned}
\int_{\mathbb{R}} (y-\mu_2)\psi_{\mu,\Sigma}(x,y)\,dy &= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\varrho^2)}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\frac{1}{\sqrt{2\pi\sigma_2^2(1-\varrho^2)}} \\
&\qquad \int_{\mathbb{R}} y\exp\left(-\frac{1}{2\sigma_2^2(1-\varrho^2)}\left(y-\frac{\sigma_2}{\sigma_1}\varrho(x-\mu_1)\right)^2\right)dy \\
&= f_X(x)\frac{\sigma_2}{\sigma_1}\varrho(x-\mu_1),
\end{aligned}$$

which completes the Example.

## 5.4. Exercises

1. Assume $c \in [-\frac{1}{5}, \frac{1}{5}]$ and let $(X,Y)$ be a random vector with values in $\{0,1,2\}\times\{0,1,2\}$ and the distribution according following table:

| $P_{X,Y}$ | $\{Y=0\}$ | $\{Y=1\}$ | $\{Y=2\}$ |
|---|---|---|---|
| $\{X=0\}$ | $\frac{1}{5}$ | $\frac{1}{5}+c$ | $0$ |
| $\{X=1\}$ | $0$ | $\frac{1}{5}-c$ | $\frac{1}{5}+c$ |
| $\{X=2\}$ | $0$ | $0$ | $\frac{1}{5}-c$ |

   a) Calculate the marginal distributions of $(X,Y)$.
   b) Calculate the expectations $E[X]$, $E[Y]$, variances $\operatorname{Var}[X]$, $\operatorname{Var}[Y]$ and covariance $\operatorname{Cov}(X,Y)$.

2. Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be a random vector with the probability density

$$f\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f(x,y) = \begin{cases} c(x+y), & \text{if } x,y \in [0,1], \\ 0, & \text{else,} \end{cases}$$

for a constant $c \in \mathbb{R}$.

   (a) Determine the constant $c$.
   (b) Calculate the densities of $X$ and $Y$.
   (c) Calculate the covariance of $X$ and $Y$.

3. Proof the analogue result of Theorem 4.2.2 for random vectors $X\colon \Omega \to \mathbb{R}^d$ and a measurable function $h\colon \mathbb{R}^d \to \mathbb{R}$.

# 6 Independence

## 6.1. Conditional probabilities

In some random experiments, one knows already a part of the outcome before the experiment is finished. For example, consider the subsequent rolling of a die twice and you know already the result of the first toss.

**Example 6.1.1.** *Urn with red and black balls*
We consider the urn from Example 5.1.6: the urn contains $r$ red and $b$ black balls. One draws 2 balls without replacement. By numbering the red balls with $1, \ldots, r$ and the black balls with $r+1, \ldots, r+b$, we describe the state space by

$$\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \ldots, r+b\}, \, \omega_1 \neq \omega_2\}.$$

We choose the probability measure $P$ as the uniform distribution on $\Omega$,

$$P(\{(\omega_1, \omega_2)\}) = \frac{1}{|\Omega|} = \frac{1}{(r+b)(r+b-1)} \qquad \text{for all } (\omega_1, \omega_2) \in \Gamma;$$

see subsection 5.2.1. We consider the events

$$\begin{aligned} A &= \text{“the second ball is red”} = \{(\omega_1, \omega_2) \in \Omega : \omega_2 \leqslant r\}, \\ B &= \text{“the first ball is red”} = \{(\omega_1, \omega_2) \in \Omega : \omega_1 \leqslant r\}. \end{aligned}$$

The cardinality of the set $A$ is $r(r+b-1)$ since the second ball has $r$ possibilities and the first one has $r+b-1$ possibilities. Thus, one obtains

$$P(A) = \frac{|A|}{|\Gamma|} = \frac{r(r+b-1)}{(r+b)(r+b-1)} = \frac{r}{r+b}.$$

Assume that it is known that the first ball drawn is red. Given this additional information, what is the probability of the event $A$? One can answer this question by introducing a new state space

$$\Omega' = \{\omega' : \omega' \in \{1, \ldots, r-1+b\}\}.$$

The uniform distribution on $\Omega'$ is given by

$$P'(\{\omega'\}) = \frac{1}{|\Omega'|} = \frac{1}{r-1+b} \qquad \text{for all } \omega' \in \Omega'.$$

The new probability measure $P'$ on $\Omega'$ takes into account that only one ball is drawn and that there are only $r-1$ red balls. The event $A'$ = "the ball is red" has then the probability $P'(A') = \frac{r-1}{r-1+b}$.

However, can we find a *new* probability measure $P_B$ on the original state space $\Omega$ which takes into account that the event $B$ has already occurred? This new probability measure $P_B$ should satisfy:

(a) $P_B(B) = 1$, since $B$ has already occurred;

(b) there exists a constant $c_B > 0$ such that

$$P_B(A) = c_B P(A) \qquad \text{for all } A \in \mathscr{A}, A \subseteq B,$$

since if $A$ implies $B$, the new probability should be proportional to the original one.

It turns out that these two properties uniquely determines a new probability measure:

**Theorem 6.1.2.** *Let $(\Omega, \mathscr{A}, P)$ be an arbitrary probability space. Let $B \in \mathscr{A}$ be a set with $P(B) > 0$. Then there exists a unique probability measure $P_B$ on $(\Omega, \mathscr{A})$ satisfying:*

(a) $P_B(B) = 1$*;*

(b) $P_B(A) = c_B P(A)$ *for all* $A \in \mathscr{A}, A \subseteq B$ *and a constant* $c_B > 0$*.*

*This probability measure $P_B$ is given by*

$$P_B \colon \mathscr{A} \to [0,1], \qquad P_B(A) = \frac{P(A \cap B)}{P(B)}.$$

Instead of using the notation $P_B$ we use $P(\cdot|B)$ as introduced in the following definition:

**Definition 6.1.3.** *Let $B \in \mathscr{A}$ be a set with $P(B) > 0$. Then the probability measure*

$$P(\cdot|B) : \mathscr{A} \to [0,1], \qquad P(A|B) = \frac{P(A \cap B)}{P(B)},$$

*is called the* conditional probability given $B$.

Conditional probabilities are often interpreted in the following two ways:

- *frequentist:* if the random event is repeated often enough, then $P(A|B)$ gives approximately the number of cases that $A$ occurs among all cases in which $B$ has occurred.
- *subjective:* if $P$ represents my point of view of the random experiment, then $P(\cdot|B)$ is my guess after I was informed of the occurrence of $B$.

**Example 6.1.4.** *Urn with red and black balls (continues Example 6.1.1)*
By using the definition of the conditional probability we obtain

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{|A\cap B|}{|\Gamma|}\frac{|\Gamma|}{|B|} = \frac{r(r-1)}{r(r+b-1)} = \frac{r-1}{r+b-1}.$$

**Example 6.1.5.** *Rolling two dice (continues Example 5.2.1)*
The rolling of two dice is modelled by the uniform distribution $P$ on

$$\Gamma = \{(k,\ell) : k,\, \ell \in \{1,\dots,6\}\}.$$

We consider the events

$$\begin{aligned} A_r &:= \text{“$r$ appears on the first die”},\\ B_s &:= \text{“the sum of the two dice is $s$”}. \end{aligned}$$

By determining the cardinality of the sets one obtains

$$P(A_r|B_) = \frac{|A_r\cap B_7|}{|B_7|} = \frac{|\{(r,7-r)\}|}{|\{(1,6),(2,5),\dots,(6,1)\}|} = \frac{1}{6} = P(A_r) \quad \text{for all } r = 1,\dots,6.$$

Consequently, the additional information that $B_7$ has occurred does not provide more information on the occurrence of the event $A_r$. However, we have

$$P(A_r|B_{11}) = \frac{1}{2} \quad \text{for } r = 5,6, \qquad P(A_r|B_{11}) = 0 \quad \text{for } r \leqslant 4,$$

which both do not coincide with $P(A_r) = \frac{1}{6}$.

The following formula are helpful in many applications.

**Theorem 6.1.6.** *Let $(\Omega, \mathscr{A}, P)$ be a probability space and $\{B_k\}_{k\in I}$ for $I \subseteq \mathbb{N}$ be a collection of pairwise disjoint sets with $P(B_k) > 0$ and $\Omega = \bigcup_{i\in I} B_i$. Then it follows:*

(a) partition equation*:*

$$P(A) = \sum_{i\in I} P(A|B_i)P(B_i) \qquad \textit{for all } A \in \mathscr{A}.$$

(b) Bayes' theorem*:*

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i\in I} P(A|B_i)P(B_i)}$$

*for all $k \in I$ and $A \in \mathscr{A}$ with $P(A) > 0$.*

**Example 6.1.7.** A certain disease is found in 4% of the total population. A blood test is 90% effective, i.e. it will yield the accurate positive result in 90% of the cases where the tested patient actually suffers from the disease. However, it also yields a positive result in 20% of the cases where the disease is not present. What is the probability that a person suffers actually from the disease if the test is positive?

The state space of the random experiment is described by

$$\Omega = \{hp, hn, sp, sn\},$$

where for example $hp$ denotes the event that the person is healthy and the test is positive (this is the simultaneous occurrence of the two events not the conditional event). Define the set $B_1 = \{sp, sn\}$, which describes the event that the person is sick and define $B_2 = B_1^c = \{hp, hn\}$, which describes the event that the person is healthy. It is given that $P(B_1) = 0.04$ and thus $P(B_2) = 0.96$. Bayes' theorem implies

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \frac{0.9 \cdot 0.04}{0.9 \cdot 0.04 + 0.2 \cdot 0.96} \approx \frac{1}{6}.$$

The reason for this surprising low probability can be seen in the rather large proportion of healthy person.

$$P(B_1|A^c) = \frac{P(A^c|B_1)P(B_1)}{P(A^c|B_1)P(B_1) + P(A^c|B_2)P(B_2)} = \frac{0.1 \cdot 0.04}{0.1 \cdot 0.04 + 0.8 \cdot 0.96} \approx 0.005.$$

The last probability is rather low meaning that the likelihood of this kind of error is very small. Consequently, the test is suitable to exclude the presence of the disease, but a person with a positive test result must have some further examinations.

## 6.2. Independent events

Two events $A$ and $B$ are colloquially called *independent*, if the probability for the event $A$ does not depend on the information whether $B$ has occurred and vice versa. According to this non-formal understanding, we would expect for the conditional probabilities that

$$P(A|B) = P(A) \qquad \text{and} \qquad P(B|A) = P(B).$$

By the very Definition 6.1.3, this results in the following (now) formal definition:

**Definition 6.2.1.** *Two events $A$ and $B$ in $\mathscr{A}$ are called* independent *if*

$$P(A \cap B) = P(A)P(B).$$

**Example 6.2.2.** *Urn with red and black balls*
An urn contains $r$ red and $b$ black balls. One draws two balls with replacing them, i.e. the drawn balls are returned to the urn. The state space can be described by

$$\Omega_2 = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, \dots, r + b\}\},$$

where the numbers $1, \ldots, r$ represents the red balls and the numbers $r+1, \ldots, r+b$ the black balls. We can assume the uniform distribution on $\Omega_2$, that is

$$P_2(\{(\omega_1, \omega_2)\}) = \frac{1}{|\Omega_2|} = \frac{1}{(r+b)^2} \qquad \text{for all } (\omega_1, \omega_2) \in \Omega_2.$$

By determining the cardinality, we obtain for the events $A_2 =$“the second ball is red” and $B_2 =$“the first ball is red” that

$$P_2(A_2 \cap B_2) = \frac{r^2}{(r+b)^2} = \frac{r(r+b)}{(r+b)^2}\frac{r(r+b)}{(r+b)^2} = P_2(A_2)P_2(B_2).$$

However, if the balls are not replaced, then we are in the situation of the random experiment described in Example 6.1.1. Here, the same events $A =$“the second ball is red” and $B$=“the first ball is red”, but considered as subsets of $\Omega$ defined in Example 6.1.1, obey

$$P(A \cap B) = \frac{r(r-1)}{(r+b)(r+b-1)} < \frac{r}{r+b}\frac{r}{r+b} = P(A)P(B),$$

where $P$ denotes the uniform distribution on $\Omega$.

There are situations where events can be independent although they are causally related.

**Example 6.2.3.** Rolling two dice (continues Example 6.1.5)
The events $A_6$ and $B_7$ obey

$$P(A_6 \cap B_7) = \frac{1}{36} = P(A_6)P(B_7).$$

Although the sum of the two dice is causally related to the outcome from the first roll, the two events are independent; see Exercise 6.4.9.

**Definition 6.2.4.** *A finite collection $A_1, \ldots, A_n$ of events $A_k \in \mathscr{A}$ are called* independent*, if*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k}) \tag{6.2.1}$$

*for each $i_1, \ldots, i_k \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, n\}$.*

Note, that Definition 6.2.4 requires that the product formula (6.2.1) is established for each subsequence and not only for $A_1 \cap \cdots \cap A_n$. This fact is illustrated by the following example:

**Example 6.2.5.** Assume $\Omega = \{1, \ldots, 8\}$ and let $P$ be the uniform distribution on $\Omega$. The events

$$A = \{1, \ldots, 4\}, \quad B = \{2, \ldots, 5\}, \quad C = \{4, \ldots, 7\}$$

satisfy $P(A \cap B \cap C) = P(A)P(B)P(C)$, but $A$ and $B$ are not independent:

$$P(A \cap B) = \tfrac{3}{8} \neq \tfrac{1}{4} = P(A)P(B).$$

In order to show independence for some sets it is not sufficient to consider all pair of events, as the following example illustrates:

**Example 6.2.6.** Tossing a coin twice can be modelled by $\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{h, t\}\}$ and the uniform distribution $P$ on $\Omega$. Consider the events

$$\begin{aligned} A &= \text{“first toss lands heads up”}, \\ B &= \text{“second toss lands heads up”}, \\ C &= \text{“in both tosses the same result”}. \end{aligned}$$

By computing the cardinal numbers, we obtain

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C).$$

However, we have

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C).$$

Thus, the events $A$, $B$ and $C$ are not independent.

**Definition 6.2.7.** *Let $I$ be an arbitrary index set. An infinite collection $\{A_n\}_{n \in I}$ of events $A_n \in \mathscr{A}$ is called* independent*, if*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k}) \tag{6.2.2}$$

*for each $i_1, \ldots, i_k \in I$ and $k \in \mathbb{N}$.*

In other words, a possibly infinite and uncountable number of events are independent, if each finite subset of these events are independent according to Definition 6.2.4. If the index set $I$ is finite both Definitions 6.2.4 and 6.2.7 coincide. Note, that in Definition 6.2.7 the set $I$ is arbitrary, thus it is even not necessarily countable.

For the next result, it is helpful to introduce some notations: assume that $(A_k)_{k \in \mathbb{N}}$ is a sequence of sets in $\mathscr{A}$, and let $S$ denote the event that

$$\omega \in S \;\Leftrightarrow\; \omega \in A_k \text{ for infinitely many } k \in \mathbb{N}\,.$$

This we can rephrase in the following way:

$$\omega \in S \;\Leftrightarrow\; \text{for each } k \in \mathbb{N} \text{ there is an } n \geqslant k \text{ such that } \omega \in A_n.$$

Here, the emphasise is on the fact that $k \in \mathbb{N}$ can be chosen as big as you like and you still find $n \geqslant k$ such that $\omega \in A_n$. Using unions and sections the right hand side of the last line can be written as

$$\omega \in S \;\Leftrightarrow\; \omega \in \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n.$$

Instead of $S$ we use the notation lim sup (which should not be confused with the limit superior of functions) and it has a purely symbolic meaning:

$$\limsup_{k\to\infty} A_k := \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$$

Another related event is the following: let $I$ denote the event that

$$\omega \in I \;\Leftrightarrow\; \omega \in A_k \text{ for all } k \in \mathbb{N} \text{ with the exeption of finitely many } k.$$

This we can rephrase in the following way

$$\omega \in I \;\Leftrightarrow\; \text{there exists } k \in \mathbb{N} \text{ such that } \omega \in A_n \text{ for all } n \geqslant k.$$

Here, the understanding is that the finitely many sets $A_{i_1}, \dots, A_{i_r}$, in which $\omega$ is not an element of, have indices $i_j$ smaller than $k$. Using unions and sections the right hand side of the last line can be written as

$$\omega \in I \;\Leftrightarrow\; \omega \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n.$$

Instead of $I$ we use the notation lim inf (which should not be confused with the limit inferior of functions) and it has a purely symbolic meaning:

$$\liminf_{k\to\infty} A_k := \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n$$

**Theorem 6.2.8.** *(Borel-Cantelli)*
*Let $(\Omega, \mathscr{A}, P)$ be a probability space and $\{A_k\}_{k\in\mathbb{N}}$ be a collection events $A_k \in \mathscr{A}$.*

(a) *If* $\sum_{k=1}^{\infty} P(A_k) < \infty$ *then* $P\left(\limsup_{k\to\infty} A_k\right) = 0$.

(b) *If* $P\left(\limsup_{k\to\infty} A_k\right) = 0$ *and* $\{A_k\}_{k\in\mathbb{N}}$ *are independent, then* $\sum_{k=1}^{\infty} P(A_k) < \infty$.

(c) *If* $\sum_{k=1}^{\infty} P(A_k) = \infty$ *and* $\{A_k\}_{k\in\mathbb{N}}$ *are independent, then* $P\left(\limsup_{k\to\infty} A_k\right) = 1$.

*Proof.* Note that (c) is just the negation of (b). □

## 6.3. Independent random variables

In the last step the notion of independence of sets are carried over to random variables.

**Definition 6.3.1.** *Let $I$ be an arbitrary index set. A collection $\{X_n\}_{n\in I}$ of random variables $X_n\colon \Omega \to \mathbb{R}$ is called* independent*, if*

$$P(X_{i_1} \in B_{i_1}, \dots, X_{i_k} \in B_{i_k}) = P(X_{i_1} \in B_{i_1}) \cdot \ldots \cdot P(X_{i_k} \in B_{i_k}) \tag{6.3.3}$$

*for each $i_1, \dots, i_k \in I$, $B_{i_1}, \dots, B_{i_k} \in \mathfrak{B}(\mathbb{R})$ and each $k \in \mathbb{N}$.*

The argument of the probability measure on the left hand side is short hand:

$$\begin{aligned} &\{X_{i_1} \in B_{i_1}, \dots, X_{i_k} \in B_{i_k}\} \\ &\qquad := \{\omega \in \Omega : \{X_{i_1}(\omega) \in B_{i_1}\} \cap \dots \cap \{X_{i_k}(\omega) \in B_{i_k}\}\}. \end{aligned}$$

Thus, a collection of random variables $\{X_k\}_{k\in I}$ are independent according to Definition 6.3.1, if for each $i_1, \dots, i_k \in I$ and event $B_{i_1}, \dots, B_{i_k} \in \mathfrak{B}(\mathbb{R})$ and each $k \in \mathbb{N}$ the sets

$$\{X_{i_1} \in B_{i_1}\}, \dots, \{X_{i_k} \in B_{i_k}\}$$

are independent according to Definition 6.2.4.

Even for finitely many random variables it might be rather difficult to establish Condition (6.3.3), since it requires to consider each possible subsets of indices and sets. Using the cumulative distribution function simplifies the situation significantly:

**Theorem 6.3.2.** *Let $X_k\colon \Omega \to \mathbb{R}$ be a random variable with cumulative distribution function $F_{X_k}\colon \mathbb{R} \to [0,1]$ for $k = 1, \dots, n$. Then $X_1, \dots, X_n$ are independenet if and only if*

$$P\big(X_1 \leqslant x_1, \dots, X_n \leqslant x_n\big) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n) \qquad \text{for all } x_1, \dots, x_n \in \mathbb{R}.$$

In the two special cases of discrete or absolutely continuous random vectors we can simplify the condition significantly.

**Theorem 6.3.3.** *Assume that $X = (X_1, \dots, X_n)\colon \Omega \to \mathbb{R}^n$ is a discrete random vector taking values in $\Gamma \subseteq \mathbb{R}^n$. Then the following are equivalent:*

(a) *the random variables $X_1, \dots, X_n$ are independent;*

(b) *for each $(x_1, \dots, x_n) \in \Gamma$ we have:*

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n).$$

The analogue result of Theorem 6.3.3 can be derived in the case that the random vector $X := (X_1, \dots, X_n)\colon \Omega \to \mathbb{R}^n$ has a density.

**Theorem 6.3.4.** *Let $X := (X_1, \dots, X_n)\colon \Omega \to \mathbb{R}^n$ be a random vector with density $f_X\colon \mathbb{R}^n \to \mathbb{R}_+$. Then the following are equivalent:*

(a) *the random variables $X_1, \dots, X_n$ are independent;*

(b) *the density* $f_X$ *obeys*

$$f_X(x_1,\dots,x_n) = f_{X_1}(x_1)\cdot\ldots\cdot f_{X_n}(x_n) \qquad \text{for all } x_1,\dots,x_n\in\mathbb{R},$$

*where* $f_{X_k}\colon \mathbb{R}\to\mathbb{R}$ *is the density of* $X_k$ *for* $k=1,\dots,n$.

Recall that if $(X,Y)$ is random vector and $X$ and $Y$ have a density, than $(X,Y)$ may not have a density, see discussion before Theorem 5.1.5. However, in the case of independent random variables $X$ and $Y$, this is true.

**Corollary 6.3.5.** *Let* $X := (X_1,\dots,X_n)\colon \Omega\to\mathbb{R}^n$ *be a random vector. If each random variable* $X_k\colon\Omega\to\mathbb{R}$ *has a density* $f_{X_k}$ *and the random variables* $X_1,\dots,X_n$ *are independent, then* $X$ *has a density* $f_X$*, which is given by*

$$f_X(x_1,\dots,x_n) = f_{X_1}(x_1)\cdot\ldots\cdot f_{X_n}(x_n) \qquad \text{for all } x_1,\dots,x_n\in\mathbb{R}\,.$$

Assuming independence in modelling is often a strong assumption. But independence of random variables has many important consequences and it often simplifies the situation. One of consequences is that the expectation of the product of two random variable is the product of the expectation.

**Theorem 6.3.6.** *Let* $X$, $Y:\Omega\to\mathbb{R}$ *be random variables. Then the following are equivalent:*

(a) $X$ *and* $Y$ *are independent;*

(b) *for all bounded, measurable functions* $f,g\colon\mathbb{R}\to\mathbb{R}$ *we have*

$$E\big[f(X)g(Y)\big] = E\big[f(X)\big]E\big[g(Y)\big].$$

**Corollary 6.3.7.** *Let* $X$, $Y:\Omega\to\mathbb{R}$ *be independent random variables in* $\mathscr{L}^2(\Omega,P)$*. Then it follows that*

$$E\big[XY\big] = E\big[X\big]E\big[Y\big].$$

We finish the section with showing the fact that independence of two random variables implies that the covariance and thus the correlation of the two random variables is zero.

**Corollary 6.3.8.** *Let* $X$, $Y:\Omega\to\mathbb{R}$ *be independent random variables with* $E[X^2]<\infty$ *and* $E[Y^2]<\infty$*. Then it follows that:*

(a) $\mathrm{Var}[X+Y] = \mathrm{Var}[X]+\mathrm{Var}[Y]$*;*

(b) $\mathrm{Cov}(X,Y)=0$*.*

We will later need the meaning of independence of a random variable and a $\sigma$-algebra. Revising the previous definitions this is straightforward:

**Definition 6.3.9.** *Let* $X:\Omega\to\mathbb{R}$ *be a random variable and* $\mathscr{D}$ *a* $\sigma$*-algebra with* $\mathscr{D}\subseteq\mathscr{A}$*. Then* $X$ *and* $\mathscr{D}$ *are called independent if*

$$P(B\cap D) = P(B)P(D) \quad \text{for all } B\in\sigma(X), D\in\mathscr{D}.$$

**Example 6.3.10.** Let $X$ and $Y$ be random variables. Then $X$ and $Y$ are independent if and only if $X$ and $\sigma(Y)$ are independent.

## 6.4. Exercises

1. Let $X$ and $Y$ be independent, geometrically distributed random variables with the same parameter $\alpha > 0$ Find the conditional probability of $X = k$ under $X + Y = n$.

2. Before car drivers have to pass a blood test they are first subject to a breath test. Only after a positive breath test result a driver is taken to the blood test. The breath test yields a positive result with probability 0.95, if the tested person is above the legal alcohol limit; it yields a negative result with probability 0.9, if the tested person is below the legal alcohol limit. From historical data, it is known that 5% of all car drivers are above the legal alcohol limit. What is the probability that
   (a) the test result is positive?
   (b) the alcohol percentage of a tested person is below the legal alcohol limit, although the test is positive?
   (c) the alcohol percentage of a tested person is above the legal alcohol limit, although the test is negative?

3. *Bertrand's box paradox*
   Each of 3 identical cabinets has 2 drawers, and in each drawer there is a coin. In one of the cabinet both coins are golden, in the other cabinet both coins are silver and the last cabinet contains one golden and one silver coin. A cabinet is randomly selected and one of its drawers is randomly opened.
   (a) What is the probability that one finds a golden coin?
   (b) Assume you have found a silver coin. What is the probability that the other drawer of the same cabinet contains a golden coin?

4. (∗) *Monty Hall problem*
   The guest of a game show is given the choice of three doors: behind one door is a car, whereas behind the other two doors are a goat, respectively. After the guest has chosen one of the doors, the host, who knows what is behind the doors, opens another door. The host offers then to the guest to revise his choice. Is it to the advantage of the guest to switch the door?

5. Let $X$ be a random variable with values in $\mathbb{N}_0$. Show that the following are equivalent:
   (a) $X$ has a geometric distribution;
   (b) the distribution of $X$ is *memoryless*:
   $$P(X = n + k | X \geqslant k) = P(X = n) \qquad \text{for all } k,\, n \in \mathbb{N}_0;$$
   (c) $P(X = n + 1 | X \geqslant 1) = P(X = n)$ for all $n \in \mathbb{N}_0$.

6. A source sends a random number $S$ of signals per time unit, which is distributed according to a Poisson distribution with parameter $\alpha > 0$. A receiver records mutually independent each of the sent signals with probability $p = 0.8$. What is the distribution of the number $Z$ of recorded signals?

7. *Modelling the aggregated claims in actuarial mathematics*
   Let $(X_k)_{k\in\mathbb{N}}$ be a sequence of independent random variables with values in $\mathbb{R}$ and identical distribution. Let $N$ be another random variable with values in $\mathbb{N}_0$ and independent from $(X_k)_{k\in\mathbb{N}}$. Define
   $$Y := \begin{cases} \sum_{k=1}^N X_k, & \text{if } N \geqslant 1,\\ 0, & \text{else.}\end{cases}$$
   (a) Show that if $E[X_1] < \infty$ and $E[N] < \infty$ then
   $$E[Y] = E[X_1]E[N].$$
   (b) Assume that $E[X_1^2] < \infty$ and $E[N^2] < \infty$. Show that
   $$\operatorname{Var}[Y] = \operatorname{Var}[X_1]E[N] + \big(E[X_1]\big)^2 \operatorname{Var}[N].$$
8. (The monkey theorem)
   A (immortal) monkey is seated at a typewriter and randomly hits the keys. Show that with probability one, the monkey will write the complete works by Shakespeare.
9. Assume that $A \in \mathscr{A}$ satisfies $P(A) = 0$ or $P(A) = 1$. Show that $A$ is independent of $A$.
10. Let $X, Y : \Omega \to \mathbb{R}$ be independent random variables and $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. Show that $f(X)$ and $f(Y)$ are independent.
11. For sets $A, B, C \in \mathscr{A}$ show the following:
    (a) $A$ and $B$ are independent if and only if $A$ and $B^c$ are independent.
    (b) If $P(B) = 0$ or $P(B) = 1$ then $A$ and $B$ are independent.
    (c) If $A$, $B$ and $C$ are independent then $A \cup B$ and $C$ are independent.
12. Let $(X, Y) : \Omega \to \mathbb{R}^2$ be a normally distributed random vector. Show that $X$ and $Y$ are independent if and only if $\operatorname{Cov}(X, Y) = 0$.
13. Let $X_1, \dots, X_n$ be independent, identically distributed random variables with $E[X_1] = \mu$ and $\operatorname{Var}[X_1] = \sigma^2 < \infty$. Define the random variables
    $$\bar{X} := \frac{1}{n}\sum_{k=1}^n X_k, \qquad S^2 := \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2.$$
    Show that
    (a) $E[\bar{X}] = \mu$;
    (b) $E[S^2] = \sigma^2$. (Hint: consider $(X_k - \mu) - (\bar{X} - \mu)$).
    (c) $\operatorname{Cov}(\bar{X}, X_i - \bar{X}) = 0$ for all $i = 1, \dots, n$.
    If $X_1$ has a $\mathrm{N}(\mu, \sigma^2)$ distribution one can conclude from (c) that $\bar{X}$ and $S^2$ are independent by Exercise 6.4.12. For this, one needs to know that $(\bar{X}, X_i - \bar{X})$ is normally distributed as a random vector, which needs a result we do not have.

# 7 Functions of random vectors

In this chapter we investigate the random variable $h(X)\colon \Omega \to \mathbb{R}$ where $X\colon \Omega \to \mathbb{R}^d$ is a random vector and $h\colon \mathbb{R}^d \to \mathbb{R}$ is a measurable function. To be precise, we have first to make sure that $h(X)$ is a random variable, which is a simple extension of Corollary 3.1.12. Our main focus is, assuming that $X$ is discrete or absolutely continuous, what can we say about the probability distribution of the random variable $h(X)$? We always assume that the underlying probability space is denoted by $(\Omega, \mathscr{A}, P)$.

## 7.1. Transformation

As mentioned we begin with a generalisation of Corollary 3.1.12:

**Corollary 7.1.1.** *If $X\colon \Omega \to \mathbb{R}^d$ is a random vector and $h\colon \mathbb{R}^d \to \mathbb{R}$ is measurable, then*

$$h \circ X\colon \Omega \to \mathbb{R} \ \text{ is a random variable.}$$

Recall that the measurability of the mapping $X\colon \Omega \to \mathbb{R}^d$ is defined in Definition 5.1.1 whereas the measurability of the function $h\colon \mathbb{R}^d \to \mathbb{R}$ is covered by Definition 3.1.1 applied to $\Omega = \mathbb{R}^d$ and $\mathscr{A} = \mathfrak{B}(\mathbb{R}^d)$.

**Theorem 7.1.2.** *Let $X\colon \Omega \to \mathbb{R}^d$ be a discrete random vector with values in $\{\gamma_1, \gamma_2, \dots\} \subseteq \mathbb{R}^d$ and $h\colon \mathbb{R}^d \to \mathbb{R}$ a measurable function. Then $h(X)$ is a discrete random variable with values in $\{h(\gamma_1), h(\gamma_2), \dots\}$ and its probability distribution is given by*

$$P_{h(X)}\colon \mathfrak{B}(\mathbb{R}) \to [0,1], \qquad P_{h(X)}(A) = \sum_{k=1}^{\infty} p_k^h \delta_{h(\gamma_k)}(A) \qquad \text{for all } A \in \mathfrak{B}(\mathbb{R}),$$

*where $p_k^h = P\big(\{\omega \in \Omega : h(X(\omega)) = h(\gamma_k)\}\big)$.*

For discrete random vectors, it is easy to derive a formula for the probability distribution of $h(X)$ in Theorem 7.1.2. However, in many cases the formula does not provide much simplifications.

**Example 7.1.3.** Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be a discrete random vector with values in $\mathbb{N}\times\mathbb{N}$. Then the random variable $S = X+Y$ is discrete with

$$P(S=k) = \sum_{j=1}^{k} P(X=k-j, Y=j) \qquad \text{for all } k\in\mathbb{N}\,.$$

**Example 7.1.4.** Let $X_1,\dots X_n$ be independent, identically distributed random variables distributed according to a Bernoulli distribution with parameter $p\in(0,1)$. Then the random variable $S = X_1+\cdots+X_n$ is binomially distributed with parameter $n$ and $p$.

In the case of an absolutely continuous random variable the situation is not that easy. Exercise 2.5.6 shows that even for a simple function $h\colon \mathbb{R}\to\mathbb{R}$ the random variable $g(X)$ may have a discontinuous cumulative distribution function and thus, cannot have a density.

In the one-dimensional case, the existence of the density and its form often can be derived from the fundamental theorem of calculus (you might have seen this in your undergraduates). We rephrase this result in our setting of cumulative distribution functions:

**Theorem 7.1.5.** *Let $F_Y\colon \mathbb{R}\to[0,1]$ be the cumulative distribution function of a random variable $Y\colon\Omega\to\mathbb{R}$. If $F_Y$ is continuously differentiable except at finitely many points then*

$$F_Y(x) = \int_{-\infty}^{x} F_Y'(u)\,du \qquad \text{for all } x\in\mathbb{R},$$

*where $F_Y'(u)$ is the derivative of $F_Y$ in $u$ if $F_Y$ is continuously differentiable in $u$ and otherwise it can be defined arbitrarily. In particular, $F_Y'$ is the density of $Y$.*

**Example 7.1.6.** Let $X$ be standard normally distributed and $h\colon\mathbb{R}\to\mathbb{R}$ be defined by $h(x)=x^2$. Then the cumulative distribution function $F_Y$ of $Y:=h(X)$ is given by

$$F_Y\colon\ \mathbb{R}\to[0,1], \qquad F_Y(y) = \begin{cases} \Phi(\sqrt{y}\,) - \Phi(-\sqrt{y}\,), & \text{if } y>0,\\ 0, & \text{else,}\end{cases}$$

where $\Phi$ denotes the cumulative distribution function of $X$. Since $F_Y$ is continuously differentiable, we obtain by Theorem 7.1.5 that $Y$ has the density

$$f_Y\colon\ \mathbb{R}\to\mathbb{R}_+, \qquad f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}\,y^{-1/2}e^{-\frac{1}{2}y}, & \text{if } y>0,\\ 0, & \text{else.}\end{cases}$$

The probability distribution of $Y$ is called *Chi-squared distribution with* 1 *degree of freedom*.

Let $X$ be a random variable with cumulative distribution function $F_X$ and $h\colon\mathbb{R}\to\mathbb{R}$ a measurable function. If the function $h$ has an inverse function $u\colon\mathbb{R}\to\mathbb{R}$ we obtain

$$F_Y(x) = P\big(h(X)\leqslant x\big) = P\big(X\leqslant u(x)\big) = F_X(u(x)).$$

Consequently, if $x\mapsto F_X(u(x))$ is continuously differentiable we can apply Theorem 7.1.5 to obtain the density of $Y$. More precisely, we have the following result:

**Corollary 7.1.7.** Transformation formula one-dimensional
*Let $X\colon \Omega \to \mathbb{R}$ be an absolutely continuous random variable with density $f\colon \mathbb{R} \to \mathbb{R}_+$ and let $h : N \to \mathbb{R}$ for $N \subseteq \mathbb{R}$ be a function with*

(i) *$N \subseteq \mathbb{R}$ is open and $P(X \in N) = 1$;*

(ii) *$h$ is injective (one-to-one) with inverse function $u\colon M \to N$, where $M := h(N)$;*

(iii) *$h$ and $u$ are continuously differentiable.*

*Then the random variable $h(X)$ has the density $g\colon \mathbb{R} \to \mathbb{R}_+$ which is given by*

$$g(y) = \begin{cases} |u'(y)|\, f(u(y)), & y \in M, \\ 0, & \textit{else}. \end{cases}$$

Note, that Corollary 7.1.7 cannot be applied in Example 7.1.6 as the function $h(x) = x^2$ is not injective. But the following example is tailor made for an application of Corollary 7.1.7:

**Example 7.1.8.** Assume that $X$ is normally distributed with expectation $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $h\colon \mathbb{R} \to \mathbb{R}$ be defined by $h(x) = \exp(x)$. The function $h$ has the inverse function $u\colon M \to \mathbb{R}$ defined by $u(y) = \ln(y)$ where $M = (0, \infty) = h(\mathbb{R})$ and $N = \mathbb{R}$. Then the random variable $g(X) = \exp(X)\colon \Omega \to \mathbb{R}$ is absolutely continuous with density

$$g(y) = |u'(y)|\, \psi_{\mu,\sigma^2}(u(y)) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(\ln y - \mu)^2\right) \quad \text{for all } y > 0,$$

and $g(y) = 0$ for all $y \leqslant 0$. Thus, $g$ is the probability density function of the log-normal distribution introduced in Subsection 2.3.4.

**Theorem 7.1.9.** Transformation formula multi-dimensional
*Let $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ be a random vector with density $f\colon \mathbb{R}^d \to \mathbb{R}_+$ and let $h : N \to \mathbb{R}^d$ for $N \subseteq \mathbb{R}^d$ be a function with*

(i) *$N \subseteq \mathbb{R}^d$ is open and $P((X_1, \dots, X_d) \in N) = 1$;*

(ii) *$h$ is injective (one-to-one) with inverse function $u = (u_1, \dots, u_d)\colon M \to N$, where $M := h(N)$;*

(iii) *$h$ and $u$ are continuously differentiable.*

*Then the random vector $h(X_1, \dots, X_d)$ has the density $g\colon \mathbb{R}^d \to \mathbb{R}_+$ which is given by*

$$g(y) = \begin{cases} |detJ(y)|\, f(u_1(y), \dots, u_d(y)), & y \in M, \\ 0, & \textit{else}, \end{cases}$$

*where* $J(y) = \left(\dfrac{\partial u_i}{\partial y_j}(y)\right)_{i,j=1}^{d}$.

**Example 7.1.10.** Let$(X, Y)\colon \Omega \to \mathbb{R}^2$ be a random vector with density $f\colon \mathbb{R}^2 \to \mathbb{R}_+$ and marginal probability densities $f_X$ and $f_Y$. Does $X+Y$ has a density? Define the function

$$h\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad h(x_1, x_2) = \begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix}.$$

The function $h$ has the inverse function $u$ defined by

$$u : \mathbb{R}^2 \to \mathbb{R}^2, \qquad u(y_1, y_2) = \begin{pmatrix} y_1 - y_2 \\ y_2 \end{pmatrix}$$

Theorem 7.1.9 implies that $(X+Y, Y)$ has the density

$$g : \mathbb{R}^2 \to \mathbb{R}_+, \qquad g(y_1, y_2) = f(y_1 - y_2, y_2).$$

The density of $X+Y$ is the marginal density $g_1$ of $g$:

$$g_1 : \mathbb{R} \to \mathbb{R}_+, \qquad g_1(y) = \int_{\mathbb{R}} f(y - y_2, y_2)\, dy_2.$$

**Example 7.1.11.** $Z = (Z_1, \dots, Z_d)\colon \Omega \to \mathbb{R}^d$ be a random vector distributed according to a $N(\mu, \Sigma)$ distribution for $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. If $\Delta \in \mathbb{R}^{m \times d}$, then the random vector $\Delta Z\colon \Omega \to \mathbb{R}^m$ is distributed according to a $N(\Delta\mu, \Delta\Sigma\Delta^T)$ distribution.

In particular for $d = 2$ it follows that if $(Z_1, Z_2)$ is normally distributed with expectation $\mu = (\mu_1, \mu_2)$ and covariance matrix $\Sigma = (s_{i,j})_{i,j=1}^2$, then for every $\alpha_1, \alpha_2 \in \mathbb{R}$ the random variable $\alpha_1 Z_1 + \alpha_2 Z_2$ is normally distributed with expectation $\alpha_1\mu_1 + \alpha_2\mu_2$ and variance $s_{1,1} + s_{2,2} + 2s_{1,2}$

In the previous example, it is essential that the random vector $(X_1, \dots, X_d)$ has a normal distribution. It is not enough to assume that $X_1, \dots, X_d$ are normally distributed. This, and something more, is shown by the following example:

**Example 7.1.12.** For a random variable $X\colon \Omega \to \mathbb{R}$ with a $N(0,1)$ normal distribution define

$$Y := \begin{cases} X, & \text{if } |X| < c, \\ -X, & \text{if } |X| \geqslant c, \end{cases}$$

for a constant $c > 0$. Symmetry of the normal distribution implies for every $B \in \mathfrak{B}(\mathbb{R})$

$$\begin{aligned} P(Y \in B) &= P\big(X \in B \cap (-c, c)\big) + P\big(-X \in B \cap \big((-\infty, c] \cup [c, \infty)\big)\big) \\ &= P\big(X \in B \cap (-c, c)\big) + P\big(X \in B \cap \big((-\infty, c] \cup [c, \infty)\big)\big) \\ &= P(X \in B), \end{aligned}$$

and thus $Y$ is also normally distributed. However for $S := X + Y$ one obtains

$$P(S = 0) = P(Y = -X) = P(|X| \geqslant c) \neq 0.$$

Thus, $S$ cannot have a density according to Corollary 2.3.4. Thus, it is essential in Example 7.1.11 to assume that the random vector $Z$ is normally distributed.

This example also shows that the random vector $(X, Y)\colon \Omega \to \mathbb{R}^2$ cannot have a density, since otherwise $S$ would have a density, according to Example 7.1.10. Thus, we have an example of a random vector $(X, Y)\colon \Omega \to \mathbb{R}^2$ which is not absolutely continuous but both marginals $X$ and $Y$ are absolutely continuous.

Even more, this example also shows that in Exercise 6.4.12 it is essential to require that the random vector $(X, Y)$ is a normally distributed and it is not enough to assume that $X$ and $Y$ are normally distributed. This follows since the definition of $Y$ implies that the correlation between $X$ and $Y$ equals 1 for $c = 0$ whereas it equals -1 for $c = \infty$. One can deduce that there exists $c \in (0, \infty)$ such that $\operatorname{Cov}(X, Y) = 0$ but they are not independent and both, $X$ and $Y$, are normally distributed.

## 7.2. Random vectors with independent margins

If the random vector $X = (X_1, \dots, X_d)$ has independent margins $X_1, \dots, X_d$, the distribution of $h(X)$ can often be derived explicitly.

**Corollary 7.2.1.**

(a) *Let $X$ and $Y$ be a discrete random variables both with support in $\mathbb{N}$. Then the random variable $S := X + Y$ is discrete with probability*

$$P(S = k) = \sum_{i=1}^{k} P(X = k - i)P(Y = i) \qquad \text{for all } k \in \mathbb{N}\,.$$

(b) *Let $X$ and $Y$ be absolutely continuous random variables with marginal densities $f_X$ and $f_Y$. Then the random variable $S := X + Y$ is absolutely continuous and has the density*

$$f_S\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_S(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy.$$

**Example 7.2.2.**

(a) Let $X$ and $Y$ be independent random variables distributed accordingly a $\operatorname{Bin}(m, p)$ and $\operatorname{Bin}(n, p)$ distribution. Then $X + Y$ has a $\operatorname{Bin}(m + n, p)$ distribution.

(b) Let $X$ and $Y$ be independent, random variables distributed accordingly a $\operatorname{Pois}(\alpha)$ and $\operatorname{Pois}(\beta)$ distribution. Then $X + Y$ has a $\operatorname{Pois}(\alpha + \beta)$ distribution.

(c) Let $X$ and $Y$ be independent random variables distributed accordingly a $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$ distribution. Then $X + Y$ has a $\mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distribution.

Instead of $X+Y$ in Corollary 7.2.1 one can consider other transformations such as $XY$, $X-Y$ etc. Note, that the chosen distributions in Example 7.2.2 are invariant under summation is a specific property of these distributions. For example, the exponential distribution does not have this property, see Exercise 7.3.3.

**Corollary 7.2.3.** *Let $X_1, \dots, X_n \colon \Omega \to \mathbb{R}$ be independent random variables with cumulative distribution function $F_{X_k} \colon \mathbb{R} \to [0,1]$ for $k = 1, \dots, n$. Then we have*

(a) $P\big(\max\{X_1, \dots, X_n\} \leqslant x\big) = F_{X_1}(x) \cdot \dots \cdot F_{X_n}(x)$ *for all* $x \in \mathbb{R}$.

(b) $P\big(\min\{X_1, \dots, X_n\} \leqslant x\big) = 1 - \big(1 - F_{X_1}(x)\big) \cdot \dots \cdot \big(1 - F_{X_n}(x)\big)$ *for all* $x \in \mathbb{R}$.

**Example 7.2.4.**

(a) Let $X_1, \dots X_n$ be independent random variables with $X_k$ distributed according to a $\mathrm{Geo}(\alpha_i)$ distribution. Then $Y = \min\{X_1, \dots, X_n\}$ has a $\mathrm{Geo}(\alpha_1 \cdot \dots \cdot \alpha_n)$ distribution.

(b) Let $X_1, \dots X_n$ be independent random variables with $X_k$ distributed according to an exponential distribution with parameter $\lambda_k > 0$. Then $Y = \min\{X_1, \dots, X_n\}$ has an exponential distribution with parameter $\lambda_1 + \dots + \lambda_n$.

**Theorem 7.2.5.** *Let $B_1, B_2, \dots$ be independent, Bernoulli distributed random variables with parameter $p \in (0,1)$. Then*

$$T := \min\{n \in \mathbb{N} :\ B_1 + \ldots + B_n = 0, B_{n+1} = 1\}$$

*is geometrically distributed with parameter* $1 - p$.

The random variable $T$ in Theorem 7.2.5 can be interpreted as the number of failures before success in the independent repetition of a Bernoulli trial.

**Example 7.2.6.** Russian roulette describes the lethal game where a player inserts a bullet into a revolver with 6 chambers . He spins the drum, points it to his head and pulls the trigger. If the trigger falls to an empty chamber he lives, he re-spins the drum and tries again.

Theorem 7.2.5 implies that the number $T$ of trials the player survives is geometrically distributed with parameter $\frac{5}{6}$. Consequently, Example 4.2.5 shows that the expectation of $T$ is 5.

## 7.3. Exercises

1. Let $(X, Y) \colon \Omega \to \mathbb{R}$ be a discrete random vector with support $\mathbb{N} \times \mathbb{N}$. Show that

   (a) $P(X + Y = k) = \displaystyle\sum_{i=1}^{k} P(X = k - i,\, Y = i)$ for all $k \in \mathbb{N}$.

   (b) $P(XY = k) = \displaystyle\sum_{i=1}^{k} P(X = k/i,\, Y = i)$ for all $k \in \mathbb{N}$.

2. Let $(X, Y) \colon \Omega \to \mathbb{R}$ be a random vector with support in $\{0, 1, 2\} \times \{0, 1, 2\}$ and distribution according following table:

| $P_{X,Y}$ | $\{Y=0\}$ | $\{Y=1\}$ | $\{Y=2\}$ |
|---|---|---|---|
| $\{X=0\}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ |
| $\{X=1\}$ | $\frac{2}{12}$ | 0 | 0 |
| $\{X=2\}$ | $\frac{2}{12}$ | $\frac{1}{12}$ | $\frac{4}{12}$ |

(a) Prove or disprove that $X$ and $Y$ are independent.

(b) Calculate expectation of $X$ and $Y$

(c) Derive the probability distribution of $X - Y$.

(d) Calculate $E[(X-Y)^2]$.

3. (a) Let $X$ and $Y$ be independent, random variables distributed accordingly a $\mathrm{Bin}(m,p)$ and $\mathrm{Bin}(n,p)$ distribution. Show that $X+Y$ has a $\mathrm{Bin}(m+n,p)$ distribution.

(b) Let $X$ and $Y$ be independent, random variables distributed accordingly a $\mathrm{Pois}(\alpha)$ and $\mathrm{Pois}(\beta)$ distribution. Show that $X+Y$ has a $\mathrm{Pois}(\alpha+\beta)$ distribution.

(c) Let $X$ and $Y$ be independent, random variables distributed accordingly an exponential distribution with parameters $\lambda_1$ and $\lambda_2$. Derive the density of $X+Y$ and explain that this is not an exponential distribution.

(d) Let $X$ and $Y$ be independent, random variables distributed accordingly a $\mathrm{Pois}(\alpha)$ and $\mathrm{Pois}(\beta)$ distribution. What can you say about the distribution of $X - Y$?

4. Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be a random vector with density $f\colon \mathbb{R}^2 \to \mathbb{R}_+$.

(a) Determine the density of the random vector $\big(XY,\, Y\big)\colon \Omega \to \mathbb{R}^2$.

(b) Let $X$ and $Y$ be independent and with densities

$$f_X\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_X(x) = \frac{1}{\pi\sqrt{1-x^2}}\mathbb{1}_{(-1,1)}(x),$$
$$f_Y\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_Y(y) = y\exp(-\tfrac{y^2}{2})\mathbb{1}_{(0,\infty)}(y).$$

Show that the random vector $\big(XY,\, Y\big)\colon \Omega \to \mathbb{R}^2$ has the density

$$g\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad g(v,w) = \begin{cases} \frac{1}{\pi\sqrt{1-\frac{v^2}{w^2}}}\exp(-\frac{w^2}{2}), & \text{if } w > |v|\,, \\ 0, & \text{else.} \end{cases}$$

(c) For the random variables $X$ and $Y$ defined in part (b), determine the density of the random variable $XY\colon \Omega \to \mathbb{R}$.

(d) For the random variables $X$ and $Y$ defined in part (b), calculate $E\big[XY^2\big]$.

5. An object appears on a circular radar screen with radius $r = 1$ as a point with coordinates $X$ and $Y$. Assume that the random vector $(X, Y)$ is uniformly distributed on the circular radar screen.
   (a) Determine the density of $(X, Y)$.
   (b) Calculate the densities of the marginals $X$ and $Y$.
   (c) Calculate the covariance of $X$ and $Y$.
   (d) Denote by $R := \sqrt{X^2 + Y^2}$ the distance to the centre of the radar screen.
      (i) Calculate the density of $R$;
      (ii) Calculate the expectation and variance of $R$.

6. Let the random variable $Z = (X, Y)\colon \Omega \to \mathbb{R}$ be uniformly distributed on the disk $D := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$. Show that the random variable $X/Y$ has density

$$f\colon \mathbb{R} \to \mathbb{R}_+, \qquad f(x) = \frac{1}{\pi(1 + x^2)}.$$

7. Let $X = (X_1, X_2, X_3)\colon \Omega \to \mathbb{R}$ be a random vector and assume that $X_1, X_2, X_3$ are independent and each $X_i$ is exponentially distributed with parameter $\lambda > 0$. Define the random variables

$$Y_1 = X_1 + X_2 + X_3, \quad Y_2 = \frac{X_1 + X_2}{Y_1} \quad Y_3 = \frac{X_1}{X_1 + X_2}.$$

   (a) Give reason why $X$ has a density and derive its form.
   (b) Given reason why $Y := (Y_1, Y_2, Y_3)$ has a density and derive its form.
   (c) Prove or disprove that $Y_1, Y_2, Y_3$ are independent.

8. Box-Muller simulation of normally distributed random variables

   Let $X_1$ and $X_2$ be two independent random variables, uniformly distributed on $[0, 1]$ and define

$$Y_1 := \sqrt{-2 \ln X_1}\, \cos(2\pi X_2), \qquad Y_2 := \sqrt{-2 \ln X_1}\, \sin(2\pi X_2)$$

   Show that $Y_1$ and $Y_2$ are independent, standard normally distributed random variables.

9. Let $X_1, \dots, X_n$ be independent, identically distributed random variables with exponential distribution with parameter $\alpha > 0$. Define the random variables

$$Y := \max\{X_1, \dots, X_n\}, \qquad Z := \min\{X_1, \dots, X_n\}.$$

   (a) Determine the cumulative distribution function of $Y$ and $Z$.
   (b) Determine the density of $Y$ and $Z$.

10. Let $X_1$ and $X_2$ be independent, exponentially distributed random variables with parameters $\alpha > 0$ and $\beta > 0$, respectively. Define the random variables

$$Y_1 := X_1 + X_2, \qquad Y_2 := \frac{X_1}{X_2}.$$

(a) Determine the density of the random vector $(Y_1, Y_2)\colon \Omega \to \mathbb{R}^2$.

(b) Show that $Y_1$ and $Y_2$ are independent, if $\alpha = \beta$.

# 8 Conditional expectation

In this chapter, we introduce conditional expectation. Although the name bears the word expectation it should not be confused with the expectation introduced in Section **??**.

The main motivation for conditional expectation is the following approximation or prediction problem: assume that a random variable $X$ cannot be observed but we want to predict its value by observing the values of another random variable $Y$. Equivalently, instead of $Y$, we can also think that we have only limited information available which our prediction of $X$ is based on. One can think of $X$ as the share price tomorrow and $Y$ as the share price today.

## 8.1. Inner product and norm in the Euclidean space

In this subsection we review some well known properties of vectors in $\mathbb{R}^d$. Let $\alpha = (\alpha_1, \dots, \alpha_d)$ and $\beta = (\beta_1, \dots, \beta_d)$ be vectors in $\mathbb{R}^d$. The inner product is defined by

$$\langle \alpha, \beta \rangle = \alpha_1 \beta_1 + \cdots + \alpha_d \beta_d.$$

The norm of the vector $\alpha$ is defined by

$$|\alpha| = \left( \alpha_1^2 + \cdots + \alpha_d^2 \right)^{1/2}.$$

Clearly, if $d = 2$ or $d = 3$ the norm $|\alpha|$ gives the length of the vector $\alpha$ but also in other dimensions we speak from the the length of the vector. The norm satisfies the triangle inequality and parallelogram law:

$$|\alpha + \beta| \leqslant |\alpha| + |\beta| \tag{8.1.1}$$

$$|\alpha + \beta|^2 + |\alpha - \beta|^2 = 2\left( |\alpha|^2 + |\beta|^2 \right) \tag{8.1.2}$$

for all $\alpha$, $\beta \in \mathbb{R}^d$. If $\vartheta \in [0, 2\pi)$ denotes the angle between the vectors $\alpha$ and $\beta$ we have

$$\langle \alpha, \beta \rangle = |\alpha|\,|\beta| \cos \vartheta, \tag{8.1.3}$$

which implies the Cauchy-Schwartz inequality in $\mathbb{R}^d$:

$$|\langle \alpha, \beta \rangle| \leqslant |\alpha|\,|\beta| \qquad \text{for all } \alpha, \beta \in \mathbb{R}^d.$$

For illustration purpose, we restrict ourselves to the case $d = 3$. Let $S$ be a line trough 0 in $\mathbb{R}^3$ and $x$ be a fixed point in $\mathbb{R}^3$ possibly outside of $S$. If we want to find the point $m$ on $S$ which has the minimal distance to $x$ we drop the perpendicular from $x$ to the line $S$. The intersection with $S$ gives the point $m$ with minimal distance to $x$. One also calls $m$ the projection $\pi_S(x)$ of $x$ onto the line $S$ and we have

$$|x - \pi_S(x)| = \inf\{|x - y| : y \in S\}.$$

A point $z$ is the projection of $x$ if and only if the vector from $x$ to $z$, that is $z - x$, is perpendicular to all vectors on $S$. Because of (8.1.3) we obtain:

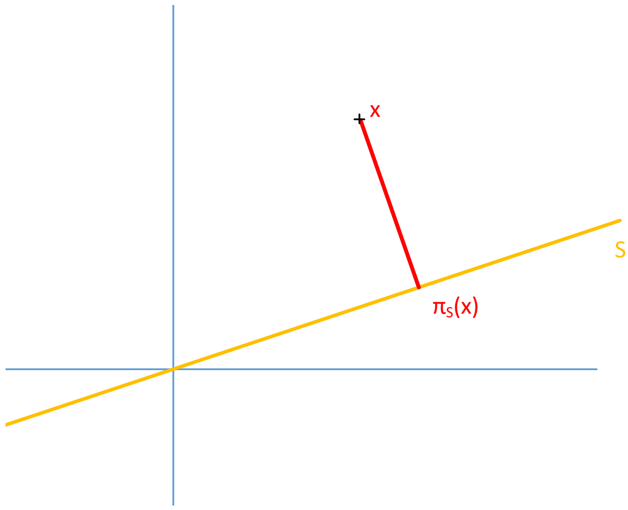



Figure 8.1.: projection in $\mathbb{R}^2$

$$z = \pi_S(x) \iff \langle z - x, y \rangle = 0 \quad \text{for all } y \in S. \tag{8.1.4}$$

These obvious facts in $\mathbb{R}^3$ can be easily generalised to each dimension $d$. Surprisingly, some infinite dimensional spaces exhibit similar structures.

Finally, we call the obvious fact, that convergence of vectors is defined by means of the norm. If $(\alpha_n)_{n \in \mathbb{N}}$ is a sequence of vectors $\alpha_n$ when we say that $\alpha_n$ converges to $\alpha$ in $\mathbb{R}^d$ if $|\alpha_n - \alpha| \to 0$ as $n \to \infty$, i.e. the distance between $\alpha_n$ and $\alpha$ decreases to 0.

## 8.2. The space of square-integrable random variables

The underlying probability space is denoted by $(\Omega, \mathscr{A}, P)$. Recall from Section 4.2 the definition of the space of square-integrable random variables:

$$\mathscr{L}^2(\Omega, P) := \{X : \Omega \to \mathbb{R} \text{ random variable such that } E[|X|^2] < \infty\}.$$

It is a linear space, which means that the sum of two random variables in $\mathscr{L}^2(\Omega, P)$ is again in $\mathscr{L}^2(\Omega, P)$ and the same applies to scalar multiplication. Some further rather obvious properties are required from a linear space; for the precise mathematical definition of a linear space see Appendix E.

**Definition 8.2.1.** *The function*

$$\langle\cdot,\cdot\rangle\colon \mathscr{L}^2(\Omega,P)\times\mathscr{L}^2(\Omega,P)\to\mathbb{R},\qquad \langle X,Y\rangle = E\big[XY\big]$$

*is called* inner product *on* $\mathscr{L}^2(\Omega,P)$.

Recall that by $XY$ we denote the random variable defined by the pointwise product $\omega\mapsto X(\omega)Y(\omega)$. Using the integral notation instead of expectation $E[\cdot]$ we have

$$\langle X,Y\rangle = \int_\Omega X(\omega)Y(\omega)\,P(d\omega) = E[XY].$$

Cauchy-Schwartz inequality from Theorem 4.2.8 guarantees that the inner product $\langle\cdot,\cdot\rangle$ is well-defined since

$$|\langle X,Y\rangle| \leqslant E\big[\,|XY|\,\big] \leqslant \big(E\big[X^2\big]\big)^{1/2}\big(E\big[Y^2\big]\big)^{1/2}.$$

It follows for all $X,Y,Z\in\mathscr{L}^2(\Omega,P)$ and $\alpha\in\mathbb{R}$ that

$$\begin{aligned}
&\langle X,X\rangle = 0 \;\Rightarrow\; X=0 \ P\text{-a.s.}\\
&\langle X,X\rangle \geqslant 0 \ P\text{-a.s.}\\
&\langle X,Y\rangle = \langle Y,X\rangle \ P\text{-a.s.}\\
&\langle X+Y,Z\rangle = \langle X,Z\rangle + \langle Y,Z\rangle \ P\text{-a.s.}\\
&\langle \alpha X,Y\rangle = \alpha\langle X,Y\rangle \ P\text{-a.s.}
\end{aligned}$$

**Definition 8.2.2.** *The function*

$$\|\cdot\|_2 : \mathscr{L}^2(\Omega,P)\to\mathbb{R}_+,\qquad \|X\|_2 = \Big(E\big[\,|X|^2\,\big]\Big)^{1/2}$$

*is called the* norm *and* $\|X\|_2$ *is called* norm of $X$.

Analogously to the Euclidean norm in (8.1.1) and (8.1.2), the norm in $\mathscr{L}^2(\Omega,P)$ satisfies the triangle inequality and parallelogram law:

**Lemma 8.2.3.** *Let $X$ and $Y$ be in $L^2(\Omega,P)$. Then we have*

(a) $\|X+Y\|_2 \leqslant \|X\|_2 + \|Y\|_2$.

(b) $\|X+Y\|_2^2 + \|X-Y\|_2^2 = 2\left(\|X\|_2^2 + \|Y\|_2^2\right)$.

Naturally, we interpret $\|X-Y\|$ as the distance between two random variables $X$ and $Y$ in $\mathscr{L}^2(\Omega,P)$, although this is not a distance in a geometric sense. The analogy with the Euclidean norm carries over to the meaning of convergence:

**Definition 8.2.4.** *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables in $\mathscr{L}^2(\Omega,P)$ converges to a random variable $X$ in $\mathscr{L}^2(\Omega,P)$ if*

$$\lim_{n\to\infty}\|X_n - X\|_2 = 0.$$

*We use the notation $X_n\to X$ in $\mathscr{L}^2(\Omega,P)$.*

The following property of $\mathscr{L}^2(\Omega, P)$ is rather advanced. We will admire its full power later when considering stochastic integration (FM04).

**Theorem 8.2.5.** *The linear space $\mathscr{L}^2(\Omega, P)$ is complete, which means the following: if a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables $X_n \in \mathscr{L}^2(\Omega, P)$ satisfies*

$$\lim_{m,n\to\infty} \|X_m - X_n\|_2 = 0,$$

*then there exists a random variable $X \in \mathscr{L}^2(\Omega, P)$ such that $X_n \to X$ in $\mathscr{L}^2(\Omega, P)$.*

The converse statement is rather simple: if the sequence $(X_n)_{n\in\mathbb{N}}$ converges to $X$ in $\mathscr{L}^2(\Omega, P)$, then Lemma 8.2.3.a implies:

$$\|X_m - X_n\|_2 = \|(X_m - X) + (X - X_n)\|_2 \leqslant \|X_m - X\|_2 + \|X - X_n\|_2 \to 0 \tag{8.2.5}$$

as $m, n \to \infty$.

In the remaining part we will extend the meaning of shortest distance and projection to the infinite dimensional space $\mathscr{L}^2(\Omega, P)$. Here, the role of the line in Section 8.1 is taken over by so-called closed subspaces defined in the following definition. Our main example of closed subspaces are introduced in Example 8.2.7.

**Definition 8.2.6.** *A subset $S$ of $\mathscr{L}^2(\Omega, P)$ is called a* closed subspace *if $S$ is a linear space and satisfies the condition:*

$$\textit{if } \{X_n\}_{n\in\mathbb{N}} \subseteq S \textit{ converges to } X \textit{ in } \mathscr{L}^2(\Omega, P) \textit{ then } X \in S.$$

**Example 8.2.7.** The following example of a closed subspaces will be used in the following subsections. Recall that the underlying probability space is $(\Omega, \mathscr{A}, P)$. For a $\sigma$-algebra $\mathscr{D}$ contained in $\mathscr{A}$ define

$$\mathscr{L}^2(\Omega, \mathscr{D}, P) := \big\{X : \Omega \to \mathbb{R} \text{ is } \mathscr{D}\text{-measurable and } E[|X|^2] < \infty\big\}.$$

For $\mathscr{D} = \mathscr{A}$ we obtain our previous definition, i.e. $\mathscr{L}^2(\Omega, P) = \mathscr{L}^2(\Omega, \mathscr{A}, P)$.

Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be $\sigma$-algebras with $\mathscr{D}_1 \subseteq \mathscr{D}_2 \subseteq \mathscr{A}$. Recall that $\mathscr{D}_1$-measurable means that $X^{-1}(B) \in \mathscr{D}_1$ in contrast to $X^{-1}(B) \in \mathscr{D}_2$ for all $B \in \mathfrak{B}(\mathbb{R})$ for $\mathscr{D}_2$-measurability. Thus, $\mathscr{D}_1$-measurable implies $\mathscr{D}_2$-measurable as $\mathscr{D}_1 \subseteq \mathscr{D}_2$, and we conclude

$$\mathscr{L}^2(\Omega, \mathscr{D}_1, P) \subseteq \mathscr{L}^2(\Omega, \mathscr{D}_2, P).$$

In particular, it follows that $\mathscr{L}^2(\Omega, \mathscr{D}, P)$ is a subset of $\mathscr{L}^2(\Omega, P)$ for each $\sigma$-algebra $\mathscr{D} \subseteq \mathscr{A}$.

Using our previous observation that $\mathscr{L}^1(\Omega, P)$ is a linear space shows that $\mathscr{L}^1(\Omega, \mathscr{D}, P)$ is a linear space. If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of random variables in $\mathscr{L}^2(\Omega, \mathscr{D}, P)$, converging to $X$ in $\mathscr{L}^2(\Omega, \mathscr{D}, P)$, then (8.2.5) shows that $\|X_m - X_n\|_2$ converges to 0 as $m, n \to \infty$. Consequently, applying Theorem 8.2.5 (for $(\Omega, \mathscr{A}, P)$ replaced by $(\Omega, \mathscr{D}, P)$) shows that $X \in \mathscr{L}^2(\Omega, \mathscr{D}, P)$. Consequently, $\mathscr{L}^2(\Omega, \mathscr{D}, P)$ is a closed subspace as defined in Definition 8.2.6.

**Theorem 8.2.8.** *Let $S$ be a closed subspace of $\mathscr{L}^2(\Omega, P)$. Then for each $X \in \mathscr{L}^2(\Omega, P)$ there exists a unique element $M \in S$ such that*

$$\|X - M\| = \inf\{\|X - Y\| : Y \in S\}.$$

The result of Theorem 8.2.8 can be described in words that for each random variable $X \in \mathscr{L}^2(\Omega, P)$ there exists a unique random variable $M \in S$ with the smallest distance to $X$, where the distance is measured by the norm. Such a result is not very surprising in $\mathbb{R}^2$ or $\mathbb{R}^d$ as observed in Section 8.1 but in an infinite dimensional space as $\mathscr{L}^2(\Omega, P)$ its proof requires some work. As in Section 8.1 we call the point $M$ the *projection of $X$ to the subspace $S$*:

**Definition 8.2.9.** *Let $S$ be a closed subspace of $\mathscr{L}^2(\Omega, P)$. Then the function*

$$\pi_S \colon \mathscr{L}^2(\Omega, P) \to S, \quad \textit{where } \|X - \pi_S(X)\| = \inf\{\|X - Y\| : Y \in S\}$$

*is called the* projection onto $S$. *That is, $\pi_S(X)$ is defined as the unique $M \in S$ described in Theorem 8.2.8.*

Note, that Theorem 8.2.8 guarantees that the projection is well defined. Thus, the projection $\pi_S$ maps any element $X \in \mathscr{L}^2(\Omega, P)$ to the element $\pi_S(X)$ in the closed subspace $S$ with the smallest distance to $X$. Clearly, if $X \in S$ then $\pi_S(X) = X$.

**Example 8.2.10.** (continues Example 8.2.7)
If $\mathscr{D}$ is a sub-$\sigma$-algebra of $\mathscr{A}$, that means $\mathscr{D}$ is a $\sigma$-algebra contained in $\mathscr{A}$, we can define the projection

$$\pi_{\mathscr{D}} \colon \mathscr{L}^2(\Omega, P) \to \mathscr{L}^2(\Omega, \mathscr{D}, P), \quad \text{where } \|X - \pi_S(X)\| = \inf\big\{ \|X - Y\| : Y \in \mathscr{L}^2(\Omega, \mathscr{D}, P)\big\}.$$

Thus, $\pi_{\mathscr{D}}(X)$ is the $\mathscr{D}$-measurable, square-integrable random variable with the smallest distance $\|X - \pi_{\mathscr{D}}(X)\|$ to $X$. In this sense, $\pi_D(X)$ approximates $X$.

The following result helps us later to identify a projection.

**Theorem 8.2.11.** *Let $S$ be a closed subspace of $\mathscr{L}^2(\Omega, P)$. Then for each $X \in \mathscr{L}^2(\Omega, P)$ and $Z \in S$ the following are equivalent:*

(a) *$Z = \pi_S(X)$;*

(b) *$\langle Z - X, Y\rangle = 0$ for all $Y \in S$.*

## 8.3. Conditional expectation

**Definition 8.3.1.** *Let $(\Omega, \mathscr{A}, P)$ be a probability space and $X \colon \Omega \to \mathbb{R}$ a random variable.*

(a) *If $\mathscr{D}$ is a sub-$\sigma$-algebra of $\mathscr{A}$, i.e. $\mathscr{D} \subseteq \mathscr{A}$, then a function $Z \colon \Omega \to \mathbb{R}$ is called a* conditional expectation of $X$ given $\mathscr{D}$ *if:*

  (1) *$Z$ is $\mathscr{D}$-measurable and $E[|Z|] < \infty$;*

  (2) *$E\big[Z \mathbb{1}_D\big] = E\big[X \mathbb{1}_D\big]$ for all $D \in \mathscr{D}$.*

  *Notation: $E[X|\mathscr{D}] := Z$.*

(b) *If $Y : \Omega \to \mathbb{R}$ is a random variable, then the* conditional expectation of $X$ given $Y$ *is defined as the conditional expectation of $X$ under $\sigma(Y)$.*
*Notation: $E[X|Y] := E[X|\sigma(Y)]$. (Recall Definition 3.2.1 of $\sigma(Y)$.)*

Note, that neither existence nor uniqueness of a conditional expectation is guaranteed at this stage.

**Remark 8.3.2.** We can interpret Condition (1) in Part (a) of Definition 8.3.1 that $Z$ is only based on information available in the $\sigma$-algebra $\mathscr{D}$. More specifically, if $\mathscr{D} = \sigma(Y)$ for a random variable as in Part (b), then Exercise 4.3.10 shows that there exists a measurable function $g\colon \mathbb{R} \to \mathbb{R}$ such that

$$E\big[X|Y\big] = g(Y),$$

i.e. the conditional expectation is only based on information obtained from observing $Y$. Condition (b) can be seen as the requirement that $Z$ approximates $X$ on sets $D \in \mathscr{D}$.

**Example 8.3.3.**

(a) Let $X$ be in $\mathscr{L}^1(\Omega, P)$ and $\mathscr{D} = \{\emptyset, \Omega\}$. Then it follows that

$$E[X|\mathscr{D}] = E[X] \quad P\text{-a.s.}$$

(b) Let $X$ be in $\mathscr{L}^1(\Omega, P)$ and $\mathscr{D} = \mathscr{A}$. Then it follows that

$$E[X|\mathscr{D}] = X \quad P\text{-a.s.}$$

(c) Let $X$ be in $\mathscr{L}^1(\Omega, P)$ and $\mathscr{D} = \sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$ for a set $A \subseteq \Omega$ with $P(A) \in (0,1)$. Then it follows that

$$E[X|\mathscr{D}](\omega) = \begin{cases} \frac{1}{P(A)} \int_A X \, dP, & \text{if } \omega \in A, \\ \frac{1}{P(A^c)} \int_{A^c} X \, dP, & \text{if } \omega \in A^c, \end{cases}.$$

(d) Example (c) can be generalised to the following: let $X$ be in $\mathscr{L}^1(\Omega, P)$ and

$$\mathscr{D} := \Big\{ \bigcup_{k \in I} C_k : \text{ for each } I \subseteq \mathbb{N} \Big\},$$

where $\{C_k : k \in \mathbb{N}\}$ is a partition of $\Omega$, i.e. the sets are pairwise disjoint and $\Omega = \cup_{k \in \mathbb{N}} C_k$, with $P(C_k) \in (0,1)$. Example 1.2.7 shows that $\mathscr{D}$ is a $\sigma$-algebra.

Then it follows that

$$E[X|\mathscr{D}] = \sum_{k=1}^{\infty} M_{C_k}(X) \mathbb{1}_{C_k},$$

where $M_{C_k}(X) := \frac{1}{P(C_k)} \int_{C_k} X \, dP$.

*Proof.* Define the random variable

$$Y := \sum_{k=1}^{\infty} M_{C_k}(X) \mathbb{1}_{C_k}.$$

Since $C_k \in \mathscr{D}$ for all $k \in I$ it follows that $Y$ is $\mathscr{D}$-measurable. Moreover,

$$E[|Y|] \leqslant \sum_{k=1}^{\infty} |M_{C_k}| \, P(C_k) \leqslant \sum_{k=1}^{\infty} \int_{C_k} |X| \, dP = E[|X|] < \infty.$$

For each $\ell \in I$ we have

$$\begin{aligned} \int_{C_\ell} Y \, dP &= \sum_{k=1}^{\infty} \frac{1}{P(C_k)} \int_{C_k} X \, dP \int_{C_\ell} \mathbb{1}_{C_k} \, dP \\ &= \frac{1}{P(C_\ell)} \int_{C_\ell} X \, dP \int_{C_\ell} \mathbb{1}_{C_\ell} \, dP = \int_{C_\ell} X \, dP. \end{aligned}$$

As all sets in $\mathscr{D}$ are countable unions of sets $C_i$ for $i \in \mathbb{N}$, linearity of the $P$-integral implies that

$$\int_C Y \, dP = \int_C X \, dP \qquad \text{for all } C \in \mathscr{D},$$

which shows that $Y = E[X|\mathscr{D}]$ $P$-a.s. □

The four examples illustrates the interpretation of the conditional expectation as a projection: the smaller the $\sigma$-Algebra $\mathscr{D}$ is the less values are attained by the conditional expectation.

**Proposition 8.3.4.** *Let $Z$ and $\tilde{Z}$ be two conditional expectations of a random variable $X \in \mathscr{L}^1(\Omega, P)$ given the sub-$\sigma$-algebra $\mathscr{D} \subseteq \mathscr{A}$. Then it follows that $P(Z = \tilde{Z}) = 1$, i.e. $Z = \tilde{Z}$ $P$-a.s.*

Proposition 8.3.4 shows that the conditional expectation is unique in a $P$-a.s. sense.

**Theorem 8.3.5.** *Let $X$ be in $\mathscr{L}^2(\Omega, \mathscr{A}, P)$ and $\mathscr{D} \subseteq \mathscr{A}$ a sub-$\sigma$-algebra. Then the conditional expectation of $X$ given $\mathscr{D}$ exists and satisfies*

$$E[X|\mathscr{D}] = \pi_{\mathscr{D}}(X) \quad P\text{-a.s.},$$

*where $\pi_{\mathscr{D}}$ is the projection $\pi_{\mathscr{D}} : L^2(\Omega, \mathscr{A}, P) \to L^2(\Omega, \mathscr{D}, P)$ (see Example 8.2.10).*

For a random variable $X$ in $\mathscr{L}^2(\Omega, \mathscr{A}, P)$ the conditional expectation $Z := E[X|\mathscr{D}]$ coincides with the projection $\pi_{\mathscr{D}}(X)$ according to Theorem 8.3.5. Thus, Definition 8.2.9 of a projection yields

$$E[|X - Z|^2] = \inf \left\{ E\left[ |X - Y|^2 \right] : Y \in \mathscr{L}^2(\Omega, \mathscr{D}, P) \right\}. \tag{8.3.6}$$

In other words, the conditional expectation is the best approximation of $X$ under all random variables in $\mathscr{L}^2(\Omega, \mathscr{D}, P)$. Here the quantifier "best" is measured by the norm in $\mathscr{L}^2(\Omega, \mathscr{A}, P)$, i.e. $E[|X - Z|^2]$. This interpretation is important in particular, if we think of $\mathscr{D}$ as the set of information which is available, and we approximate $X$ by random variables which carry less information.

Establishing the existence of the conditional expectation for $X \in \mathscr{L}^2(\Omega, P)$ follows immediately from the existence of the projection in $\mathscr{L}^2(\Omega, P)$, i.e. Theorem 8.2.8. However, by some limiting arguments one can improve this result to the case of only integrable random variables, i.e. $X \in \mathscr{L}^1(\Omega, P)$; recall that $\mathscr{L}^2(\Omega, P) \subseteq \mathscr{L}^1(\Omega, P)$; see (4.2.11). Although one cannot easily speak of projections in $\mathscr{L}^1(\Omega, P)$, as there is not an inner product $\langle \cdot, \cdot \rangle$, one can always think of the conditional expectation of $X$ as "coarsening" the random variable $X$.

**Theorem 8.3.6.** *Let $X$ be in $\mathscr{L}^1(\Omega, P)$ and $\mathscr{D} \subseteq \mathscr{A}$ a sub-$\sigma$-algebra. Then the conditional expectation $E[X|\mathscr{D}]$ exists and is a random variable in $\mathscr{L}^1(\Omega, \mathscr{D}, P)$.*

Theorem 8.3.6 improves Theorem 8.3.5 since we only have to assume that $X$ has finite expectation in order to guarantee the existence of the conditional expectation of $X$. However, if $X$ has a finite second moment, i.e. $X \in \mathscr{L}^2(\Omega, P)$ then Theorem 8.3.5 shows that the conditional expectation also has a finite second moment, whereas in the general situation in Theorem 8.3.6 the conditional expectation has only a finite expectation.

## 8.4. Properties of conditional expectation

**Theorem 8.4.1.** *Let $X$ and $Y$ be in $\mathscr{L}^1(\Omega, P)$ and $\mathscr{D} \subseteq \mathscr{A}$ a sub-$\sigma$-algebra. Then we have the following:*

(a) $E\big[E[X|\mathscr{D}]\big] = E[X]$ *$P$-a.s.;*

(b) *if $X$ is $\mathscr{D}$-measurable then $E[X|\mathscr{D}] = X$ $P$-a.s.*

(c) *linearity: $E[\alpha X + \beta Y|\mathscr{D}] = \alpha E[X|\mathscr{D}] + \beta E[Y|\mathscr{D}]$ $P$-a.s. for all $\alpha$, $\beta \in \mathbb{R}$.*

(d) *monotonicity: if $X \geqslant 0$ $P$-a.s. then $E[X|\mathscr{D}] \geqslant 0$ $P$-a.s.*

(e) *projection: if $\mathscr{D}_1 \subseteq \mathscr{D}_2$ sub-$\sigma$-algebras then $E\big[E[X|\mathscr{D}_2]\big|\mathscr{D}_1\big] = E[X|\mathscr{D}_1]$ $P$-a.s.*

(f) *taking out what is known: if $Y$ is $\mathscr{D}$-measurable then $E[XY|\mathscr{D}] = YE[X|\mathscr{D}]$ $P$-a.s.*

(g) *if $X$ is independent of $\mathscr{D}$ then $E[X|\mathscr{D}] = E[X]$ $P$-a.s.*

Also the limit theorems in Section 4.1 for integrals have analogue results for the conditional expectation.

**Theorem 8.4.2.** *Let $\mathscr{D} \subseteq \mathscr{A}$ be a sub-$\sigma$-algebra. Assume that $\{X_n\}_{n\in\mathbb{N}}$ is an increasing sequence of random variables in $\mathscr{L}^1(\Omega, P)$ with $X_n \geqslant 0$ $P$-a.s. for all $n \in \mathbb{N}$ and define $X := \sup_{n\in\mathbb{N}} X_n$. Then it follows that*

$$\lim_{n\to\infty} E[X_n|\mathscr{D}] = E[X|\mathscr{D}] \qquad P\text{-a.s.}$$

**Theorem 8.4.3.** *Let $\mathscr{D} \subseteq \mathscr{A}$ be a sub-$\sigma$-algebra. Assume that $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of random variables converging $P$-a.s. to a random variable $X$. If there exists $Y \in \mathscr{L}^1(\Omega, P)$ such that*

$$|X_n(\omega)| \leqslant Y(\omega) \qquad \text{for all } \omega \in \Omega \text{ and } n \in \mathbb{N},$$

*then the conditional expectation of $X$ exists and satisfies*

$$\lim_{n\to\infty} E[X_n|\mathscr{D}] = E[X|\mathscr{D}] \qquad P\text{-a.s.}$$

## 8.5. Computing conditional expectation

We derive some formulas to compute the conditional expectation in case of discrete or absolutely continuous random variables.

**Theorem 8.5.1.** *Let $X, Y$ be discrete random variables with $X$ taking values in $\{\gamma_1, \gamma_2, \dots\}$. Assume that $h\colon \mathbb{R} \to \mathbb{R}$ is a measurable function such that $E[|h(X)|] < \infty$. Define the function*

$$g\colon \mathbb{R} \to \mathbb{R}, \qquad g(y) := \begin{cases} \displaystyle\sum_{k=1}^{\infty} h(\gamma_k) P\big(X = \gamma_k | Y = y\big), & \text{if } P(Y=y) > 0, \\ 0, & \text{else.} \end{cases}$$

*Then it follows that*

$$E\big[h(X)|Y\big] = g(Y) \quad P\text{-a.s.}$$

For the above theorem, recall Definition 6.1.3 of conditional probability. We can rewrite the formula in Theorem 8.5.1 as:

$$\begin{aligned} \sum_{k=1}^{\infty} h(\gamma_k) P\big(X = \gamma_k | Y = y\big) &= \frac{1}{P(Y=y)} \sum_{k=1}^{\infty} h(\gamma_k) P(X = \gamma_k, Y = y) \\ &= \frac{1}{P(Y=y)} E\big[h(X)\mathbb{1}_{\{Y=y\}}\big]. \end{aligned}$$

The equality in the last line follows directly from Corollary 4.2.4. The last line does not refer any more to the assumption that $X$ is discrete and in fact, Theorem 8.5.1 can be generalised to general random variables $X$ in $\mathscr{L}^1(\Omega, P)$.

**Theorem 8.5.2.** *Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be an absolutely continuous random vector with a density $f\colon \mathbb{R}^2 \to \mathbb{R}_+$. Assume that $h\colon \mathbb{R} \to \mathbb{R}$ is a measurable function such that $E[|h(X)|] < \infty$. Define the function*

$$g\colon \mathbb{R} \to \mathbb{R}, \qquad g(y) := \begin{cases} \displaystyle\frac{1}{f_Y(y)} \int_{\mathbb{R}} h(x) f(x,y)\,dx, & \text{if } f_Y(y) > 0, \\ 0, & \text{else,} \end{cases}$$

*where $f_Y$ denotes the density of $Y$. Then it follows that*

$$E[h(X)|Y] = g(Y) \quad P\text{-a.s.}$$

If the random variable $X$ is $P$-integrable, Theorem 8.5.2 can be applied to determine the conditional expectation $E[X|Y]$.

## 8.6. Exercises

1. Let $X$ and $Y$ be independent random variables with Poisson distributions with parameters $\alpha >$ and $\beta > 0$. Calculate the conditional expectation $E[X|X+Y]$ by
   (a) using Example 8.3.3.d.
   (b) using Theorem 8.5.1.
2. Let $X$ and $Y$ be two independent random variables which are Bernoulli distributed with parameter $p \in (0,1)$. Compute for the random variable $Z := \mathbb{1}_{\{X+Y=0\}}$ the conditional expectation $E[X|Z]$ by
   (a) using Example 8.3.3.d.
   (b) using Theorem 8.5.1.
3. For random variables $X, Y \in \mathscr{L}^2(\Omega, P)$ one can define the *conditional variance* by
$$\operatorname{Var}[X|Y] := E\big[(X - E[X|Y])^2|Y\big].$$
   (a) Give reason that the conditional variance is well defined.
   (b) Show that we have
$$\operatorname{Var}[X|Y] = E[X^2|Y] - \big(E[X|Y]\big)^2 \quad P\text{-a.s.}$$
   (c) Show that we have
$$\operatorname{Var}[X] = E\big[\operatorname{Var}[X|Y]\big] + \operatorname{Var}\big[E[X|Y]\big].$$
4. Let $X$ and $Y$ be independent random variables which are exponentially distributed with parameter $\alpha > 0$. Determine a version of $E[X|X+Y]$.
5. Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be an absolutely continuous random vector with density
$$f\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f(x,y) = \frac{4y}{x^3}\mathbb{1}_{(0,1)}(x)\mathbb{1}_{(0,x^2)}(y).$$
   (a) Calculate the densities of $X$ and $Y$.
   (b) Determine the conditional expectation $E[Y|X]$.
   (c) Determine the conditional expectation $E[X^2|Y]$.
   (d) Determine the conditional expectation $E[X^2|Y^2]$.
6. Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be a normally distributed random vector with parameters
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\varrho \\ \sigma_1\sigma_2\varrho & \sigma_2^2 \end{pmatrix},$$
   for $\sigma_1, \sigma_2 > 0$ and $\varrho \in (-1,1)$. Show that
$$E[X|Y] = \varrho\frac{\sigma_1}{\sigma_2}Y \quad P\text{-a.s.}$$

7. Let $(X,Y)\colon \Omega \to \mathbb{R}^2$ be a random variable with joint density

$$f\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f(x,y) = xe^{-x(y+1)}\mathbb{1}_{[0,\infty)}(x)\mathbb{1}_{[0,\infty)}(y).$$

   (a) Find the densities of $X$ and $Y$.
   (b) Determine a version of the conditional expectation $E[Y|X]$.

# 9 Modes of convergence

In this chapter the underlying probability space is denoted by $(\Omega, \mathscr{A}, P)$.

## 9.1. Three modes of convergence

### 9.1.1. Almost sure convergence

A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is nothing else than a sequence of measurable functions $X_n\colon \Omega \to \mathbb{R}$. Consequently, one might consider the concept of pointwise convergence of functions as it is known from calculus:

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \qquad \text{for all } \omega \in \Omega.$$

However, in the probabilistic situation this convergence is too strong, as illustrated by the following example:

**Example 9.1.1.** A coin is tossed infinitely often. This random experiment can be modelled by the state space $\Omega = \{(\omega_1, \omega_2, \dots) : \omega_k \in \{0,1\}\}$, where $\omega_k$ denotes the outcome in the $k$-th toss and 0 corresponds that the coin lands tails. Let $X_n\colon \Omega \to \{0,1\}$ be defined by $X_n(\omega) = \omega_n$ for all $n \in \mathbb{N}$, i.e. $X_n$ represents the outcome of the $n$-th toss. If we assume that the probability for heads-up is $\frac{1}{2}$ we expect that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k(\omega) = \frac{1}{2}. \tag{9.1.1}$$

However, for the possible outcome $\omega = (0,0,0,\dots) \in \Omega$ this convergence is not true. In fact, if we define

$$A := \{\omega \in \Omega : \text{only finitely many heads occur}\}$$

it follows that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = 0 \qquad \text{for all } \omega \in A,$$

which contradicts our intuition (9.1.1). In fact, one can show that $P(A) = 0$, which together with the next Definition 3.3.4 puts this example in perspective.

Example 9.1.1 suggests that a reasonable concept of pointwise convergence for random variables should exclude $\omega \in \Omega$ which occur with probability 0. This is the definition of $P$-a.s. convergence as introduced in Definition 3.3.4, which we recall here:

**Definition 9.1.2.** *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables $X_n \colon \Omega \to \mathbb{R}$* converges $P$-a.s. *to a random variable $X \colon \Omega \to \mathbb{R}$ if*

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

The $P$-almost sure convergence can be equivalently formulated in the following way:

**Lemma 9.1.3.** *For random variables $X$, $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, the following are equivalent:*

(a) *$(X_n)_{n\in\mathbb{N}}$ converges $P$-a.s. to $X$;*

(b) $\lim\limits_{m\to\infty} P\Big(\sup\limits_{n\geqslant m} |X_n - X| \geqslant \varepsilon\Big) = 0$ *for all $\varepsilon > 0$.*

A simple criterium guaranteeing $P$-a.s. convergence is based on an application of the Borel-Cantelli Lemma 6.2.8:

**Corollary 9.1.4.** *If random variables $X$, $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, satisfy*

$$\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty \quad \textit{for all } \varepsilon > 0,$$

*then $(X_n)_{n\in\mathbb{N}}$ converges to $X$ $P$-a.s.*

### 9.1.2. Convergence in $p$-mean

In Section 8.2, the space $\mathscr{L}^2(\Omega, P)$ of square integrable random variables is introduced. We extend this definition for $p \geqslant 1$ by

$$\mathscr{L}^p(\Omega, P) := \{X : \Omega \to \mathbb{R} \text{ random variable such that } E[|X|^p] < \infty\}.$$

As in the case $p = 2$ we introduce

$$\|X\|_p := \big(E\big[\,|X|^p\,\big]\big)^{1/p}.$$

The triangle inequality in $\mathbb{R}$ shows $|X + Y| \leqslant |X| + |Y|$ and thus $\|X + Y\|_1 \leqslant \|X\|_1 + \|Y\|_1$. The analogue inequality for $p = 2$ is given in Lemma 8.2.3. In fact, one can derive (but we do not do) for each $p \geqslant 1$ and $X, Y \in \mathscr{L}^p(\Omega, P)$ that

$$\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p\,. \tag{9.1.2}$$

We extend Definition 8.2.4 to the case of arbitrary $p \geqslant 1$:

**Definition 9.1.5.** *Let $p \geqslant 1$ and assume that $X$, $X_n\colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, are random variables in $\mathscr{L}^p(\Omega, P)$. The sequence $(X_n)_{n\in\mathbb{N}}$* converges in $\mathscr{L}^p(\Omega, P)$ *or* in $p$-th mean to $X$ *if*

$$\lim_{n\to\infty} \|X_n - X\|_p = 0.$$

The most important cases of convergence in $p$-mean are for $p = 1$ and $p = 2$, which are called *mean convergence* or *square-mean convergence*, respectively. For random variables $X$, $X_n$ in $\mathscr{L}^2(\Omega, P)$, inequality (4.2.10) implies

$$\|X_n - X\|_1 \leqslant \|X_n - X\|_2 \,. \tag{9.1.3}$$

Thus, convergence in square-mean implies convergence in mean.

### 9.1.3. Convergence in probability

**Definition 9.1.6.** *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables $X_n\colon \Omega \to \mathbb{R}$* converges in probability *to a random variable $X\colon \Omega \to \mathbb{R}$ if for each $n \in \mathbb{N}$ we have*

$$\lim_{n\to\infty} P\left(|X_n - X| \geqslant \varepsilon\right) = 0.$$

*Convergence in probability is denoted by $X_n \xrightarrow{P} X$.*

If a sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to a random variable $X$ then the limit $X$ is $P$-a.s. unique. This follows since if $(X_n)_{n\in\mathbb{N}}$ converges in probability to $X$ and $Y$ then it follows for each $n \in \mathbb{N}$ that

$$P\left(|X - Y| \geqslant \tfrac{1}{n}\right) \leqslant P\left(|X - X_n| \geqslant \tfrac{1}{2n}\right) + P\left(|X_n - Y| \geqslant \tfrac{1}{2n}\right).$$

Thus, we have $\lim_{n\to\infty} P(|X - Y| \geqslant \frac{1}{n}) = 0$, and Part (g) in Theorem 1.2.13 implies

$$P(X = Y) = P(|X - Y| = 0) = P\left(\bigcap_{n=1}^{\infty} |X - Y| \leqslant \tfrac{1}{n}\right) = \lim_{n\to\infty} P\left(|X - Y| \leqslant \tfrac{1}{n}\right) = 1.$$

We compare convergence in probability with the two other kinds of convergence. It turns out that in both cases, convergence in probability is the weaker one.

**Theorem 9.1.7.** *For random variables $X$, $X_n\colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, the following implication holds true:*

$$\lim_{n\to\infty} X_n = X \quad P\text{-a.s.} \quad \Longrightarrow \quad X_n \xrightarrow{P} X.$$

The converse implication is not correct as the following example illustrates:

**Example 9.1.8.** Define a probability space by $\Omega = (0, 1]$, $\mathscr{A} = \mathfrak{B}(\mathbb{R}) \cap (0, 1]$ and let $P$ be the uniform distribution on $\Omega$. Define recursively for $n \in \mathbb{N}$:

$$a_1 = 0,\; b_1 = 1, \quad a_{n+1} = \begin{cases} b_n, & \text{if } b_n < 1, \\ 0, & \text{else,} \end{cases}, \quad b_{n+1} = \min\{a_{n+1} + \tfrac{1}{n+1}, 1\}.$$

Define for each $n \in \mathbb{N}$ the random variable

$$X_n \colon \Omega \to \mathbb{R}, \qquad X_n := \mathbb{1}_{(a_n, b_n]}.$$

Since for each $\varepsilon \in (0,1)$ and $n \in \mathbb{N}$ we have

$$P(|X_n| > \varepsilon) = P\big((a_n, b_n]\big) = b_n - a_n \leqslant \frac{1}{n},$$

it follows that $X_n \xrightarrow{P} 0$ for $n \to \infty$. However, since $X_n(\omega) = 1$ for infinitely many $n \in \mathbb{N}$, the sequence $(X_n(\omega))_{n\in\mathbb{N}}$ does not converge for any $\omega \in (0,1]$, and in particular, $(X_n)_{n\in\mathbb{N}}$ does not converge $P$-a.s.

Although convergence in probability does not imply $P$-a.s. convergence, one can always find a subsequence which converges $P$-a.s.

**Theorem 9.1.9.** *For random variables $X$, $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, the following are equivalent:*

(a) *$(X_n)_{n\in\mathbb{N}}$ converges in probability to $X$;*

(b) *for each subsequence $(X_{n_k})_{k\in\mathbb{N}}$ of $(X_n)_{n\in\mathbb{N}}$ there exists a further subsequence $(X_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ which converges $P$-a.s., that is*

$$\lim_{\ell\to\infty} X_{n_{k_\ell}} = X \quad P\text{-a.s.}$$

In particular, Theorem 9.1.9 implies that for each sequence $(X_n)_{n\in\mathbb{N}}$, converging in probability, there exists a subsequence which converges $P$-a.s. This result enables us to conclude that convergence in probability is invariant under continuous transformation:

**Corollary 9.1.10.** *Let $X$, $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, be random variables and $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then the following implication holds true:*

$$X_n \xrightarrow{P} X \implies f(X_n) \xrightarrow{P} f(X).$$

It remains to compare convergence in probability with convergence in $\mathscr{L}^p(\Omega, P)$. Also in this case it turns out that convergence in probability is the weaker notion.

**Theorem 9.1.11.** *For random variables $X$, $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$, and $p \geqslant 1$, the following implication holds true:*

$$\lim_{n\to\infty} \|X_n - X\|_p = 0 \implies X_n \xrightarrow{P} X.$$

The converse implication of Theorem 9.1.11 is not correct as the following example illustrates:

**Example 9.1.12.** Define a probability space by $\Omega = (0,1]$, $\mathscr{A} = \mathfrak{B}(\mathbb{R}) \cap (0,1]$ and let $P$ be the uniform distribution on $(0,1]$. Define for each $n \in \mathbb{N}$ the random variable

$$X_n \colon \Omega \to \mathbb{R}, \qquad X_n(\omega) := n\mathbb{1}_{(0,1/n]}(\omega).$$

Since $X_n(\omega) = 0$ for all $n \geqslant \frac{1}{\omega}$ it follows that $X_n \to 0$ $P$-a.s. and thus, in probability. However, since

$$E\left[|X_n|^p\right] = \int |X_n|^p \, dP = n^p P\big((0, \tfrac{1}{n}]\big) = n^{p-1},$$

the sequence $(X_n)_{n\in\mathbb{N}}$ does not converge in $L^p(\Omega, P)$ for any $p \geqslant 1$.

**Remark 9.1.13.** All definitions and results in Subsections 9.1.1, 9.1.2 and 9.1.3 can be extended to cover the case of random vectors $X_n$, $X\colon \Omega \to \mathbb{R}^d$.

## 9.2. Laws of large numbers

If a random experiment is independently repeated several times, one expects that the arithmetic mean of the random outcomes stabilises around the average. In this subsection, we give a precise meaning to what is meant by stabilising.

In this section, let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables $X_n\colon \Omega \to \mathbb{R}$. The $n$-partial sum is defined by

$$S_n := \sum_{k=1}^{n} X_k.$$

By stabilising we mean the convergence of the arithmetic mean:

$$\lim_{n\to\infty} \frac{1}{n} S_n. \tag{9.2.4}$$

If the convergence in (9.2.4) takes place $P$-a.s. we call it the *strong law of large numbers*, and if the convergence is in probability, it is called the *weak law of large numbers*. The convergence can also be additionally in $\mathscr{L}^p(\Omega, P)$, which depends strongly on the existence of finite moments of the random variables $(X_n)_{n\in\mathbb{N}}$.

**Theorem 9.2.1.** *(weak law of large numbers)*
*Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of uncorrelated random variables $X_n\colon \Omega \to \mathbb{R}$ with $E[X_k] = \mu$ and $\operatorname{Var}[X_k] \leqslant c$ for all $k \in \mathbb{N}$ for a constant $c > 0$. Then it follows that*

$$\lim_{n\to\infty} \frac{1}{n} S_n = \mu \qquad \text{in probability and in } \mathscr{L}^2(\Omega, P).$$

In Theorem 9.2.1, the random variables $X_n$ are not required to have the same distribution, but only the same expectation. Even the variances might differ but they must be uniformly bounded. Given this comment, it is obvious that the prerequisites in the following result are stronger but also implying a stronger result.

**Theorem 9.2.2.** *(strong law of large numbers)*
*Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables $X_n\colon \Omega \to \mathbb{R}$ with $E[X_1] = \mu$ and $\operatorname{Var}[X_1] = \sigma^2 < \infty$. Then it follows that*

$$\lim_{n\to\infty} \frac{1}{n} S_n = \mu \qquad P\text{-a.s. and in } \mathscr{L}^2(\Omega, P).$$

Because of the assumption $E[X_k^2] < \infty$ for all $k \in \mathbb{N}$ we obtain in the Theorems 9.2.1 and 9.2.2 that the convergence of the partial sum is not only in probability or $P$-a.s., but also in square-mean. In fact, for the strong law of large numbers one can abandon this condition and one obtains the following:

**Theorem 9.2.3.** *(strong law of large numbers by Kolmogorov)*
*Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables $X_n\colon \Omega \to \mathbb{R}$ with $E[X_1] = \mu$. Then it follows that*

$$\lim_{n\to\infty} \frac{1}{n} S_n = \mu \qquad P\text{-a.s. and in } \mathscr{L}^1(\Omega, P).$$

Theorem 9.2.3 can be proved very elegantly by martingales, see Stochastic Analysis. Due to a result by Etemadi (1981), one can even improve Theorem 9.2.3 and only require that the random variables $(X_n)_{n\in\mathbb{N}}$ are uncorrelated.

**Remark 9.2.4.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables $X\colon \Omega \to \mathbb{R}$. Then Theorem 9.2.3 implies for each $B \in \mathfrak{B}(\mathbb{R})$ that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_B(X_k) = P(X_1 \in B) \qquad P\text{-a.s.}$$

Thus, the left hand side, the so-called *empirical probability of the event $B$* or *relative frequency of $B$*, converges to the probability $P(B)$. In some sense, this justifies the axioms by Kolmogorov and enables to interpret the probabilities $P(B)$ in the frequentists' approach. (see Wikipedia frequentist probability).

Theorem 9.1.7 shows that the strong law of large numbers always implies the weak law of numbers. However, the converse conclusion is not correct as the following example illustrates.

**Example 9.2.5.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables $X_n\colon \Omega \to \mathbb{R}$ with

$$P(X_n = n) = P(X_n = -n) = \frac{1}{2n\ln(n+1)}, \qquad P(X_n = 0) = 1 - \frac{1}{n\ln(n+1)}$$

for all $n \in \mathbb{N}$. Then $(X_n)_{n\in\mathbb{N}}$ obeys the weak law of large numbers but not the strong law of large numbers.

There are numerous generalisations and modifications of the weak and strong law of large numbers. Some of them are the following:

(a) the requirement of identical distributions can be replaced by a condition on the asymptotic behaviour of the variances $\operatorname{Var}[X_n]$ for $n \to \infty$.

(b) instead of the scaling factor $1/n$ one can use other factors; see e.g the law of large numbers by Marcinkiewicz-Zygmund.

(c) there are law of large numbers for dependent random variables, if the dependency of the random variables is of a certain structure. In this case, the limit is often a random variable and not a constant.

(d) there are even weak laws of large numbers for random variables without a finite expectation; see Exercise 9.4.2.

## 9.3. Applications of the law of large numbers

#### Consistent estimators

One part of statistics treats the estimation of parameters of the unknown distribution by observing the independent repetitions of a random experiment. Often one knows that the probability distribtion of a random experiment is in a certain class, e.g. the normal distribution, but one does not know for example the expectation and variance.

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $E[X_1^2] < \infty$. The realisation $X_n(\omega)$ describes the outcome of the $n$-th repetition of the random experiment. It follows from the strong law of large number in Theorem 9.2.2 that

$$\bar{S}_n := \frac{1}{n} S_n \to E[X_1] \quad P\text{-a.s. and in } L_P^2(\Omega);$$
$$S_n^2 := \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{S}_n)^2 \to \sigma^2 \quad P\text{-a.s.}$$

Consequently, after observing $n$ outcomes of the random experiment, the values $\bar{S}_n(\omega)$ and $S_n^2(\omega)$ are reasonable estimates of $\mu$ and $\sigma^2$. The random variables $\bar{S}_n$ and $S_n^2$ are called *consistent estimators for the expectation and variance.*

Recall from Exercise 6.4.12 that $E[\bar{S}_n] = E[X_1]$ and $E[S_n^2] = \text{Var}[X_1]$.

#### Monte-Carlo approximation I

In finance, but also in many other applications, one often has to determine the expectation $E[X]$ of a random variable $X$. However, in many cases, the probability distribution of $X$ is unknown and it is impossible to calculate the expectation $E[X]$. Instead, one can simulate the random variable by independent repetition $X_1, \dots, X_n$ and the strong law of large numbers guarantees

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n X_k = E[X] \qquad P\text{-a.s. and in } \mathscr{L}^1(\Omega, P).$$

#### Monte-Carlo approximation II

In order to approximate the value of an integral $\int f(y)\,dy$ for a deterministic function $f\colon \mathbb{R} \to \mathbb{R}^d$, one often applies the so-called *Monte-Carlo approximation*. In many cases, these are multidimensional integrals, that is $d > 1$ but we restrict ourselves to the case for a function $f : [0,1] \to [0,1]$, and we want to approximate its integral

$$I(f) := \int_0^1 f(y)\,dy.$$

The strong law of large numbers enables us to approximate the value $I(f)$ by the simulation of random variables.

Let $X_1, Y_1, X_2, Y_2, \dots$ be independent random variables which are uniformly distributed on $[0,1]$. The random variables

$$Z_n \colon \Omega \to \mathbb{R}, \qquad Z_n := 1_{\{X_n \leqslant f(Y_n)\}},$$

define a sequence $(Z_n)_{n\in N}$ of independent, identically distributed random variables in $\mathscr{L}_2(\Omega, P)$ with

$$E[Z_1] = P\big(X_1 \leqslant f(Y_1)\big) = \iint_{[0,1]^2} \mathbb{1}_{\{0\leqslant x\leqslant f(y)\}}(x,y)dx\,dy = \int_0^1 f(y)\,dy.$$

The strong law of large numbers in Theorem 9.2.2 implies

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n Z_k = \int_0^1 f(y)\,\lambda(dy) \qquad P\text{-a.s. and in } \mathscr{L}^2(\Omega, P).$$

Since it is very easy to simulate uniformly distributed random variables, one obtains in this way an approximation of the integral $I(f)$. In typical application, the integrand is multidimensional where alternative deterministic numerical integration algorithms requires a lot computation time.

#### Empirical cumulative distribution function

One can approximate the cumulative distribution function $F_X$ of a random variable $X$ by observing the outcomes of independent random variables $X_n \colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ with the same probability distribution as $X$. The empirical cumulative distribution function is defined for each $x \in \mathbb{R}$ and $n \in \mathbb{N}$ by

$$F_n(x)\colon \Omega \to [0,1], \qquad F_n(x) := \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{X_k \leqslant x\}}.$$

By defining $Y_n := \mathbb{1}_{\{X_n \leqslant x\}}$ for fixed $x \in \mathbb{R}$ one obtains a sequence $(Y_n)_{n\in\mathbb{N}}$ of independent, identically distributed random variables with Bernoulli distribution with parameter $p := F(x)$. The strong law of large numbers in Theorem 9.2.2 implies

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n Y_k = F(x) \qquad P\text{-a.s.}$$

With some additional work one can even establish that the convergence is uniformly, that is

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0 \quad P\text{-a.s.}$$

This is the Theorem by Glivenko-Cantelli.

**Renewal theory**

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables in $\mathscr{L}^1(\Omega, P)$ with $\mu := E[X_1] > 0$. For each $t > 0$, the random variable

$$\tau(t)\colon \Omega \to \mathbb{R}_+ \cup \{\infty\}, \qquad \tau(t) = \inf\{n \in \mathbb{N} : S_n > t\},$$

denotes the first time when the partial sum $S_n := X_1 + \cdots + X_n$ exceeds $t$. With some additional work, the strong law of large numbers implies

$$\lim_{t\to\infty} \frac{\tau(t)}{t} = \frac{1}{\mu} \quad P\text{-a.s.}$$

A typical example of this random experiment is the model of a bulb, which is immediately replaced after it burnt out. In this model, the random variable $X_k$ denotes the lifetime of the $k$-th bulb, $S_n = X_1 + \cdots + X_n$ describes the total lifetime of the lamp with $n$ replacements, and $\tau(t)$ models the total number of bulbs, which have burnt out by time $t$.

## 9.4. Exercises

1. Let $X$, $X_n\colon \Omega \to \mathbb{R}^d$ be random variables in $\mathscr{L}_p(\Omega, P)$ for $p \geqslant 1$. Show the following implications:
   (a) $\lim_{n\to\infty} \|X_n - X\|_p = 0 \;\Rightarrow\; \lim_{n\to\infty} E\,|X_n|^p = E\,|X|^p$.
   (b) $\lim_{n\to\infty} \|X_n - X\|_1 = 0 \;\Rightarrow\; \lim_{n\to\infty} EX_n = EX$.
   Is the converse of this implication true?
   (c) $\lim_{n\to\infty} \|X_n - X\|_2 = 0 \;\Rightarrow\; \lim_{n\to\infty} \operatorname{Var} X_n = \operatorname{Var} X$.
2. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables and assume that each $X_n$ is absolutely continuous with density
   $$f_n\colon\ \mathbb{R} \to \mathbb{R}_+, \qquad f_n(x) = \frac{n}{\pi(1+n^2x^2)}.$$
   Show that
   (a) Show that $E[|X_n|^p] = \infty$ for all $n \in \mathbb{N}$ and for $p = 1$ and $p = 2$.
   (b) Show that $X_n \to 0$ in probability.
   (c) Show that $(X_n)_{n\in\mathbb{N}}$ does not converge $P$-a.s.
3. $(*)$ Let $X_n\colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be independent random variables which are distributed according to an exponential distribution with parameter $\alpha > 0$. Show that
   (a) $P\left(\dfrac{X_n}{\ln n} \geqslant \dfrac{1}{\alpha} \text{ infinitely often}\right) = 1$.
   (b) $P\left(\dfrac{X_n}{\ln n} \leqslant \dfrac{1}{\alpha} \text{ infinitely often}\right) = 1$.

(c) $\lim_{n\to\infty} \dfrac{X_n}{\ln n} = \dfrac{1}{\alpha} \quad P$-a.s.

4. (∗) Let $X_n\colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be random variables. Show that the sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability if and only if

$$\lim_{n\to\infty} \sup_{m\geqslant n} P\left(|X_m - X_n| \geqslant \varepsilon\right) = 0 \qquad \text{for all } \varepsilon > 0.$$

5. Let $X$, $X_n\colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be random variables.

   Show that the following two statements are equivalent:

   (1) $(X_n)_{n\in\mathbb{N}}$ converges in probability to $X$.

   (2) $\lim_{n\to\infty} E[|X_n - X| \wedge 1] = 0$.

   This result shows that $(X_n)$ converges in probability to 0 if and only if $d(X_n, X) := E[|X_n - X| \wedge 1] \to 0$. The function $d$ defines a so-called metric on the space of all random variables and thus, convergence in probability can be described by this metric.

6. Let $p_n \in (0,1)$ and $X_n\colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be independent random variables with

$$P(X_n = 1) = p_n, \quad P(X_n = 0) = 1 - p_n.$$

   Find for the following statements an equivalent condition in terms of the sequence $(p_n)_{n\in\mathbb{N}}$

   (a) $(X_n)_{n\in\mathbb{N}}$ converges to 0 in probability .

   (b) $(X_n)_{n\in\mathbb{N}}$ converges to 0 in $L^p_P(\Omega)$ for $p \geqslant 1$.

   (c) $(X_n)_{n\in\mathbb{N}}$ converges $P$-a.s.to 0.

7. Let $X_n\colon \Omega \to \mathbb{R}$ for $n \in \mathbb{N}$ be independent random variables and assume that each $X_n$ is uniformly distributed on the interval $[-\frac{1}{n}, \frac{1}{n}]$ for $n \in \mathbb{N}$. Show that $(X_n)_{n\in\mathbb{N}}$ converges in probability to 0.

8. Let $U\colon \Omega \to \mathbb{R}$ be a random variable with a uniform distribution on the interval $[1,2]$ and define for each $n \in \mathbb{N}$ a random variable by

$$X_n\colon \Omega \to \mathbb{R}, \qquad X_n(\omega) = \begin{cases} 1, & \text{if } U(\omega) \in \left[1, 2 - \frac{1}{n}\right], \\ 0, & \text{else.} \end{cases}$$

   Show that $(X_n)_{n\in\mathbb{N}}$ converges in $\mathscr{L}^2(\Omega, P)$.

9. Let $X$ and $X_n$ for $n \in \mathbb{N}$ be random variables in $\mathscr{L}^1(\Omega, P)$. Show that if $X_n \to X$ in $\mathscr{L}^1(\Omega, P)$, then $E[X_n|\mathscr{D}] \to E[X|\mathscr{D}]$ in $\mathscr{L}^1(\Omega, P)$ for any $\sigma$-algebra $\mathscr{D} \subseteq \mathscr{A}$.

# A Set operations

Let $\Omega$ be a non-empty set, e.g. any example of a sample space in Section 1.1.

Relations between two sets $A$, $B$ are the following:

- the set $A$ is a subset of set $B$ if the following is true:

$$\omega \in A \;\Rightarrow\; \omega \in B.$$

  The notation is $A \subseteq B$.

- the set $A$ equals the set $B$ if the following is true:

$$\omega \in A \;\Leftrightarrow\; \omega \in B.$$

  The notation is $A = B$.

Some standard set operations are the following:

- the complement of a set $A \subseteq \Omega$ is $A^c := \{\omega \in \Omega : \omega \notin A\}$;
- the intersection of two sets $A$, $B \subseteq \Omega$ is $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$;
- the union of two sets $A$, $B \subseteq \Omega$ is $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$;
- the set difference of two sets $A$, $B \subseteq \Omega$ with $A \subseteq B$ is $B \backslash A = \{\omega \in \Omega : \omega \in B \text{ and } \omega \notin A\}$;

Set operations can alwyas illustrated by a so-called Venn diagram. An example is presented in figure A.1.

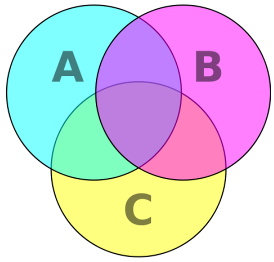



Figure A.1.: Venn diagram for 3 sets

Set operations satisfy some well known properties. Some of those are obvious, whereas others requires some thoughts. Let $A$, $B$, $C \subseteq \Omega$:

- Complements: $(A^c)^c = A$;
  $\emptyset^c = \Omega$;
  $\Omega^c = \emptyset$.
- Commutativity: $A \cap B = B \cap A$, $A \cup B = B \cup A$;
  $A \cap \emptyset = \emptyset$, $A \cup \emptyset = A$;
  $A \cap A = A, ; A \cup A = A$
  $A \cap \Omega = A$, $A \cup \Omega = \Omega$;
  $A \cap A^c = \emptyset$, $A \cup A^c = \Omega$.
- Associativity: $(A \cap B) \cap C = A \cap (B \cap C)$;
  $(A \cup B) \cup C = A \cup (B \cup c)$.
- Distributivity: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$;
  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.
- De Morgan's laws: $(A \cap B)^c = A^c \cup B^c$;
  $(A \cup B)^c = A^c \cap B^c$.

# B Indicator function

Let $\Omega$ be a non-empty set and $A$ a subset of $\Omega$. The indicator function is defined by

$$\mathbb{1}_A \colon \Omega \to \mathbb{R}, \qquad \mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

In many cases $\Omega = \mathbb{R}$, in which $A$ could be an interval $(a, b]$ or just an elementary set $\{a\}$.

Between set operations on the sets and algebra operations on the indicator function, one has the followin relations for all subsets $A$, $B \subseteq \Omega$:

$$\begin{aligned} \mathbb{1}_{A^c} &= 1 - \mathbb{1}_A & \qquad \mathbb{1}_{B \setminus A} &= \mathbb{1}_B(1 - \mathbb{1}_A) \\ \mathbb{1}_{A \cap B} &= \min\{\mathbb{1}_A, \mathbb{1}_B\} = \mathbb{1}_A \mathbb{1}_B & \qquad \mathbb{1}_{A \cup B} &= \max\{\mathbb{1}_A, \mathbb{1}_B\}. \end{aligned}$$

# C
# Functions and pre-images

Let $E$ and $F$ be some non-empty sets.

**Definition C.0.1.** *A function from a set $E$ to a set $F$ is an object $f$ such that every $x \in E$ is uniquely associated with an element $f(x) \in V$. The set $E$ is called the* domain *and the set $E$ is called the* codomain*, while the set $f(E)$ is called the* range of $f$.

In this note, a function $f$ is often denoted by

$$f\colon E \to F, \qquad f(x) = \text{the rule}.$$

Typical examples are functions which are defined on $E = \mathbb{R}$ with values in $F = \mathbb{R}^d$. But the domain $E$ can be each set. For example, a probability measure is defined on a set of sets, i.e. the $\sigma$-algebra.

Let $f\colon E \to F$ be a function. The pre-image under $f$ of a subset $B \subseteq F$ is defined as

$$f^{-1}(B) := \{x \in E :\, f(x) \in B\}.$$

Note, that $f^{-1}(B)$ is a subset of the domain $E$, where the mapping $f$ is defined.

**Lemma C.0.2.** *Let $f\colon E \to F$ be a function. Then we have:*

(a) $f^{-1}(F) = E$;

(b) $f^{-1}(\emptyset) = \emptyset$;

(c) *If $B_1$, $B_2 \subseteq F$ are disjoint, then $f^{-1}(B_1)$ and $f^{-1}(B_2)$ are disjoint;*

(d) *If $B$ is a subset of $F$ then $f^{-1}\big(B^c\big) = \big(f^{-1}(B)\big)^c$;*

(d) *If $B_1 \subseteq B_2 \subseteq F$ then $f^{-1}(B_1) \subseteq f^{-1}(B_2)$;*

(e) *if $B_1, B_2, \ldots$ are subsets of $F$ then*

$$f^{-1}\left(\bigcup_{k=1}^{\infty} B_k\right) = \bigcup_{k=1}^{\infty} f^{-1}(B_k);$$

(f) *if $B_1, B_2, \ldots$ are subsets of $F$ then*

$$f^{-1}\left(\bigcap_{k=1}^{\infty} B_k\right) = \bigcap_{k=1}^{\infty} f^{-1}(B_k);$$

(g) *if $G$ is another space and $g\colon F \to G$ is another function then $(g \circ f)^{-1}(C) = f^{-1}(g^{-1})(C)$ for all $C \subseteq G$.*

# D Sums and matrices

**Definition D.0.1.** *For a sequence $(\alpha_k)_{k\in\mathbb{N}}$ of real numbers define $s_n := \alpha_1 + \cdots + \alpha_n$.*

(a) *the series $\sum_{k=1}^{\infty} \alpha_k$ is called convergent if $s_n$ converges to a finite limit for $n \to \infty$.*

(b) *the series $\sum_{k=1}^{\infty} \alpha_k$ is called absolutely convergent if the series $\sum_{k=1}^{\infty} |\alpha_k|$ is convergent.*

**Definition D.0.2.** *For $n \in \mathbb{N}$ and $k \in \{0, \ldots, n\}$ the binomial coefficient is defined by*

$$\binom{n}{k} := \frac{n!}{(n-k)!k!} \qquad \textit{read "n choose k".}$$

*The number $\binom{n}{k}$ is the number of ways to choose $k$ elements from a set with $n$ members.*

Apart from its application in combinatorics, the binomial coefficient appears in the binomial formula:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \qquad \text{for all } x,\, y \in \mathbb{R}.$$

**Definition D.0.3.** *The determinant $det(M)$ of a matrix $M = (m_{i,j})_{i,j=1}^{d}$ in $\mathbb{R}^{d\times d}$ is defined by*

$$det[M] := \sum_{\sigma=(\sigma_1,\ldots,\sigma_d)} \operatorname{sgn}(\sigma) \prod_{i=1}^{d} m_{i,\sigma_i},$$

*where the sum is taken over all permutations $(\sigma_1, \ldots, \sigma_d)$ of $\{1, \ldots, d\}$ and $sgn(\sigma)$ is $+1$ if the permutation $\sigma$ can be achieved by by successively interchanging two entries an even number of times, and $-1$ whenever it can be achieved by an odd number of such interchanges.*

**Theorem D.0.4.** *Let $M$ be a matrix in $\mathbb{R}^{d\times d}$.*

(a) $det[M] = det\left[M^T\right]$.

(b) *If $M$ is invertible then* $det\left[M^{-1}\right] = \dfrac{1}{det[M]}$.

**Definition D.0.5.** *Let $M$ be a matrix in $\mathbb{R}^{d\times d}$.*

(a) *Let $M$ be of the form $M = (m_{i,j})_{i,j=1}^d$. Then the transpose $M^T$ of $M$ is defined by*

$$M^T := (m_{j,i})_{j,i=1}^d.$$

*(Reflect $M$ over its main diagonal).*

(b) *The matrix $M$ is called* positive definite *if*

$$v^T M v > 0 \qquad \textit{for all } v \in \mathbb{R}^d \backslash \{0\}.$$

(a) *The matrix $M$ is called* symmetric *if $M = M^T$.*

# E
# Linear space

You probably call many sets intuitively *spaces* but below you find the mathematical definition.

**Definition E.0.1.** *A linear space is a set $V$ together with two operations $+$ (addition) and $\cdot$ (scalar multiplication) such that for all $u, v, w \in V$ and $\alpha, \beta \in \mathbb{R}$ the following is satisfied:*

(a) $u + v \in V$;

(b) $\alpha \cdot v \in V$;

(c) $u + v = v + u$;

(d) $u + (v + w) = (u + v) + w$;

(e) *there exists an element in $V$, denoted by $0$, such that: $u + 0 = 0 + u = u$;*

(f) *for every $u \in V$ there is another object, denoted by $-u$ such that $u + (-u) = 0$;*

(g) $\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$;

(h) $(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$;

(i) $\alpha \cdot (\beta \cdot u) = (\alpha\beta) \cdot u$;

(j) $1 \cdot u = u$;

# 10 Solutions

## 10.1. Chapter 1

1. (a) Suppose $\omega \in (A_1 \cap A_2)^c$. Then $\omega \notin A_1 \cap A_2$, which implies $\omega \notin A_1$ or $\omega \notin A_2$. Hence, $\omega \in A_1^c$ or $\omega \in A_2^c$ which is $\omega \in A_1^c \cup A_2^c$.

   Suppose $\omega \in (A_1^c \cup A_2^c)$. Then $\omega \in A_1^c$ or $\omega \in A_2^c$, which implies $\omega \notin A_1$ or $\omega \notin A_2$. Hence, $\omega \notin A_1 \cap A_2$, which is $\omega \in (A_1 \cap A_2)^c$.

   The second relation follows analogously.

   (b) It is basically the same arguments as in part (a).

   Suppose $\omega \in (\cap A_k)^c$. Then there exists $k_0 \in \mathbb{N}$ such that $\omega \notin A_{k_0}$. This implies that $\omega \in A_{k_0}^c$ and thus $\omega \in \cup A_k^c$.

   Suppose $\omega \in \cup A_k^c$. Then there exists $k_0 \in \mathbb{N}$ such that $\omega \in A_{k_0}^c$. This implies that $\omega \notin A_{k_0}$, which implies $\omega \notin \cap A_k$ and thus $\omega \in (\cap A_k)^c$.

   The second relation follows analogously.

2. (a) $\displaystyle\bigcup_{k=1}^{n} A_k$. (b) $\displaystyle\bigcap_{k=1}^{n} A_k^c$.

   (c) Let $B_k$ denote the event that only the $k$-th memory is broken but all other $n-1$ memory sticks work. It follows for each $k \in \{1, \ldots, n\}$ that

$$B_k = A_k \bigcap \Big( \bigcap_{\substack{j=1 \\ j \neq k}}^{n} A_j^c \Big).$$

It follows that the event of exactly one memory stick is broken is described by

$$\bigcup_{k=1}^{n} B_k = \bigcup_{k=1}^{n} \Big(A_k \bigcap \Big(\bigcap_{\substack{j=1\\ j\neq k}}^{n} A_j^c\Big)\Big).$$

(d) By part (c) one obtains

$$(\bigcup_{k=1}^{n} B_k) \bigcup (\bigcap_{k=1}^{n} A_k^c) = \bigcup_{k=1}^{n} \Big(A_k \bigcap \Big(\bigcap_{\substack{j=1\\ j\neq k}}^{n} A_j^c\Big)\Big) \bigcup (\bigcap_{k=1}^{n} A_k^c)$$

3. (a) A male student who smokes or a male student who does not live in the student hall.

(b) All male students do not smoke and live in a student hall.

(c) There is a student who does not live in a student hall but smokes or there is a student who does not smoke but lives in a student hall.

(d) $A^c = B$ means that all female students do not smoke and that all male students smoke (with the assumption that there are either male or female students). The event that all male students smoke is described by $A \subseteq B^c$ and it is a strict subset if there exists one female student who smokes.

4. (a) $\Omega = \{\omega = (\omega_1, \ldots, \omega_4) : \omega_k \in \{h, t\} \text{ for } k = 1, \ldots, 4\}$.

(b) $A = \{(\omega_1, \ldots, \omega_4) \in \Omega : \omega_1 = \omega_2 = t\}$ and $B = \{(\omega_1, \ldots, \omega_4) \in \Omega : \omega_1 = \omega_3 = t\}$.

(c) The set $A \cup B$ describes the event that the coin lands with tail up in the first two tosses or in the first and third toss. The set $A \cap B$ describes the event that the coin lands with tail up in the first three tosses.

$A \cup B = \{(t,t,t,t), (t,t,t,h), (t,t,h,h), (t,t,h,t), (t,h,t,h), (t,h,t,t)\}$

$A \cap B = \{(t,t,t,h), (t,t,t,t)\}$.

5. (a) For any $\omega \in \Omega$ it follows that:

$$\omega \in f^{-1}(B^c) \;\Leftrightarrow\; \exists x \in B^c : f(\omega) = x \;\Leftrightarrow\; f(\omega) \notin B \;\Leftrightarrow\; \omega \in (f^{-1}(B))^c.$$

(b) For any $\omega \in \Omega$ we have:

$$\begin{aligned}
\omega \in f^{-1}\Big(\bigcup_{k\in I} B_k\Big) \;&\Leftrightarrow\; \exists x \in \bigcup_{k\in I} B_k : f(\omega) = x \\
&\Leftrightarrow\; \exists k \in I \,\exists x \in B_k : f(\omega) = x \\
&\Leftrightarrow\; \exists k \in I : f(\omega) \in B_k \\
&\Leftrightarrow\; \omega \in \bigcup_{k\in I} f^{-1}(B_k).
\end{aligned}$$

$$\begin{aligned}\omega \in f^{-1}\left(\bigcap_{k\in I} B_k\right) &\Leftrightarrow \exists x \in \bigcap_{k\in I} B_k : f(\omega) = x\\ &\Leftrightarrow \forall k \in I\, \exists x \in B_k : f(\omega) = x\\ &\Leftrightarrow \forall k \in I : f(\omega) \in B_k\\ &\Leftrightarrow \omega \in \bigcap_{k\in I} f^{-1}(B_k).\end{aligned}$$

(c) Since $E \in \mathscr{E}$ and $f^{-1}(E) = \Omega \in \mathscr{A}$ we have $E \in \mathscr{C}$. Let $B \in \mathscr{C}$. Then $f^{-1}(B) \in \mathscr{A}$ and thus $(f^{-1}(B))^c \in \mathscr{A}$ since $\mathscr{A}$ is a $\sigma$-algebra. By part (a) it follows that $f^{-1}(B^c) = (f^{-1}(B))^c \in \mathscr{A}$ and thus $B^c \in \mathscr{C}$. If $B_1, B_2, \ldots \in \mathscr{C}$ then $f^{-1}(B_1), f^{-1}(B_2), \cdots \in \mathscr{A}$ and thus $\cup f^{-1}(B_k) \in \mathscr{A}$ since $\mathscr{A}$ is a $\sigma$-algebra. By part (b) it follows that

$$f^{-1}\left(\bigcup_{k=1}^{\infty} B_k\right) = \bigcup_{k=1}^{\infty} f^{-1}(B_k) \in \mathscr{A},$$

which shows that $\cup B_k \in \mathscr{C}$.

6. Since $B = \Omega \cap B$ it follows that $B \in \mathscr{A} \cap B$. If $C \in \mathscr{A} \cap B$ then there is $A \in \mathscr{A}$ such that $C = A \cap B$. It follows that $C^c = A^c \cap B$ (since we take the complement in $B$, that is $C^c = B \backslash C$) and thus $C^c \in \mathscr{A} \cap B$ as $A^c \in \mathscr{A}$. If $C_1, C_2, \ldots \in \mathscr{A} \cap B$ then there exists $A_1, A_2, \ldots \in \mathscr{A}$ such that $C_k = A_k \cup B$. It follows that

$$\bigcup_{k=1}^{\infty} C_k = \bigcup_{k=1}^{\infty} (A_k \cap B) = \left(\bigcup_{k=1}^{\infty} A_k\right) \cap B.$$

Hence, $\cup C_k \in \mathscr{A} \cap B$.

7. $\sigma(\mathscr{C}) = \{\emptyset, E, \{a, b\}\, \{c, d, e\}, \{f\}, \{a, b, f\}, \{c, d, e, f\}, \{a, b, c, d, e\}\}$.

8. Recall that the Borel $\sigma$-algebra is generated by all open intervals $(a, b)$ for $a \leqslant b$.

(a) The interval $(-\infty, b)$ is in $\mathfrak{B}(\mathbb{R})$ as $(-\infty, b) = \cup_{n=1}^{\infty}(-n, b)$ and $(-n, b) \in \mathfrak{B}(\mathbb{R})$ for all $n \in \mathbb{N}$. It follows that $[b, \infty)$ is in $\mathfrak{B}(\mathbb{R})$ as $[b, \infty) = (-\infty, b)^c$. Consequently, we have $[a, b) \in \mathfrak{B}(\mathbb{R})$ as $[a, b) = \big((-\infty, a) \cup [b, \infty)\big)^c$. Since $[a, b] = \cap_{n=1}^{\infty}[a, b + \frac{1}{n})$ we have $[a, b] \in \mathfrak{B}(\mathbb{R})$.

(b) Since the interval $[x, x + \frac{1}{n}]$ is in $\mathfrak{B}(\mathbb{R})$ for each $x \in \mathbb{R}$ according to $(a)$ we conclude that $\{x\} = \bigcap_{n=1}^{\infty} \left[x, x + \frac{1}{n}\right] \in \mathfrak{B}(\mathbb{R})$.

(c) The representation $\mathbb{N} = \cup_{n=1}^{\infty}\{n\}$ implies $\mathbb{N} \in \mathfrak{B}(\mathbb{R})$ by part (a).

(d) The representation $\mathbb{Q} = \cup_{p=-\infty}^{\infty} \cup_{q=1}^{\infty} \left\{\frac{p}{q}\right\}$ implies $\mathbb{Q} \in \mathfrak{B}(\mathbb{R})$ by part (a).

9. Since $\Omega = f^{-1}(\mathbb{R})$ and $\mathbb{R} \in \mathfrak{B}(\mathbb{R})$ it follows that $\Omega \in \mathscr{A}$. Let $A$ be in $\mathscr{A}$. Then there exists $B \in \mathfrak{B}(\mathbb{R})$ with $A = f^{-1}(B)$. It follows (see Exercise 1.3.5) that $A^c = f^{-1}(B^c)$ and

thus $A^c \in \mathscr{A}$ since $B^c \in \mathfrak{B}(\mathbb{R})$. If $A_1, A_2, \ldots \in \mathscr{A}$ then there exist $B_1, B_2, \ldots \in \mathfrak{B}(\mathbb{R})$ with $A_k = f^{-1}(B_k)$ for all $k \in \mathbb{N}$. Exercise 1.3.5 implies that

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} f^{-1}(B_k) = f^{-1}\left(\bigcup_{k=1}^{\infty} B_k\right).$$

Since $\cup B_k \in \mathfrak{B}(\mathbb{R})$ it follows that $\cup A_k \in \mathscr{A}$.

10. 1.(a) $\sigma(\{A,B\}) = \{\emptyset, \Omega, A, B, A^c, B^c, A\cup B, (A\cup B)^c\}$.

    1.(b) $\sigma(\{A,B\}) = \{\emptyset, \Omega, A\backslash B, B\backslash A, A\cap B, (A\cup B)^c, (A\cup B)\backslash(A\cap B), A, (A\backslash B)\cup(A\cup B)^c, B, (B\backslash A)\cup(A\cup B)^c, (A\cap B)\cup(A\cup B)^c, A\cup B, (A\cap B)^c, ....\}$.

    (2) Define $\mathscr{D} := \left\{\bigcup_{i\in I} C_i : \text{ for each } I \subseteq \mathbb{N}\right\}$. First, we show that $\mathscr{D}$ is a $\sigma$-algebra: Since $\Omega = \cup_{i\in\mathbb{N}} C_i$ it follows that $\Omega \in \mathscr{D}$. If $A \in \mathscr{D}$ then there exists $I \subseteq \mathbb{N}$ such that $A = \cup_{i\in I} C_i$. It follows that

    $$A^c = \Omega\backslash A = \left(\bigcup_{i\in\mathbb{N}} C_i\right)\backslash\left(\bigcup_{i\in I} C_k\right) = \bigcup_{i\in N\backslash I} C_i,$$

    and thus $A^c \in \mathscr{E}$. If $A_1, A_2, \ldots \in \mathscr{D}$ are pairwise disjoint then there exits $I_1, I_2, \ldots \subseteq I$ pairwise disjoint such that $A_\ell = \cup_{i\in I_\ell} C_i$. It follows that

    $$\bigcup_{\ell=1}^{\infty} A_\ell = \bigcup_{\ell=1}^{\infty}\bigcup_{i\in I_\ell} C_i = \bigcup_{i\in J_1\cup J_2\cup\ldots} C_i,$$

    and thus $\cup A_\ell \in \mathscr{E}$.

    Second, note that $\mathscr{C} \subseteq \mathscr{D}$. For the last step, assume that $\mathscr{A}$ is a $\sigma$-algebra containing $\mathscr{C}$. Then, the property of a $\sigma$-algebra implies

    $$\bigcup_{i\in I} A_i \in \mathscr{A}$$

    for each subset $I \subseteq \mathbb{N}$. Thus, we have $\mathscr{D} \subseteq \mathscr{A}$ and we can conclude that $\sigma(\mathscr{C}) = \mathscr{D}$.

11. (a) $\Omega \in \mathscr{A}$ since $\Omega^c = \emptyset$ is countable. If $A \in \mathscr{A}$ then $A$ or $A^c$ is countable. Thus, $A^c \in \mathscr{A}$. Let $A_1, A_2, \ldots,$ be in $\mathscr{A}$. We consider two cases:

    (i) Assume that for each $k \in \mathbb{N}$ the set $A_k$ is countable. Then it follows that $\cup A_k$ is countable since it is the countable union of countable sets. Thus, $\cup A_k \in \mathscr{A}$.

    (ii) Assume that there exists at least one $k_0 \in \mathbb{N}$ such that $A_{k_0}$ is not countable. Then $A_{k_0}^c$ is countable since $A_{k_0} \in \mathscr{A}$. Since

    $$B := \cap_{k=1}^{\infty} A_k^c \subseteq A_{k_0}^c,$$

    it follows that $B$ is countable as a subset of a countable set. Since $B = \left(\cup A_k\right)^c$ it follows that $\cup A_k \in \mathscr{A}$.

(b) Since each elementary set $\{\omega\}$ is countable it follows that $\mathscr{C} \subseteq \mathscr{A}$. Assume that $\mathscr{B}$ is a $\sigma$-algebra of $\Omega$ with $\mathscr{C} \subseteq \mathscr{B}$. Then the countable union of elementary sets $\{\omega\}$ is in $\mathscr{B}$, and thus countable sets are elements of $\mathscr{B}$. Since $\mathscr{B}$ is a $\sigma$-algebra, also the complements of countable sets are in $\mathscr{B}$, which results in $\mathscr{A} \subseteq \mathscr{B}$. Thus, we have $\mathscr{A} = \sigma(\mathscr{C})$.

(c) Since $\Omega^c = \emptyset$ is countable we have $P(\Omega) = 1$. Let $A_1, A_2, \ldots$ be pairwise disjoint sets in $\mathscr{A}$.

(i) Assume that for each $k \in \mathbb{N}$ the set $A_k$ is countable. Then it follows that $\cup A_k$ is countable and thus we have

$$0 = P(\cup A_k) \quad \text{and} \quad \sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} 0 = 0.$$

(ii) Assume that there exists at least one $k_0 \in \mathbb{N}$ such that $A_{k_0}$ is not countable but $A_{k_0}^c$ is countable. Since the sets $A_1, A_2, \ldots$ are pairwise disjoint it follows that $A_k \subseteq A_{k_0}^c$ for all $k \in \mathbb{N}$ and $k \neq k_0$. Consequently, the set $A_k$ is countable for each $k \in \mathbb{N}$ and $k \neq k_0$ and thus we obtain

$$\sum_{k=1}^{\infty} P(A_k) = P(A_{k_0}) = 1.$$

Since $\cup A_k$ is not countable (as $A_{k_0} \subseteq \cup A_k$) it follows that $1 = P(\cup A_k)$ which completes the proof.

12. (a) Since $\emptyset \in \mathfrak{N}$ and $\Omega = \Omega \cup \emptyset$ we have $\Omega \in \mathscr{A}'$. For $C \in \mathscr{A}'$ there exists $A \in \mathscr{A}$ and $F \in \mathfrak{N}$ such that $C = A \cup F$. Since $F \in \mathfrak{N}$ there exists $N \in \mathscr{A}$ with $F \subseteq N$ and $P(N) = 0$. Define $\tilde{A} := A^c \cap N^c$ which is in $\mathscr{A}$ and define $\tilde{F} := F^c \cap A^c \cap N$ which is in $\mathfrak{N}$ as $\tilde{F} \subseteq N$. By using standard set-theoretic manipulation it follows that

$$\begin{aligned} C^c = A^c \cap F^c &= (A^c \cap F^c \cap N^c) \cup (A^c \cap F^c \cap N) \\ &= \big(A^c \cap N^c\big) \cup \big(A^c \cap F^c \cap N\big) = \tilde{A} \cup \tilde{F}, \end{aligned}$$

which shows that $C^c \in \mathscr{A}'$. For $C_k \in \mathscr{A}'$ for $k \in \mathbb{N}$ there exists $A_k \in \mathscr{A}$, $F_k \in \mathfrak{N}$ and $N_k \in \mathscr{A}$ with $F_k \subseteq N_k$ and $P(N_k) = 0$ such that $C_k = A_k \cup F_k$. It follows that

$$\bigcup_{k=1}^{\infty} C_k = \bigcup_{k=1}^{\infty} (A_k \cup F_k) = \left(\bigcup_{k=1}^{\infty} A_k\right) \bigcup \left(\bigcup_{k=1}^{\infty} F_k\right) =: A \cup F.$$

Clearly, $A \in \mathscr{A}$ since $\mathscr{A}$ is a $\sigma$-algebra. The set $F$ is in $\mathfrak{N}$ since part (e) of Theorem 1.2.13 implies

$$F \subseteq \bigcup_{k=1}^{\infty} N_k \quad \text{and} \quad P\left(\bigcup_{k=1}^{\infty} N_k\right) \leqslant \sum_{k=1}^{\infty} P(N_k) = 0.$$

(b) Since $\Omega = \Omega \cup \emptyset$ it follows that $P'(\Omega) = P(\Omega) = 1$. Let $C_1, C_2 \in \mathscr{A}$ be pairwise disjoint sets. Then there exists pairwise disjoint sets $A_1, A_2, \ldots \in \mathscr{A}$ and $F_1, F_2, \ldots \in \mathfrak{N}$ with $C_k = A_k \cup F_k$. As in (a) it follows that $\cup F_k \in \mathfrak{N}$, which yields

$$P'\left(\bigcup_{k=1}^{\infty} C_k\right) = P'\left(\left(\bigcup_{k=1}^{\infty} A_k\right) \bigcup \left(\bigcup_{k=1}^{\infty} F_k\right)\right) = P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} P'(C_k).$$

It is missing in this exercise to show that actually the probability measure $P'$ is well defined.

## 10.2. Chapter 2

1. (a) By using properties of probability measures, see Theorem 1.2.13, we conclude

$$P\big((a,b]\big) = P\big((-\infty,b]\backslash(-\infty,a]\big) = P\big((-\infty,b]\big) - P\big((-\infty,a]\big) = F(b) - F(a).$$

(b) By Theorem 1.2.13 and Part (a) it follows

$$\begin{aligned} P\big(\{b\}\big) = P\left(\bigcap_{n=1}^{\infty} (b - \tfrac{1}{n}, b]\right) &= \lim_{n\to\infty} P\big((b-\tfrac{1}{n}, b]\big) = F(b) - \lim_{n\to\infty} F(b - \tfrac{1}{n}) \\ &= F(b) - F(b-). \end{aligned}$$

(c) From Part (b) we derive

$$\begin{aligned} P\big([a,b]\big) = P\big((-\infty,b]\backslash(-\infty,a)\big) &= P\big((-\infty,b]\big) - P\big((-\infty,a)\big) \\ &= P\big((-\infty,b]\big) - P\big((-\infty,a]\big) + P\big(\{a\}\big) \\ &= F(b) - F(a) + F(a) - F(a-) = F(b) - F(a-). \end{aligned}$$

2. (a) For every $k \in \mathbb{N}$ we have

$$\frac{P(\{k\})}{P(\{k-1\})} = \frac{\alpha}{k}.$$

Consequently, the largest value is attained for $k = [\alpha]$, where $[x]$ denotes the largest integer smaller than $x$.

(b) Define for fixed $k \in \mathbb{N}_0$ the function

$$f\colon [0,\infty) \to \mathbb{R}, \qquad f(\alpha) = e^{-\alpha}\frac{\alpha^k}{k!}.$$

Differentiating yields that the largest value is attained for $\alpha = k$.

(c) For every $k \in \{1, \ldots, n\}$ we obtain

$$\frac{P(\{k\})}{P(\{k-1\})} = \frac{n-k+1}{k}\,\frac{p}{1-p}.$$

Since

$$(n-k+1)p \leqslant k(1-p) \iff (n+1)p \leqslant k,$$

it follows that the largest values is attained for $k = [(n+1)p]$.

3. For every $n \in \mathbb{N}$ define $\alpha_n := np_n$. Then we obtain for $k \in \{0, 1, \ldots, n\}$ that

$$\begin{aligned} P(\{k\}) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{1}{k!} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k-1}{n} \alpha_n^k \left(1 - \frac{\alpha_n}{n}\right)^n \left(1 - \frac{\alpha_n}{n}\right)^{-k}. \end{aligned} \tag{10.2.1}$$

Since $k$ is fixed and $p_n \to 0$ it follows for all $j \in \{0, \ldots, k+1\}$ that

$$\frac{n-j}{n} \to 1 \quad \text{and} \quad \left(1 - \frac{\alpha_n}{n}\right)^{-k} \to 1 \quad \text{for } n \to \infty.$$

A well known result from calculus shows that

$$\left(1 - \frac{\alpha_n}{n}\right)^n \to e^{-\alpha} \quad \text{for } n \to \infty.$$

By applying the two limits above to (10.2.1) and recalling that $\alpha_n \to \alpha$ one obtains

$$P(\{k\}) \to \frac{1}{k!} \alpha^k e^{-\alpha} \quad \text{for } n \to \infty,$$

which completes the exercise.

4. By using the formula for the geometric sum, we obtain for each $x \geqslant 0$ that,

$$P((-\infty, x]) = P(\{0, 1, \ldots, [x]\}) = \sum_{k=0}^{[x]} \alpha^k (1-\alpha) = (1-\alpha) \frac{1 - \alpha^{[x]+1}}{1-\alpha} = 1 - \alpha^{[x]+1}.$$

As the support of $P$ is $\mathbb{N} \cup \{0\}$ this shows the claimed form of the cumulative distribution function.

5. (a) Since we have

$$\int_{-\infty}^{\infty} f(u)\, du = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+u^2}\, du = \frac{1}{\pi} \left(\frac{1}{2}\pi + \frac{1}{2}\pi\right) = 1,$$

Theorem 2.3.3 implies that there exists an absolutely continuous probability measure $P$ with density $f$.

(b) For each $x \in \mathbb{R}$ we obtain

$$F_P(x) = \int_{-\infty}^{x} f(u)\, du = \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+u^2}\, du = \frac{1}{\pi} \left(\arctan x + \frac{1}{2}\pi\right).$$

6. We obtain

$$P(\mathbb{R}) = cP_1(\mathbb{R}) + (1-c)P_2(\mathbb{R}) = c + (1-c) = 1.$$

Let $B_1, B_2, \ldots \in \mathfrak{B}(\mathbb{R})$ be pairwise disjoint sets. Using $\sigma$-additivity of $P_1$ and $P_2$ we obtain

$$\begin{aligned} P\left(\bigcup_{k=1}^{\infty} B_k\right) &= cP_1\left(\bigcup_{k=1}^{\infty} B_k\right) + (1-c)P_2\left(\bigcup_{k=1}^{\infty} B_k\right) \\ &= c\sum_{k=1}^{\infty} P_1(A_k) + (1-c)\sum_{k=1}^{\infty} P_2(A_k) \\ &= \sum_{k=1}^{\infty} \big(cP_1(A_k) + (1-c)P_2(A_k)\big) = \sum_{k=1}^{\infty} P(A_k), \end{aligned}$$

which shows that $P$ is a probability measure on $\mathfrak{B}(\mathbb{R})$.

## 10.3. Chapter 3

1. Fix $\omega_0 \in \Omega$ and define $a := X(\omega_0)$. Since measurability requires $X^{-1}(\{a\}) \in \{\emptyset, \Omega\}$, but $\omega_0 \in X^{-1}(\{a\})$ and thus $X^{-1}(\{a\}) \neq \emptyset$, we can only have $X^{-1}(\{a\}) = \Omega$. Consequently, $X(\omega) = a$ for all $\omega \in \Omega$.

2. (a) For every $B \in \mathfrak{B}(\mathbb{R})$ we have

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B, \text{ and } 2 \notin B \\ \{1,3,5\}, & \text{if } 0 \notin B, \text{ and } 2 \in B \\ \{7,9\}, & \text{if } 0 \in B, \text{ and } 2 \notin B \\ \Omega, & \text{if } 0 \in B, \text{ and } 2 \in B. \end{cases}$$

Consequently, $X^{-1}(B) \in \mathscr{A}$ for all $B \in \mathfrak{B}(\mathbb{R})$, and thus $X$ is $\mathscr{A}$-measurable.

(b) $\sigma(X) = \{\emptyset, \Omega, \{1,3,5\}, \{7,9\}\}$.

3. (a) By considering $\subseteq$ and $\supseteq$ one shows

$$\{X > Y\} = \bigcup_{q \in \mathbb{Q}} \{X > q\} \cap \{Y < q\}. \tag{10.3.2}$$

Since $\{X > q\} = X^{-1}((q, \infty))$ and $\{Y < q\} = Y^{-1}((-\infty, q))$ are in $\mathscr{A}$ and one takes the union of countable sets in (10.3.2) it follows that the union is in $\mathscr{A}$, and thus we have $\{X > Y\} \in \mathscr{A}$.

(b) Follows from (a) since $\{X \neq Y\} = \{X < Y\} \cup \{X > Y\} \in \mathscr{A}$.

4. For a set $A \in \sigma(Y)$ there exists a set $B \in \mathfrak{B}(\mathbb{R})$ such that $A = Y^{-1}(B)$. It follows from Lemma C.0.2 that
$$A = Y^{-1}(B) = \big(g(X)\big)^{-1}(B) = X^{-1}\big(g^{-1}(B)\big).$$
Since measurability of $g$ guaranties that $g^{-1}(B) \in \mathfrak{B}(\mathbb{R})$ it follows that $A \in \sigma(X)$.

5. Exercise 2.5.1 implies for every $y \in \mathbb{R}$ that
$$F_Y(y) = P(|X| \leqslant y) = P(-y \leqslant X \leqslant y) = P_X([-y, y]) = F_X(y) - F_X(y-),$$
where $P_X$ denotes the probability distribution of $X$.

6. Let $F_Y$ denote the cumulative distribution function of $Y$. For $y < 0$ we have $F_Y(y) = P(\max\{X, 0\} \leqslant y) = 0$. For $y \geqslant 0$ we obtain
$$\begin{aligned} F_Y(y) &= P\big(\max\{X,0\} \leqslant y\big) \\ &= P\big(\{\max\{X,0\} \leqslant y\} \cap \{X < 0\}\big) + P\big(\{\max\{X,0\} \leqslant y\} \cap \{X \geqslant 0\}\big) \\ &= P\big(X < 0\big) + P\big(0 \leqslant X \leqslant y\big) = \int_{-1}^{0} \tfrac12\, du + \int_0^y \tfrac12\, du = \frac{1+y}{2}. \end{aligned}$$

7. For every $k \in \{1, \dots, n\}$ and $\alpha \in \mathbb{R}$ the property (i) implies
$$\{\omega \in A_k : \; X(\omega) \leqslant \alpha\} = \{\omega \in \Omega : \; X(\omega) \leqslant \alpha\} \cap A_k \in \mathscr{A} \cap A_k.$$
Theorem 3.1.7 yields part (ii).

For the reverse implication note that
$$\{\omega \in \Omega : \; X(\omega) \leqslant \alpha\} = \bigcup_{k=1}^{n} \{\omega \in A_k : \; X(\omega) \leqslant \alpha\}. \tag{10.3.3}$$
Condition (ii) implies that $\{\omega \in A_k : X(\omega) \leqslant \alpha\} \in \mathscr{A} \cap A_k$ for all $k = 1, \dots, n$. Since $A_k \in \mathscr{A}$ the very definition of $\mathscr{A} \cap A_k$ implies that $\mathscr{A} \cap A_k \subseteq \mathscr{A}$ for all $k = 1, \dots, n$. Consequently, we have $\{\omega \in A_k : X(\omega) \leqslant \alpha\} \in \mathscr{A}$ for all $k = 1, \dots, n$, which yields $\{\omega \in \Omega : X(\omega) \leqslant \alpha\} \in \mathscr{A}$ for all $\alpha \in \mathbb{R}$ by (10.3.3) Again, an application of Theorem 3.1.7 yields part (i).

8. For a set $A \subseteq \mathbb{R}$ which is not in $\mathfrak{B}(\mathbb{R})$ define
$$X\colon \mathbb{R} \to \mathbb{R}, \qquad X(x) = \begin{cases} 1, & \text{if } x \in A, \\ -1, & \text{if } x \in A^c. \end{cases}$$
It follows that $X$ is not measurable due to Example 3.1.2. However, since $X^2(x) = 1$ for all $x \in \mathbb{R}$ the function $X^2$ is measurable.

9. (a) $\sigma(\mathscr{D}) = \{\emptyset, \Omega, [0, \frac{1}{4}), [\frac{1}{4}, \frac{1}{2}), [\frac{1}{2}, 1), [0, \frac{1}{2}), [0, \frac{1}{4}) \cup [\frac{1}{2}, 1), [\frac{1}{4}, 1)\}$

(b) For each $B \in \mathfrak{B}(\mathbb{R})$ we have

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } 2 \notin B,\ 3 \notin B,\ -1.1 \notin B, \\ \Omega, & \text{if } 2 \in B,\ 3 \in B,\ -1.1 \in B, \\ [0, \frac{1}{4}), & \text{if } 2 \in B,\ 3 \notin B,\ -1.1 \notin B, \\ [\frac{1}{4}, \frac{1}{2}), & \text{if } 2 \notin B,\ 3 \in B,\ -1.1 \notin B, \\ [\frac{1}{2}, 1), & \text{if } 2 \notin B,\ 3 \notin B,\ -1.1 \in B, \\ [0, \frac{1}{2}), & \text{if } 2 \in B,\ 3 \in B,\ -1.1 \notin B, \\ [0, \frac{1}{4}) \cup [\frac{1}{2}, 1), & \text{if } 2 \in B,\ 3 \notin B,\ -1.1 \in B, \\ [\frac{1}{4}, 1), & \text{if } 2 \notin B,\ 3 \in B,\ -1.1 \in B. \end{cases}$$

Thus, we have $X^{-1}(B) \in \sigma(\mathscr{D})$ for all $B \in \mathfrak{B}(\mathbb{R})$, which shows that $X$ is $\sigma(\mathscr{D})$–measurable.

(c) The function $Y$ is not $\sigma(\mathscr{D})$–measurable since $[\frac{1}{4}, \frac{9}{16}) \in \mathfrak{B}(\mathbb{R})$ but

$$Y^{-1}\Big([\tfrac{1}{4}, \tfrac{9}{16})\Big) = [\tfrac{1}{2}, \tfrac{3}{4}) \notin \sigma(\mathscr{D}).$$

(d) The $\sigma$-algebra $\mathscr{F}$ is given by $\mathscr{F} = \{B \cap [0,1) : B \in \mathfrak{B}(\mathbb{R})\}$.

For the proof define $\mathscr{G} := \{B \cap [0,1) : B \in \mathfrak{B}(\mathbb{R})\}$.

(i) Let $G \in \mathscr{G}$. Then $G = B \cap [0,1)$ for some $B \in \mathfrak{B}(\mathbb{R})$. It follows that $G = Y^{-1}(B^2 \cap [0,1))$ where we use the notation $B^2 := \{x^2 : x \in B\}$. Since $B^2 \in \mathfrak{B}(\mathbb{R})$ as the pre-image $r^{-1}(B)$ of the continuous function $r\colon [0,1] \to [0,1]$, $r(x) = \sqrt{x}$ , we established $\mathscr{G} \subseteq \mathscr{F}$.

(ii) Each intervals $[a,b)$ for $0 \leqslant a < b \leqslant 1$ is in $\mathscr{F}$, as $[a,b) = Y^{-1}\big([a^2, b^2)\big)$. Thus, we have

$$\mathscr{F} \subseteq \sigma\Big(\big\{[a,b) : 0 \leqslant a < b \leqslant 1\big\}\Big) = \sigma\Big(\big\{[a,b) \cap [0,1) : a,b \in \mathbb{R}\,\big\}\Big) = \mathscr{G},$$

which shows our claim.

10. Assume that $Y$ is measurable. For every set $B \in \mathfrak{B}(\mathbb{R})$ we have

$$\begin{aligned} X^{-1}(B) &= \Big(X^{-1}(B) \cap Y^{-1}(B)^c\Big) \cup \Big(X^{-1}(B) \cap Y^{-1}(B)\Big) \\ &= \Big(X^{-1}(B) \cap Y^{-1}(B)^c\Big) \cup \Big(Y^{-1}(B) \cap \big(Y^{-1}(B)^c \cup X^{-1}(B)\big)\Big) \\ &= \Big(X^{-1}(B) \cap Y^{-1}(B)^c\Big) \cup \Big(Y^{-1}(B) \cap \big(Y^{-1}(B) \cap X^{-1}(B)^c\big)^c\Big) \end{aligned} \tag{10.3.4}$$

Since the sets

$$X^{-1}(B) \cap Y^{-1}(B)^c \quad \text{and} \quad Y^{-1}(B) \cap X^{-1}(B)^c$$

are both contained in the $P$-null set $\{X \neq Y\}$, they are elements of $N'$ and thus in $\mathscr{A}'$. Thus, if $Y$ is measurable, then also $Y^{-1}(B) \in \mathscr{A}'$, which yields measurability of $X$ by (10.3.4).

## 10.4. Chapter 4

1. (a) Using properties from the $P$-integral in Theorem 4.1.6 we obtain:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= E[(X-E[X])(Y-E[Y])]\\ &= E[XY-E[X]Y-E[Y]X+E[X]E[Y]]\\ &= E\big[XY\big]-E\big[E[X]Y\big]-E\big[E[Y]X\big]+E\big[E[X]E[Y]\big]\\ &= E[XY]-E[X]E[Y]-E[X]E[Y]+E[X]E[Y]\\ &= E[XY]-E[X]E[Y].\end{aligned}$$

   (b) By part (a), Theorem 4.2.10 and Theorem 4.1.6 one obtains

$$\begin{aligned}\operatorname{Var}[X+Y] &= E[(X+Y-E[X+Y])^2]\\ &= E[(X+Y)^2]-(E[X+Y])^2\\ &= E[X^2]+E[Y^2]+2E[XY]-(E[X])^2-(E[Y])^2-2E[X]E[Y]\\ &= \operatorname{Var}[X]+\operatorname{Var}[Y]+2\operatorname{Cov}(X,Y).\end{aligned}$$

2. Follows immediately from Theorem 4.1.6 and Theorem 4.2.10.

3. Since the random variable $X$ is defined on $\Omega$ we have

$$\Omega = A_1\cup\cdots\cup A_n = B_1\cup\cdots\cup B_m.$$

   Thus, we obtain the representations

$$A_i=\bigcup_{k=1}^{m}(A_i\cap B_k)\quad\text{and}\quad B_j=\bigcup_{k=1}^{n}(A_k\cap B_j)$$

   for all $i=1,\dots,n$ and $j=1,\dots,m$. Since the sets $A_i\cap B_1,\dots,A_i\cap B_m$ and the sets $A_1\cap B_j,\dots,A_n\cap B_j$ are pairwise disjoint, respectively, the additivity of the measures implies

$$\mu(A_i)=\sum_{j=1}^{m}\mu\big(A_i\cap B_j\big)\quad\text{and}\quad \mu(B_j)=\sum_{i=1}^{n}\mu\big(A_i\cap B_j\big).$$

   Consequently, we obtain

$$\begin{aligned}\sum_{i=1}^{n}\alpha_i\mu(A_i) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\mu(A_i\cap B_j),\\ \sum_{j=1}^{m}\beta_j\mu(B_j) &= \sum_{j=1}^{m}\sum_{i=1}^{n}\beta_i\mu(A_i\cap B_j)=\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_i\mu(A_i\cap B_j).\end{aligned}$$

   Since both simple random variable represents the same function we must have $\alpha_i=\beta_j$ for all $i$ and $j$ with $A_i\cap B_j\neq\emptyset$, which shows the claim.

4. (a) Let the simple function $Z$ be of the form

$$Z(\omega) = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}(\omega) \qquad \text{for all } \omega \in \Omega,$$

where $\alpha_k \in \mathbb{R}_+$ and $A_k \in \mathscr{A}$. For a constant $c \in (0,1)$ define for each $n \in \mathbb{N}$ the set $B_n := \{X_n \geqslant cZ\}$ which is in $\mathscr{A}$ due to the measurability of $X_n$ and $Z$. Since $X_n(\cdot) \geqslant cZ(\cdot)\mathbb{1}_{B_n}(\cdot)$ it follows from monotonicity in Theorem 4.1.6 that

$$\int X_n \, dP \geqslant c \int Z \mathbb{1}_{B_n} \, dP.$$

On the other hand since $X_n$ is increasing and $Z \leqslant \sup X_n$ it follows that $B_1 \subseteq B_2 \subseteq \dots$ and $\Omega = \cup_{n \geqslant 1} B_n$. Thus, we have $A_k \cap B_1 \subseteq A_k \cap B_2 \subseteq \dots$ and $A_k = \cup_{n \geqslant 1} A_k \cap B_n$. Part (f) of Theorem 1.2.13 implies

$$\begin{aligned}
\int Z \, dP = \sum_{k=1}^{m} \alpha_k P(A_k) &= \sum_{k=1}^{m} \alpha_k P\left( \bigcup_{n \geqslant 1} A_k \cap B_n \right) \\
&= \lim_{n \to \infty} \sum_{k=1}^{m} \alpha_k P(A_k \cap B_n) = \lim_{n \to \infty} \int Z \mathbb{1}_{B_n} \, dP.
\end{aligned}$$

Consequently, we have

$$\sup_{n \in \mathbb{N}} \int X_n \, dP \geqslant \sup_{n \in \mathbb{N}} c \int Z \mathbb{1}_{B_n} \, dP = c \lim_{n \to \infty} \int Z \mathbb{1}_{B_n} \, dP = c \int Z \, dP.$$

Since $c \in (0,1)$ is arbitrary, the proof is completed.

(b) For every $m \in \mathbb{N}$ we have $Y_m \leqslant \sup X_n$ and $X_m \leqslant \sup Y_n$. Part (a) implies for every $m \in \mathbb{N}$ that

$$\int Y_m \, dP \leqslant \sup_{n \in \mathbb{N}} \int X_n \, dP \quad \text{and} \quad \int X_m \, dP \leqslant \sup_{n \in \mathbb{N}} \int Y_n \, dP,$$

which completes the proof.

5. Let $\omega \in \Omega$ and $n \in \mathbb{N}$. If $X(\omega) \geqslant n$ then

$$X_n(\omega) = n = \frac{n2^{n+1}}{2^{n+1}} \leqslant X_{n+1}(\omega).$$

If $X(\omega) < n$ then there exists $k \in \{0, \dots, n2^n - 1\}$ such that

$$\frac{k}{2^n} \leqslant X(\omega) < \frac{k+1}{2^n}.$$

It follows that

$$X_n(\omega) = \frac{k}{2^n} = \frac{2k}{2^{n+1}} \leqslant X_{n+1}(\omega),$$

since $X_{n+1} \geqslant \frac{2k}{2^{n+1}}$. Thus, the monotonicity is shown.

(b) Let $\omega \in \Omega$. Since $X(\omega) < \infty$ there exists an integer $n_0 \in \mathbb{N}$ such that $X(\omega) < n_0$. For every $n \geqslant n_0$ the definition of $X_n$ yields

$$0 \leqslant X(\omega) - X_n(\omega) \leqslant \frac{1}{2^n}.$$

Consequently, it follows $X_n(\omega) \to X(\omega)$ for $n \to \infty$.

6. a) From the assumption it follows that $Q(\Omega) = E[X] = 1$. Let $A_1, A_2, \ldots \in \mathscr{A}$ be pairwise disjoint sets. Exercise 4.3.8 yields

$$Q\left(\bigcup A_k\right) = E\left[\mathbb{1}_{\cup A_k} X\right] = E\left[\sum \mathbb{1}_{A_k} X\right] = \sum E[\mathbb{1}_{A_k} X] = \sum Q(A_k).$$

b) Since $P(X > 0) = 1$ it follows that

$$P(\mathbb{1}_X X = 0,\, X = 0) \leqslant P(X = 0) = 0.$$

Consequently, we obtain

$$\begin{aligned} P(\mathbb{1}_A X = 0) &= P(\mathbb{1}_A X = 0,\, X = 0) + P(\mathbb{1}_A X = 0,\, X > 0) \\ &= P(\mathbb{1}_A X = 0,\, X > 0) = P(\mathbb{1}_A = 0) = P(A^c). \end{aligned} \tag{10.4.5}$$

Hence, if $P(A) = 0$ then $P(\mathbb{1}_A X = 0) = 1$. Theorem 4.1.6.d implies $E[\mathbb{1}_A X] = 0$. If $Q(A) = 0$ then Theorem 4.1.6.d yields $P(\mathbb{1}_A X = 0) = 1$. It follows from (10.4.5) that $P(A) = 0$.

c) First assume that $Y$ is a simple functions of the form

$$Y(\omega) = \sum_{k=1}^{m} \alpha_k \mathbb{1}_{A_k}(\omega) \qquad \text{for all } \omega \in \Omega,$$

for $\alpha_k \in \mathbb{R}$ and disjoint sets $A_1, \ldots, A_m \in \mathscr{A}$. It follows from linearity of expectation that

$$E_Q[Y] = \sum_{k=1}^{m} \alpha_k Q(A_k) = \sum_{k=1}^{m} \alpha_k E[\mathbb{1}_{A_k} X] = E[XY].$$

For a non-negative random variable $Y$ let $Y_n$ be an approximating sequence. Then

$$E_Q[Y] = \lim_{n\to\infty} E_Q[Y_n] = \lim_{n\to\infty} E[Y_n X] = E[YX],$$

where the last equality is guaranteed by the monotone convergence Theorem 4.1.9. This shows that $Y$ is $Q$-integrable if and only if $YX$ is $P$-integrable. The second part follows by the decomposition $Y = Y^+ - Y^-$ and the same calculation.

d) Follows from part (c).

e) By the definition of $R$ it follows that $R(A) = E_Q[\mathbb{1}_A \frac{1}{X}]$. Thus, part (a) implies that $R$ is a probability measure, since $E_Q[\frac{1}{X}] = E[\frac{1}{X}X] = 1$ by part (c). Moreover,

$$R(A) = E_Q[\mathbb{1}_A \frac{1}{X}] = E[\mathbb{1}_A \frac{1}{X} X] = E[\mathbb{1}_A] = P(A) \qquad \text{for all } A \in \mathscr{A},$$

implies $R = P$.

f) No. Counterexample: $\Omega = [0,1]$ and $P$ the uniform distribution on $\Omega$ and

$$X : \Omega \to \mathbb{R}, \qquad X(\omega) = \begin{cases} 2, & \text{if } \omega \in [0, \frac{1}{2}], \\ 0, & \text{else.} \end{cases}$$

Then $Q((\frac{1}{2}, 1]) = 0$ but $P((\frac{1}{2}, 1]) = \frac{1}{2}$.

7. For every $\varepsilon > 0$ it follows that $\mathbb{1}_{\{X=\infty\}} \leqslant \varepsilon X$. Thus, monotonicity of the expectation implies

$$P(X = \infty) = E[\mathbb{1}_{\{X=\infty\}}] \leqslant \varepsilon E[X].$$

Since $E[X] < \infty$ and $\varepsilon > 0$ is arbitrary, we can conclude that $P(X = \infty) = 0$.

8. (a) Define $S_n := X_1 + \cdots + X_n$ for all $n \in \mathbb{N}$. Linearity of the $P$-integral implies

$$E[S_n] = E\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} E\left[X_k\right].$$

Since $X_k$ is non-negative for all $k \in \mathbb{N}$, the sum $S_n$ is increasing and converging pointwise to $S := \sum_{k=1}^{\infty} X_k$, where the latter might equal $+\infty$. The monotone convergence Theorem 4.1.9 implies

$$E\big[S\big] = E\big[\lim_{n\to\infty} S_n\big] = \lim_{n\to\infty} E\big[S_n\big] = \sum_{k=1}^{\infty} E\left[X_k\right], \tag{10.4.6}$$

which completes part (a).

(b) By applying (a) to $|X_k|$ it follows from (10.4.6) that $U := \sum_{k=1}^{\infty} |X_k|$ obeys $E[U] < \infty$. Exercise 4.3.7 implies $P(U < \infty) = 1$, which shows (i). Thus, the partial sum $U_n := \sum_{k=1}^{n} |X_k|$ converge a.s. to $U$. Since

$$\left|\sum_{k=1}^{n} X_k\right| \leqslant \sum_{k=1}^{n} |X_k| \leqslant U \quad \text{and} \quad E[U] < \infty,$$

Lebesgue's Theorem 4.1.10 of dominated convergence implies

$$E\left[\sum_{k=1}^{\infty} X_k\right] = E\left[\lim_{n\to\infty} \sum_{k=1}^{n} X_k\right] = \lim_{n\to\infty} E\left[\sum_{k=1}^{n} X_k\right] = \lim_{n\to\infty} \sum_{k=1}^{n} E\big[X_k\big] = \sum_{k=1}^{\infty} E[X_k],$$

which completes the exercise.

9. Note that $|X| \mathbb{1}_{\{X>c\}} \leqslant |X|$ and $E[|X|] < \infty$ by assumption. Since

$$\lim_{c\to\infty} X(\omega)\mathbb{1}_{\{X>c\}}(\omega) = 0 \qquad \text{for all } \omega \in \Omega,$$

it follows from Lebesgue's Theorem 4.1.10 of dominated convergence that

$$\lim_{c\to\infty} E[|X| \mathbb{1}_{\{X>c\}}] = 0. \tag{10.4.7}$$

Since we have for each $k \in \mathbb{N}$ and $c > 0$ that

$$E[X\mathbb{1}_{A_k}] = E[X\mathbb{1}_{A_k}\mathbb{1}_{\{X>c\}}] + E[X\mathbb{1}_{A_k}\mathbb{1}_{\{X\leqslant c\}}] \leqslant E[|X| \mathbb{1}_{\{X>c\}}] + cP(A_k),$$

the exercise is completed by (10.4.7).

10. Assume first that $Z$ is a simple function, that is

$$Z = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}, \tag{10.4.8}$$

for some $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, $A_1, \dots, A_n \in \sigma(Y)$ and $n \in \mathbb{N}$. For each $k \in \{1, \dots, n\}$ there exists a set $B_k \in \mathfrak{B}(\mathbb{R})$ such that $A_k = Y^{-1}(B_k)$. Define a function by

$$g\colon \mathbb{R} \to \mathbb{R}, \qquad g(x) = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{B_k}(x).$$

Obviously, $g$ is $(\mathscr{E}\text{-}\mathfrak{B}(\mathbb{R}))$-measurable and satisfies $g(Y) = Z$.

Assume that $Z$ is non-negative. Then Theorem 4.1.3 applied for the $\sigma$-algebra $\sigma(Y)$ implies that there exists simple functions $Z_n$ of the form (10.4.8) such that $\sup_{n\in\mathbb{N}} Z_n(\omega) = Z(\omega)$ for all $\omega \in \Omega$. From the first part it follows that there exists measurable, increasing functions $g_n$ satisfying $g_n(Y) = Z_n$ for all $n \in \mathbb{N}$. Define the set $D := \{\sup_{n\in\mathbb{N}} g_n < \infty\}$ and define a function by

$$g\colon E \to \mathbb{R}, \qquad g(x) = \mathbb{1}_D(x) \sup_{n\in\mathbb{N}} g_n(x).$$

Then the function $g$ is measurable by Corollary 3.1.10. Since

$$\Big\{\sup_{n\in\mathbb{N}} g_n(Y) < \infty\Big\} = \Big\{\sup_{n\in\mathbb{N}} Z_n < \infty\Big\} = \Big\{Z < \infty\Big\} = \Omega,$$

it follows that

$$g(Y) = \mathbb{1}_D \sup_{n\in\mathbb{N}} g_n(Y) = \sup_{n\in\mathbb{N}} g_n(Y) = \sup_{n\in\mathbb{N}} Z_n = Z,$$

which completes the solution.

Applying the second part to the decomposition $Z = Z^+ - Z^-$ shows the claim for an arbitrary function $Z$

## 10.5. Chapter 5

1. (a) One obtains

$$\begin{aligned}
P(X=0) &= P(X=0,\,Y=0)+P(X=0,\,Y=1)+P(X=0,\,Y=2)=\tfrac{2}{5}+c,\\
P(X=1) &= P(X=1,\,Y=0)+P(X=1,\,Y=1)+P(X=1,\,Y=2)=\tfrac{2}{5},\\
P(X=2) &= P(X=2,\,Y=0)+P(X=2,\,Y=1)+P(X=2,\,Y=2)=\tfrac{1}{5}-c,\\
P(Y=0) &= P(X=0,\,Y=0)+P(X=1,\,Y=0)+P(X=2,\,Y=0)=\tfrac{1}{5},\\
P(Y=1) &= P(X=0,\,Y=1)+P(X=1,\,Y=1)+P(X=2,\,Y=1)=\tfrac{2}{5},\\
P(Y=2) &= P(X=0,\,Y=2)+P(X=1,\,Y=2)+P(X=2,\,Y=2)=\tfrac{2}{5}.
\end{aligned}$$

(b) With part (a) one obtains

$$\begin{aligned}
E[X] &= 0P(X=0)+1P(X=1)+2P(X=2)=\tfrac{4}{5}-2c,\\
E[X^2] &= 0^2P(X=0)+1^2P(X=1)+2^2P(X=2)=\tfrac{6}{5}-4c,\\
E[Y] &= 0P(Y=0)+1P(Y=1)+2P(Y=2)=\tfrac{6}{5},\\
E[Y^2] &= 0^2P(Y=0)+1^2P(Y=1)+2^2P(Y=2)=2,\\
E[XY] &= 1\cdot 1P(X=1,Y=1)+1\cdot 2P(X=1,Y=2)+2\cdot 2P(X=2,Y=2)\\
&= \tfrac{7}{5}-3c.
\end{aligned}$$

Consequently we obtain

$$\begin{aligned}
\mathrm{Var}[X] &= E[X^2]-\big(E[X]\big)^2=-4c^2-\tfrac{4}{5}c+\tfrac{14}{25},\\
\mathrm{Var}[Y] &= E[Y^2]-\big(E[Y]\big)^2=\tfrac{14}{25},\\
\mathrm{Cov}(X,Y) &= E[XY]-E[X]E[Y]=-\tfrac{3c}{5}+\tfrac{11}{25}.
\end{aligned}$$

2. (a) One obtains

$$\begin{aligned}
\int_{\mathbb{R}^2} f(x,y)\,dx\,dy = \int_0^1\int_0^1 c(x+y)\,dx\,dy &= c\int_0^1\left(x\int_0^1 dy+\int_0^1 y\,dy\right)dx\\
&= c\int_0^1 (x+\tfrac{1}{2})\,dx = c.
\end{aligned}$$

Consequently, in order that $f$ is the density of a probability measure the constant $c$ must satisfy $c=1$.

(b) One obtains

$$f_X(x)=\int_{\mathbb{R}} f(x,y)\,dy=\begin{cases}\int_0^1 (x+y)\,dy, & \text{if } x\in[0,1],\\ 0, & \text{else,}\end{cases}=\begin{cases}x+\tfrac{1}{2}, & \text{if } x\in[0,1],\\ 0, & \text{else,}\end{cases}$$

Symmetry of $f$ yields $f_Y=f_X$.

(c) With part (b) Theorem **??** yields that

$$E[X] = \int_{\mathbb{R}} x f_X(x)\, dx = \int_0^1 x(x + \tfrac{1}{2})\, dx = \tfrac{7}{12}.$$

Since the random variables $Y$ and $X$ have the same distribution it follows that $E[Y] = \frac{7}{12}$.

$$\begin{aligned} E[XY] = \int_{\mathbb{R}} \int_{\mathbb{R}} xyf(x,y)\, dy\, dx &= \int_0^1 \int_0^1 xy(x+y)\, dy\, dx \\ &= \int_0^1 \left( x^2 \int_0^1 y\, dy + \int_0^1 y^2\, dy \right) dx \\ &= \frac{1}{3}. \end{aligned}$$

Exercise 4.3.1 implies that the covariance is given by $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = \frac{-1}{144}$.

## 10.6. Chapter 6

1. For each $n \in \mathbb{N}_0$ we obtain

$$\begin{aligned} P(X+Y=n) &= \sum_{k=0}^{n} P(X=k, Y=n-k) \\ &= \sum_{k=0}^{n} P(X=k)(Y=n-k) = (1-\alpha)^2\alpha^n(n+1). \end{aligned}$$

Consequently, for each $k \in \{0,\dots,n\}$ we have

$$\begin{aligned} P(X=k|X+Y=n) &= \frac{P(X=k, Y=n-k)}{P(X+Y=n)} \\ &= \frac{P(X=k)(Y=n-k)}{P(X+Y=n)} \\ &= \frac{(1-\alpha)^2\alpha^k\alpha^{n-k}}{(1-\alpha)^2\alpha^k(n+1)} \\ &= \frac{1}{n+1}. \end{aligned}$$

2. We model the random experiment by $\Omega = \{dn, dp, sn, sp\}$, where $d$ stands for drunk, $s$ for sober, $p$ positive test and $n$ negative test. We define the sets

$$\begin{aligned} B &:= \text{driver is above the limit} = \{dn, dp\} \\ A &:= \text{test is positive} = \{dp, sp\}. \end{aligned}$$

The given probabilities are $P(B) = 0.05$, $P(A|B) = 0.95$ and $P(A^c|B^c) = 0.9$. It follows that

$$P(B^c) = 1 - P(B) = 0.95,$$
$$P(A^c|B) = 1 - P(A|B) = 0.05, \qquad P(A|B^c) = 1 - P(A^c|B^c) = 0.1.$$

(a) The partition equation in Theorem 6.1.6 implies

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = 0.1425.$$

Thus, the test is positive with a probability of 0.1425.

(b) Bayes' theorem in Theorem 6.1.6 implies

$$P(B^c|A) = \frac{P(A|B^c)P(B^c)}{P(A)} = 0.6666. \tag{10.6.9}$$

Thus, the alcohol percentage is below the limit, although the test is positive with a probability of 0.6666.

(c) Since $P(B \cap A^c) = P(A^c|B)P(B) = 0.0025$ we obtain

$$P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)} = 0.0029.$$

Thus, the alcohol percentage is above the limit, although the test is negative with a probability of 0.0029.

3. We model the random experiment with

$$\Omega := \{1,2,3\} \times \{a,b\} = \{(x,y) : x \in \{1,2,3\},\ y \in \{a,b\}\},$$

where the number denotes the cabinet and $a$ and $b$ the drawer. We assume (without any restriction) that in cabinet 1 contains 2 golden coins and cabinet 2 contains 2 silver coins. In cabinet 3 the drawer a contains the golden and the drawer b the silver coin. According to the question, a cabinet and a drawer is chosen randomly, i.e. we can assume the uniform distribution on $\Omega$ that is

$$P\colon \mathscr{P}(\Omega) \to [0,1], \qquad P(\{\omega\}) = \frac{1}{6}.$$

(a) The event of finding a golden coin is described by $G := \{(1,a),(1,b),(3,a)\}$. Consequently,

$$P(G) = \frac{|G|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}.$$

Thus, the probability to find a golden coin is $\frac{1}{2}$.

(b) Define the set

$$C := \text{third cabinet is chosen} = \{(3,a),(3,b)\},$$

and let $S := \{(2,a),(2,b),(3,b)\}$ describe the event of finding a silver coin. It follows that

$$P(C|S) = \frac{P(C \cap S)}{P(S)} = \frac{|\{(3,b)\}|}{\frac{1}{2}} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Thus, the probability that the other drawer of the cabinet contains a golden coin, after finding a silver coin in the same cabinet, is $\frac{1}{3}$.

4. We model the random experiment with

$$\Omega := \{(1,0,0,2),(1,0,0,3),(0,1,0,3),(0,0,1,2)\},$$

where the first 3 entries denote the door with the car indicated by 1, and the last entry denotes the door, opened by the host. Without any restriction we assume that the candidate chooses the first door. We define the sets

$$\begin{aligned} A_k &:= \text{car is behind the } k\text{-th door}, \\ B_k &:= \text{host opens the } k\text{-th door}. \end{aligned}$$

Each door is equally likely to be the one with the car, thus $P(A_k) = \frac{1}{3}$ for $k = 1,2,3$. Since the candidate chooses the door 1 (by our assumption) we have $P(B_1) = 0$. In the following cases the host does not have any choice and thus we have

$$P(B_2|A_2) = 0, \quad P(B_3|A_3) = 0, \quad P(B_2|A_3) = 1, \quad P(B_3|A_2) = 1.$$

In the only case when the host has a choice, we assume that she chooses equally likely either door 2 or door 3, and thus we obtain

$$P(B_2|A_1) = P(B_3|A_1) = \frac{1}{2}.$$

The partition equation in Theorem 6.1.6 implies that

$$\begin{aligned} P(B_2) &= P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2) + P(B_2|A_3)P(A_3) = \frac{1}{2}\frac{1}{3} + 0 + 1\frac{1}{3} = \frac{1}{2}, \\ P(B_3) &= P(B_3|A_1)P(A_1) + P(B_3|A_2)P(A_2) + P(B_3|A_3)P(A_3) = \frac{1}{2}\frac{1}{3} + 1\frac{1}{3} + 0 = \frac{1}{2}. \end{aligned}$$

Bayes' theorem in Theorem 6.1.6 implies

$$\begin{aligned} P(A_1|B_2) &= \frac{P(B_2|A_1)P(A_1)}{P(B_2)} = \frac{1/2 \cdot 1/3}{1/2} = \frac{1}{3}, \\ P(A_3|B_2) &= \frac{P(B_2|A_3)P(A_3)}{P(B_2)} = \frac{1 \cdot 1/3}{1/2} = \frac{2}{3}. \end{aligned}$$

Thus, the guest increases her chances to win in this situation by switching the chosen door. The remaining situation, that the host opens door 3 can be done analogously.

5. (a) $\Rightarrow$ (b) For every $k,\, n \in \mathbb{N}_0$ it follows that

$$\begin{aligned} P(X = n+k | X \geqslant k) &= \frac{P(X = n+k,\, X \geqslant k)}{P(X \geqslant k)} \\ &= \frac{P(X = n+k)}{P(X \geqslant k)} \\ &= (1-\alpha)\alpha^{n+k} \left( \sum_{i=k}^{\infty} (1-\alpha)\alpha^i \right)^{-1} \\ &= \alpha^{n+k}\alpha^{-k} \left( \frac{1}{1-\alpha} \right)^{-1} \\ &= (1-\alpha)\alpha^n \\ &= P(X = n). \end{aligned}$$

(b) $\Rightarrow$ (c) choose $k = 1$.
(c) $\Rightarrow$ (a) define the function $f(n) := P(X = n)$ for $n \in \mathbb{N}_0$. We show

$$f(n) = (1 - f(0))^n f(0) \tag{10.6.10}$$

for all $n \in \mathbb{N}_0$ by induction. For $n = 0$ the relation (10.6.10) is satisfied. Assume that (10.6.10) is true for $n$ and we show it for $n + 1$:

$$\begin{aligned} f(n+1) = P(X = n+1) &= P(X = n+1,\, X \geqslant 1) \\ &= P(X = n+1 | X \geqslant 1) P(X \geqslant 1) \\ &= f(n)(1 - f(0)) \qquad \text{by property (c)} \\ &= (1 - f(0))^{n+1} f(0) \qquad \text{by (10.6.10) for } n. \end{aligned}$$

Consequently, we can conclude that $X$ has the geometric distribution with parameter $\alpha = 1 - f(0)$.

6. According to the exercise we have

$$\begin{aligned} P(S = n) &= \frac{\alpha^n}{n!} e^{-\alpha} \quad \text{for all } n \in \mathbb{N}_0, \\ P(Z = k | S = n) &= \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for all } k \in \{0, \ldots, n\},\, n \in \mathbb{N}_0\,. \end{aligned}$$

The partition equation in Theorem 6.1.6 implies for each $k \in \mathbb{N}_0$ that

$$
\begin{aligned}
P(Z=k) &= \sum_{n=0}^{\infty} P(Z=k|S=n)P(S=n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\alpha^n}{n!} e^{-\alpha} \\
&= e^{-\alpha} \sum_{n=k}^{\infty} \frac{1}{(n-k)!k!} p^k (1-p)^{n-k} \alpha^n \\
&= e^{-\alpha} \sum_{n=0}^{\infty} \frac{1}{n!k!} p^k (1-p)^n \alpha^{n+k} \\
&= e^{-\alpha} \frac{(p\alpha)^k}{k!} \sum_{n=0}^{\infty} \frac{1}{n!} (1-p)^n \alpha^n \\
&= e^{-\alpha} \frac{(p\alpha)^k}{k!} e^{\alpha(1-p)} \\
&= \frac{(p\alpha)^k}{k!} e^{-\alpha p}.
\end{aligned}
$$

It follows that the random variable $Z$ is distributed according to a Poisson distribution with parameter $\alpha p$.

7. (a) Theorem 6.3.6 yields

$$
\begin{aligned}
E[Y] &= E\left[\left(\sum_{k=1}^{N} X_k\right) \mathbb{1}_{\{N \geqslant 1\}}\right] + E\left[0\mathbb{1}_{\{N=0\}}\right] \\
&= \sum_{n=1}^{\infty} E\left[\left(\sum_{k=1}^{n} X_k\right) \mathbb{1}_{\{N=n\}}\right] && \text{(partition of } \{N \geqslant 1\}\text{)} \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} E\left[X_k \mathbb{1}_{\{N=n\}}\right] && \text{(linearity of expectation)} \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} E\left[X_k\right] E[\mathbb{1}_{\{N=n\}}] && \text{(independence)} \\
&= E[X_1] \sum_{n=1}^{\infty} P(N=n) \sum_{k=1}^{n} 1 && \text{(identical distribution)} \\
&= E[X_1] \sum_{n=0}^{\infty} nP(N=n) \\
&= E[X_1]E[N].
\end{aligned}
$$

(b) Similarly as in (a) we obtain by Theorem 6.3.6 that

$$\begin{aligned}
E[Y^2] &= E\left[\left(\sum_{k=1}^{N} X_k\right)^2 \mathbb{1}_{\{N\geqslant 1\}}\right] + E\big[0^2\mathbb{1}_{\{N=0\}}\big] \\
&= \sum_{n=1}^{\infty} E\left[\left(\sum_{k=1}^{n} X_k^2 + 2\sum_{1\leqslant i<j\leqslant n} X_iX_j\right)\mathbb{1}_{\{N=n\}}\right] \qquad \text{(partition of } \{N \geqslant 1\}) \\
&= \sum_{n=1}^{\infty}\sum_{k=1}^{n} E\left[X_k^2\mathbb{1}_{\{N=n\}}\right] + 2\sum_{n=1}^{\infty}\sum_{1\leqslant i<j\leqslant n} E[X_iX_j\mathbb{1}_{\{N=n\}}] \quad \text{(linearity of expectation)} \\
&= \sum_{n=1}^{\infty}\sum_{k=1}^{n} E[X_k^2]E\big[\mathbb{1}_{\{N=n\}}\big] + 2\sum_{n=1}^{\infty}\sum_{1\leqslant i<j\leqslant n} E[X_i]E[X_j]E\big[\mathbb{1}_{\{N=n\}}\big] \quad \text{(independence)} \\
&= E[X_1^2]\sum_{n=1}^{\infty}\sum_{k=1}^{n} P(N=n) + 2\big(E[X_1]\big)^2\sum_{n=1}^{\infty}\sum_{1\leqslant i<j\leqslant n} P(N=n) \quad \text{(identical distribution)} \\
&= E[X_1^2]\sum_{n=1}^{\infty} nP(N=n) + 2\big(E[X_1]\big)^2\sum_{n=1}^{\infty}\binom{n}{2}P(N=n) \\
&= E[X_1^2]E[N] + \big(E[X_1]\big)^2\sum_{n=1}^{\infty} n(n-1)P(N=n) \\
&= E[X_1^2]E[N] + (E[X_1])^2\big(E[N^2]-E[N]\big).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{Var}[Y] = E[Y^2] - \big(E[Y]\big)^2 &= E[X_1^2]E[N] + (E[X_1])^2\big(E[N^2]-E[N]\big) - \big(E[X_1])^2\big(E[N]\big)^2 \\
&= \mathrm{Var}[X_1]E[N] + \big(E[X_1]\big)^2\,\mathrm{Var}[N].
\end{aligned}$$

8. Let $(\alpha_1,\dots,\alpha_N)$ be the key strokes which are necessary to write the complete works by Shakespeare (with $N$ very large). The exercise implicitly assumes that each key being hit with positive probability. Assume that the random variable $X_k$ describes the $k$-th stroke the monkey hits. Group the random variables $X_k$ as $S_k := \big(X_{kN+1},\dots,X_{(k+1)N}\big)$ for each $k \in \mathbb{N}_0$. Since the collection $\{X_k\}_{k\in\mathbb{N}}$ is independent, the random variables $\{S_k\}_{k\in\mathbb{N}_0}$ are independent, and they satisfy

$$P\big(S_k = (\alpha_1,\dots,\alpha_N)\big) = \prod_{j=1}^{N} P(X_j=\alpha_j) =: p > 0 \qquad \text{for all } k \in \mathbb{N}_0\,.$$

Consequently, we have

$$\sum_{k=1}^{\infty} P\big(S_k = (\alpha_1,\dots,\alpha_N)\big) = p\sum_{k=1}^{\infty} 1 = \infty,$$

and thus the Borel-Cantelli Lemma 6.2.8 implies

$$P\big(S_k = (\alpha_1, \dots, \alpha_N) \text{ for infinitely many } k\big) = 1.$$

9. Assume that $P(A) = 0$. Then it follows that

$$P(A \cap A) = P(A) = 0 = 0 \cdot 0 = P(A)P(A).$$

If $P(A) = 1$ then

$$P(A \cap A) = P(A) = 1 = 1 \cdot 1 = P(A)P(A).$$

Thus, in both cases the set $A$ is independent.

10. For ever $B_1$, $B_2 \in \mathfrak{B}(\mathbb{R})$ one obtains

$$\begin{aligned} P\big(f(X) \in B_1,\, f(Y) \in B_2\big) &= P\big(X \in f^{-1}(B_1), Y \in f^{-1}(B_2)\big) \\ &= P\big(X \in f^{-1}(B_1)\big)P\big(Y \in f^{-1}(B_2)\big) \\ &= P\big(f(X) \in B_1\big)P\big(g(Y) \in B_2\big), \end{aligned}$$

which shows that $f(B_1)$ and $f(B_2)$ are independent.

11. (a) Assume that $A$ and $B$ are independent. Then we obtain

$$\begin{aligned} P(A \cap B^c) = P(A) - P(A \cap B) &= P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) = P(A)P(B^c), \end{aligned}$$

which shows that $A$ and $B^c$ are independent. If $A$ and $B^c$ are independent, the same argument with $B$ replaced by $B^c$ shows that $A$ and $B$ are independent.

(b) If $P(B) = 0$ then $P(A)P(B) = 0$ and $P(A \cap B) \leqslant P(B) = 0$. Thus $A$ and $B$ are independent. If $P(B) = 1$ then $P(B^c) = 0$. The same argument as before shows that $A$ and $B^c$ are independent, which completes the question by part (a).

(c) If $A$, $B$ and $C$ are independent then we obtain

$$\begin{aligned} P(A)P(B)P(C) = P(A \cap B)P(C) &= \big(P(A) + P(B) - P(A \cup B)\big)P(C) \\ &= P(A \cap C) + P(B \cap C) + P(A \cup B)P(C). \end{aligned}$$

Consequently, we conclude

$$\begin{aligned} P(A \cup B)P(C) &= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) \\ &= P\big((A \cap C) \cup (B \cap C)\big) = P\big((A \cup B) \cap C\big), \end{aligned}$$

which shows that $A \cup B$ and $C$ are independent.

12. coming soon

13. (a) Linearity of the expectation yields

$$E[\bar{X}] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \mu.$$

(b) For every $i \in \{1, \ldots, n\}$ we have

$$X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu) = \tfrac{n-1}{n}(X_i - \mu) - \sum_{\substack{k=1 \\ k \neq i}}^{n} \tfrac{1}{n}(X_k - \mu).$$

Since both terms are independent and $E[X_i - \bar{X}] = 0$, Corollary 6.3.8 implies

$$\begin{aligned} E\left[(X_i - \bar{X})^2\right] &= \mathrm{Var}\left[\tfrac{n-1}{n}(X_i - \mu)\right] + \mathrm{Var}\left[\sum_{\substack{k=1 \\ k \neq i}}^{n} \tfrac{1}{n}(X_k - \mu)\right] \\ &= \left(\tfrac{n-1}{n}\right)^2 \mathrm{Var}\left[X_i - \mu\right] + \tfrac{1}{n^2}\sum_{\substack{k=1 \\ k \neq i}}^{n} \mathrm{Var}\left[X_k - \mu\right] = \tfrac{n-1}{n}\sigma^2. \end{aligned}$$

Consequently, it follows

$$E[S^2] = \tfrac{1}{n-1}\sum_{k=1}^{n} E\left[(X_k - \bar{X})^2\right] = \sigma^2.$$

(c) For every $i \in \{1, \ldots, n\}$ we have

$$\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = \mathrm{Cov}(\bar{X}, X_i) - \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$$

## 10.7. Chapter 7

1. coming soon
2. coming soon
3. coming soon
4. (a) Define the function

$$h\colon \mathbb{R} \times \mathbb{R}\backslash\{0\} \to \mathbb{R} \times \mathbb{R}\backslash\{0\}, \qquad h\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xy \\ y \end{pmatrix}.$$

The function $h$ is injective with the inverse function $u$ given by

$$u\colon \mathbb{R} \times \mathbb{R}\backslash\{0\} \to \mathbb{R} \times \mathbb{R}\backslash\{0\}, \qquad u\begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} u_1(v,w) \\ u_2(v,w) \end{pmatrix} = \begin{pmatrix} \frac{v}{w} \\ w \end{pmatrix}.$$

Both functions $h$ and $u$ are continuously differentiable with

$$J(v,w) = \begin{pmatrix} \frac{1}{w} & -\frac{v}{w^2} \\ 0 & 1 \end{pmatrix}, \qquad \det J(v,w) = \tfrac{1}{w}.$$

For the open set $N := \mathbb{R} \times \mathbb{R}\backslash\{0\}$ we have $P((X,Y) \in N) = 1$ as $Y$ has a density and thus satisfies $P(Y=0)=0$. Consequently, Theorem 7.1.9 implies that the random vector $(XY, Y)$ has a density $g\colon \mathbb{R}^2 \to \mathbb{R}_+$ given by

$$\begin{aligned} g(v,w) &= \begin{cases} |\det J(v,w)|\, f\big(u_1(v,w), u_2(v,w)\big), & \text{if } (v,w) \in \mathbb{R}\times\mathbb{R}\backslash\{0\}, \\ 0, & \text{else.} \end{cases} \\ &= \begin{cases} \frac{1}{|w|} f\big(\frac{v}{w}, w\big), \text{if } (v,w) \in \mathbb{R}\times\mathbb{R}\backslash\{0\}, \\ 0, & \text{else.} \end{cases} \end{aligned}$$

(b) The random vector $(X,Y)$ has the joint density $f\colon \mathbb{R}^2 \to \mathbb{R}_+$ given by $f(x,y) = f_X(x) f_Y(y)$ according to Theorem 6.3.4. Thus, we obtain from Part (a) that

$$\begin{aligned} g(v,w) &= \begin{cases} \frac{1}{|w|} \frac{1}{\pi\sqrt{1-\frac{v^2}{w^2}}} w \exp(-\frac{w^2}{2}) \mathbb{1}_{(-1,1)}(\frac{v}{w}) \mathbb{1}_{(0,\infty)}(w), & \text{if } (v,w) \in \mathbb{R}\times\mathbb{R}\backslash\{0\}, \\ 0, & \text{else,} \end{cases} \\ &= \begin{cases} \frac{1}{\pi\sqrt{1-\frac{v^2}{w^2}}} \exp(-\frac{w^2}{2}), & \text{if } w > |v| \geqslant 0, \\ 0, & \text{else,} \end{cases} \end{aligned}$$

where we have used that

$$w > 0,\ \big|\tfrac{v}{w}\big| < 1 \ \Leftrightarrow\ w > |v| \geqslant 0.$$

(c) According to Theorem 5.1.5 the density $h\colon \mathbb{R} \to \mathbb{R}_+$ of $XY$ is given by

$$\begin{aligned} h(v) &= \int_{\mathbb{R}} g(v,w)\, dw \\ &= \tfrac{1}{\pi} \int_{|v|}^{\infty} \frac{1}{\pi\sqrt{1-\frac{v^2}{w^2}}} \exp(-\tfrac{w^2}{2})\, dw \\ &= \tfrac{1}{\pi} \int_0^{\infty} \tfrac{z}{\sqrt{z^2+v^2}} \frac{1}{\pi\sqrt{1-\frac{v^2}{z^2+v^2}}} \exp(-\tfrac{z^2+v^2}{2})\, dz \\ &= \frac{1}{\pi} \exp(-\tfrac{v^2}{2}) \int_0^{\infty} \exp(-\tfrac{z^2}{2})\, dz \\ &= \tfrac{1}{\sqrt{2\pi}} \exp(-\tfrac{z^2}{2}). \end{aligned}$$

(d) As the density of $X$ is symmetric, i.e. $f_X(x) = f_X(-x)$ for all $x \in \mathbb{R}$, we have $E[X] = 0$. As the random variables $X$ and $Y$ are independent, we have

$$E[XY] = E[X]E[Y] = 0.$$

5. (a) From the very definition of the uniform distribution we obtain that the density of $(X,Y)$ is given by

$$f\colon \mathbb{R}^2 \to \mathbb{R}, \qquad f(x,y) = \frac{1}{\pi r^2}\mathbb{1}_S(x,y),$$

where $S := \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant r^2\}$ for $r = 1$.

(b) According to Theorem 5.1.5 we obtain that the density $f_X$ of $X$ is given by

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy, dx \\ &= \mathbb{1}_{[-r,r]}(x) \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} \frac{1}{\pi r^2}\,dy = \begin{cases} \frac{2}{\pi r^2}\sqrt{r^2-x^2}\ , & \text{if } x \in [-r,r], \\ 0, & \text{else.} \end{cases} \end{aligned}$$

The density $f_Y$ of $Y$ can be obtained analogously.

(c) Corollary 4.2.6 together with Part (a) guarantees that

$$E[X] = \int_{\mathbb{R}} x f_X(dx)\,dx = \frac{2}{\pi r^2}\int_{-r}^{r} x\sqrt{r^2-x^2}\ dx = 0,$$

since the integrand is odd. Corollary 5.3.3 yields

$$E[XY] = \int_{\mathbb{R}}\int_{\mathbb{R}} xy f(x,y)\,dx\,dy = \frac{1}{\pi r^2}\int_{-r}^{r} x \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} y\,dy\,dx = 0.$$

Thus, we obtain by the result of Exercise 4.3.1 that $\operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$.

(d) Define the random variables $R = \sqrt{X^2+Y^2}$ and $\Theta = \arctan Y/X$ and the function

$$h\colon \mathbb{R}^2 \to (0,\infty)\times[0,2\pi), \qquad h(x,y) = \begin{pmatrix} \sqrt{x^2+y^2} \\ \arctan\frac{y}{x} \end{pmatrix}.$$

The function $h$ is injective with the inverse function $u$ given by

$$u\colon (0,\infty)\times[0,2\pi) \to \mathbb{R}^2, \qquad u(s,\vartheta) = \begin{pmatrix} s\cos\vartheta \\ s\sin\vartheta \end{pmatrix}.$$

Both functions $h$ and $u$ are continuously differentiable with the Jacobi matrix of $u$

$$J(s,\vartheta) = \begin{pmatrix} \cos\vartheta & -s\sin\vartheta \\ \sin\vartheta & s\cos\vartheta \end{pmatrix}.$$

For the open set $N := \mathbb{R}\times\mathbb{R}$ we obviously have $P((X,Y)\in N) = 1$. Consequently, Theorem 7.1.9 implies that the random vector $(R,\Theta)$ has a density $g\colon \mathbb{R}^2 \to \mathbb{R}_+$ given by

$$\begin{aligned} g\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad g(s,\vartheta) &= |\det[J](s,\vartheta)|\, f(s\cos\vartheta, s\sin\vartheta) \\ &= \frac{s}{\pi r^2}\mathbb{1}_\Omega(s\cos\vartheta, s\sin\vartheta). \end{aligned}$$

According to Theorem 5.1.5 the density $g_R\colon \mathbb{R} \to \mathbb{R}$ of $R$ is given for all $s \in \mathbb{R}$ by

$$\begin{aligned} g_R(s) = \int_{-\infty}^{\infty} g(s,\vartheta)\, d\vartheta &= \frac{s}{\pi r^2} \int_0^{2\pi} \mathbb{1}_\Omega(s\cos\vartheta, s\sin\vartheta)\, d\vartheta \\ &= \frac{s}{\pi r^2} \int_0^{2\pi} \mathbb{1}_{[0,r)}(s)\, d\vartheta = \frac{2s}{r^2} \mathbb{1}_{[0,r)}(s). \end{aligned}$$

Corollary 4.2.6 guarantees that

$$\begin{aligned} E[R] &= \int_{\mathbb{R}} s g_R(s)\, ds = \frac{2}{3} r, \\ \operatorname{Var}[R] &= \int_{\mathbb{R}} (s - \tfrac{2}{3}r)^2 g_R(s) = \tfrac{2}{3} \int_0^r s(s - \tfrac{2}{3}r)^2\, ds = \tfrac{1}{18} r^2. \end{aligned}$$

6. The density of $(X,Y)$ is given by

$$f\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f(x,y) = \mathbb{1}_D(x,y)\frac{1}{\pi}.$$

By changing to polar coordinates we obtain for every $z \in \mathbb{R}$ that

$$P(X/Y \leqslant z) = \int_{\substack{(x,y)\in D \\ x/y \leqslant z}} \frac{1}{\pi}\, dx\, dy = \frac{1}{2} + \frac{1}{\pi} \arctan(z).$$

Theorem 7.1.5 implies that $X/Y$ has a density given by

$$g\colon \mathbb{R} \to \mathbb{R}_+, \qquad g(z) = \frac{1}{\pi(1+z^2)}.$$

7. coming soon

8. coming soon

9. (a) Theorem 6.3.2 guarantees that the cumulative distribution function $F_Y$ of $Y$ obeys for each $y \geqslant 0$:

$$\begin{aligned} F_Y(y) = P(Y \leqslant y) &= P(X_1 \leqslant y, \dots, X_n \leqslant y) \\ &= P(X_1 \leqslant y) \cdot \ldots \cdot P(X_n \leqslant y) = \big(1 - e^{-\alpha y}\big)^n. \end{aligned}$$

Analogously, the cumulative distribution function $F_Z$ of $Z$ is for each $z \geqslant 0$ given by

$$F_Z(z) = 1 - P(Z > z) = 1 - P(X_1 > z, \dots, X_n > z) = 1 - e^{-n\alpha z}$$

Cleary, $F_Y(y) = F_Z(y) = 0$ for $y < 0$.

(b) Since $F_Y$ and $F_Z$ are continuously differentiable in $\mathbb{R}\backslash\{0\}$, Theorem 7.1.5 implies that $Y$ has a density $f_Y$ given by

$$f_Y\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_Y(y) = n\alpha(1 - e^{-\alpha y})^{n-1} e^{-\alpha y} \mathbb{1}_{[0,\infty)}(y),$$

and that $Z$ has a density $f_Z$ given by

$$f_Z\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_Z(z) = n\alpha e^{-n\alpha z}.$$

10. (a) Since $X_1$ and $X_2$ are independent, Corollary 6.3.5 implies that the joint density $f_{X_1,X_2}$ of the random variable $(X_1, X_2)\colon \Omega \to \mathbb{R}^2$ is given by

$$f_{X_1,X_2}\colon \mathbb{R}^2 \to \mathbb{R}_+, \qquad f_{X_1,X_2}(x_1,x_2) = \alpha\beta e^{-\alpha x_1} e^{-\beta x_2} \mathbb{1}_{[0,\infty)}(x_1)\mathbb{1}_{[0,\infty)}(x_2).$$

Let $D := \{(x_1,x_2) \in \mathbb{R}^2 : x_1 > 0,\, x_2 > 0\}$ and note that $D$ is open and $P\big((X_1,X_2) \in D\big) = 1$. Define the function

$$h\colon D \to D, \qquad h(x_1,x_2) = \begin{pmatrix} x_1 + x_2 \\ \frac{x_1}{x_2} \end{pmatrix}.$$

The function $h$ is invertible with inverse $u = h^{-1}$ given by

$$u\colon D \to D, \qquad u(y_1,y_2) = \begin{pmatrix} \frac{y_1 y_2}{1+y_2} \\ \frac{y_1}{1+y_2} \end{pmatrix}.$$

Both functions $h$ and $u$ are continuously differentiable and the Jacobian matrix of $u$ is given by

$$J(y_1,y_2) = \begin{pmatrix} \frac{y_2}{1+y_2} & \frac{y_1(1+y_2)-y_1y_2}{(1+y_2)^2} \\ \frac{1}{1+y_2} & \frac{-y_1}{(1+y_2)^2} \end{pmatrix}.$$

Theorem 7.1.9 implies that the random vector $(Y_1,Y_2)\colon \Omega \to \mathbb{R}^2$ has a density $g\colon \mathbb{R}^2 \to \mathbb{R}_+$, and that the density $g$ is given by

$$\begin{aligned} g(y_1,y_2) &= |\det[J(y_1,y_2)]|\, f_{X_1,X_2}\big(u_1(y_1,y_2), u_2(y_1,y_2)\big) \\ &= \alpha\beta \frac{y_1}{(1+y_2)^2} \exp\left(-\alpha\left(\tfrac{y_1y_2}{1+y_2}\right)\right) \exp\left(-\beta\left(\tfrac{y_1}{1+y_2}\right)\right) \mathbb{1}_{[0,\infty)}(y_1)\mathbb{1}_{[0,\infty)}(y_2). \end{aligned}$$

(b) If $\alpha = \beta$ it follows from part (a) that the joint density $g\colon \mathbb{R}^2 \to \mathbb{R}_+$ of the random variable $(Y_1,Y_2)\colon \Omega \to \mathbb{R}^2$ is given by

$$g(y_1,y_2) = \alpha^2 \frac{y_1}{(1+y_2)^2} \exp\big(-\alpha y_1\big) \mathbb{1}_{[0,\infty)}(y_1)\mathbb{1}_{[0,\infty)}(y_2).$$

According to Theorem 5.1.5,the random variable $Y_1$ has the density $g_{Y_1}\colon \mathbb{R} \to \mathbb{R}_+$ given by

$$g_{Y_1}(y) = \int_{\mathbb{R}} g(y,y_2)\, dy_2 \quad = \alpha^2 y e^{-\alpha y}\mathbb{1}_{[0,\infty)}(y) \int_0^\infty \tfrac{1}{(1+y_2)^2}\, dy_2 = \alpha^2 y e^{-\alpha y}\mathbb{1}_{[0,\infty)}(y),$$

and the random variable $Y_2$ has the density $g_{Y_2}\colon \mathbb{R} \to \mathbb{R}_+$ given by

$$g_{Y_2}(y) = \int_{\mathbb{R}} g(y_1,y)\, dy_1 \quad = \frac{1}{(1+y)^2}\mathbb{1}_{[0,\infty)}(y)\alpha^2 \int_0^\infty y_1 e^{-\alpha y_1}\, dy_1 = \frac{1}{(1+y)^2}\mathbb{1}_{[0,\infty)}(y).$$

It follows that

$$g(y_1,y_2) = g_{Y_1}(y_1)g_{Y_2}(y_2) \qquad \text{for all } y_1,y_2 \in \mathbb{R},$$

and thus, Theorem 6.3.4 guarantees that $Y_1$ and $Y_2$ are independent.

## 10.8. Chapter 8

1. (a) Define the sets $C_k := \{X + Y = k\}$ for $k \in \mathbb{N}_0$. By independence of $X$ and $Y$ we conclude

$$\begin{aligned}\int_{C_k} X\,dP &= E\left[X\mathbb{1}_{C_k}\right]\\ &= \sum_{i=1}^{k} iP(X = i, Y = k - i)\\ &= e^{-(\alpha+\beta)}\sum_{i=1}^{k}\frac{1}{(i-1)!(k-i)!}\alpha^i\beta^{k-i}\\ &= e^{-(\alpha+\beta)}\frac{\alpha}{(k-1)!}\sum_{i=0}^{k-1}\frac{(k-1)!}{i!(k-1-i)!}\alpha^i\beta^{k-1-i}\\ &= e^{-(\alpha+\beta)}\frac{\alpha}{(k-1)!}(\alpha+\beta)^{k-1}. \end{aligned} \tag{10.8.11}$$

Since $X + Y$ has a Poisson distribution with parameter $\alpha + \beta$ due to Example 7.2.2, we conclude from (10.8.11) that for each $k \in \mathbb{N}_0$:

$$M_{C_k} = \frac{1}{P(C_k)}\int_{C_k} X\,dP = \frac{\alpha}{\alpha+\beta}k.$$

Since Example 1.2.7.c shows that

$$\sigma(X+Y) = \sigma(\{C_k : k \in \mathbb{N}\}) = \left\{\bigcup_{i\in I} C_i : \text{ for each } I \subseteq \mathbb{N}\right\},$$

Example 8.3.3.d implies

$$E[X|X+Y] = \sum_{k=0}^{\infty} M_{C_k}\mathbb{1}_{C_k} = \frac{\alpha}{\alpha+\beta}\sum_{k=0}^{\infty} k\mathbb{1}_{\{X+Y=k\}} = \frac{\alpha}{\alpha+\beta}(X+Y),$$

which completes the solution.

(b) For $y \in \mathbb{N} \cup \{0\}$ we obtain

$$\begin{aligned} g(y) &= \sum_{k=0}^{\infty} kP(X = k|X + Y = y) \\ &= \sum_{k=1}^{\infty} k \frac{P(X = k)P(Y = y - k)}{P(X + Y = y)} \\ &= \frac{1}{(\alpha+\beta)^y} \sum_{k=1}^{y} k \binom{y}{k} \alpha^k \beta^{y-k} \\ &= \frac{\alpha}{(\alpha+\beta)^y} \frac{d}{d\alpha} (\alpha+\beta)^y \\ &= \frac{\alpha}{\alpha+\beta} y. \end{aligned}$$

Consequently, Theorem 8.5.1 implies that $E[X|X+Y] = \frac{\alpha}{\alpha+\beta}(X+Y)$ $P$-a.s.

2. (a) It follows from Example 8.3.3.d that

$$\begin{aligned} E[X|Z] &= \left( \frac{1}{P(Z=0)} \int_{\{Z=0\}} X \, dP \right) \mathbb{1}_{\{Z=0\}} + \left( \frac{1}{P(Z=1)} \int_{\{Z=1\}} X \, dP \right) \mathbb{1}_{\{Z=1\}} \\ &= \frac{E[X \mathbb{1}_{\{Z=0\}}]}{P(Z=0)} \mathbb{1}_{\{Z=0\}} + \frac{E[X \mathbb{1}_{\{Z=1\}}]}{P(Z=1)} \mathbb{1}_{\{Z=1\}}. \end{aligned}$$

If $Z = 1$ then $X = 0$ and thus, the second term above vanishes since $E[X\mathbb{1}_{\{Z=1\}}] = 0$. In order to simplify the first term, we derive for the denominator

$$P(Z = 0) = 1 - P(Z = 1) = 1 - P(X = 0, Y = 0) = 1 - (1-p)^2 = p(2-p).$$

For the numerator of the first term we obtain

$$E[X\mathbb{1}_{\{Z=0\}}] = E[X\mathbb{1}_{\{Z=0\}}\mathbb{1}_{\{X=0\}}] + E[X\mathbb{1}_{\{Z=0\}}\mathbb{1}_{\{X=1\}}] = P(Z = 0, \ X = 1) = p.$$

Consequently, we obtain

$$E[X|Z] = \frac{\mathbb{1}_{\{Z=0\}}}{2-p} \quad P\text{-a.s.}$$

(b) Coming soon

3. (a) This definition only makes sense if we identify random variables which are $P$-a.s. equal. In this case, the definition is well defined since $X \in \mathscr{L}^2(\Omega, P)$ and thus Theorem 8.3.5 guarantees that $E[X|Y] \in \mathscr{L}^2(\Omega, P)$. Thus, $(X - E[X|Y])^2 \in \mathscr{L}^1(\Omega, P)$, which shows by Theorem 8.3.6 that the conditional expectation of $(X - E[X|Y])^2$ is well defined.

(b) Theorem 8.4.1 yields $P$-a.s.

$$\begin{aligned}
\operatorname{Var}[X|Y] &= E\big[X^2 - 2XE[X|Y] + \big(E[X|Y]\big)^2|Y\big] \\
&= E\big[X^2|Y\big] - 2E\big[XE[X|Y]|Y\big] + E\big[\big(E[X|Y]\big)^2|Y\big] \qquad (\text{ linearity}) \\
&= E\big[X^2|Y\big] - 2E\big[X|Y\big]E\big[X|Y\big] + \big(E[X|Y]\big)^2 E[1|Y] \qquad (\text{known out}) \\
&= E\big[X^2|Y\big] - \big(E[X|Y]\big)^2.
\end{aligned}$$

(c) By part (a) and Theorem 4.2.10 we obtain

$$\begin{aligned}
&E\big[\operatorname{Var}[X|Y]\big] + \operatorname{Var}\big[E[X|Y]\big] \\
&\quad = E\big[E[X^2|Y]\big] - E\big[\big(E[X|Y]\big)^2\big] + E\big[\big(E[X|Y]\big)^2\big] - \big(E\big[E[X|Y]\big]\big)^2 \\
&\quad = E[X^2] - \big(E[X]\big)^2 \\
&\quad = \operatorname{Var}[X].
\end{aligned}$$

4. Corollary 6.3.5 implies that the random variable $(X, Y)$ has the density

$$f\colon \mathbb{R}^2 \to \mathbb{R}, \qquad f(x,y) = \alpha^2 e^{-\alpha(x+y)} \mathbb{1}_{[0,\infty)}(x)\mathbb{1}_{[0,\infty)}(y).$$

In order to apply Theorem 8.5.2 we have to determine the joint density of $(X, X+Y)$. Define the function

$$h\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad h(x,y) = (x, x+y).$$

The function $h$ is injective with inverse function $u = h^{-1}$

$$u\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad u(x,z) = (x, z-x).$$

Since both $h$ and $u$ are continuously differentiable, Theorem 7.1.9 implies that the random variable $(X, X+Y)$ has a density given by

$$f_{X,X+Y}\colon \mathbb{R}^2 \to \mathbb{R}, \qquad f_{X,X+Y}(x,z) = \alpha^2 e^{-\alpha z} \mathbb{1}_{[0,z]}(x)\mathbb{1}_{[0,\infty)}(z).$$

Example 7.3.3 shows that the density $f_Z$ of $Z = X+Y$ is given by

$$f_Z\colon \mathbb{R} \to \mathbb{R}_+, \qquad f_Z(z) = \alpha^2 z e^{-\alpha z} \mathbb{1}_{[0,\infty)}(z).$$

For every $z > 0$ we obtain

$$\begin{aligned}
g(z) := \frac{1}{f_Z(z)} \int_{\mathbb{R}} x f_{X,Z}(x,y)\, dx &= \int_{\mathbb{R}} x \frac{\alpha^2 e^{-\alpha z}\mathbb{1}_{[0,z]}(x)\mathbb{1}_{[0,\infty)}(z)}{\alpha^2 z e^{-\alpha z}}\, dx \\
&= \frac{1}{z}\int_0^z x\, dx \\
&= \frac{1}{2} z.
\end{aligned}$$

Thus, Theorem 8.5.2 implies that

$$E[X|X+Y] = \frac{1}{2}(X+Y) \quad P\text{-a.s}$$

5. (a) The density $f_X\colon \mathbb{R} \to \mathbb{R}_+$ of $X$ is given for all $x \in \mathbb{R}$ by

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \tfrac{4}{x^3}\mathbb{1}_{(0,1)}(x) \int_0^{x^2} y\,dy = \tfrac{8}{2x^3}\mathbb{1}_{(0,1)}(x)x^4 = 2x\mathbb{1}_{(0,1)}(x).$$

The density $f_Y\colon \mathbb{R} \to \mathbb{R}_+$ of $Y$ is given for all $y \in \mathbb{R}$ by

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx = 4y \int_{\sqrt{y}}^1 \tfrac{1}{x^3}\,dx = 2y\mathbb{1}_{(0,1)}(y)\left(\tfrac{1}{y} - 1\right) = 2(1-y)\mathbb{1}_{(0,1)}(y).$$

(b) For each $x \in (0,1)$ we calculate

$$g(x) := \tfrac{1}{f_X(x)} \int_{\mathbb{R}} y f(x,y)\,dy = \tfrac{4}{2x^4} \int_0^{x^2} y^2\,dy = \tfrac{2}{3}x^2.$$

Consuently, Theorem 8.5.2 implies

$$E[Y|X] = \tfrac{2}{3}X^3 \quad P\text{-a.s.}$$

(c) For each $y \in (0,1)$ we calculate

$$g(y) := \tfrac{1}{f_Y(x)} \int_{\mathbb{R}} x^2 f(x,y)\,dx = \tfrac{4y}{2(1-y)} \int_{\sqrt{y}}^1 \tfrac{1}{x}\,dx = -\tfrac{2y}{1-y} \ln\sqrt{y}\,.$$

Consuently, Theorem 8.5.2 implies

$$E[X^2|Y] = -\tfrac{2Y}{1-Y} \ln\sqrt{Y} \quad P\text{-a.s.}$$

(d) We will show that $\sigma(Y) = \sigma(Y^2)$, which immediately establishes $E[X^2|Y^2] = E[X^2|Y]$ $P$-a.s. For this purpose, define the function

$$h\colon [0,1] \to [0,1], \qquad h(x) = x^2 \quad \text{and} \quad u\colon [0,1] \to [0,1], \qquad u(y) = \sqrt{y}\,.$$

Clearly, $u$ is the inverse function of $h$.

Let $A$ be in $\sigma(Y^2)$. Then there exists $B \in \mathfrak{B}(\mathbb{R})$ such that $A = \big(h(Y)\big)^{-1}(B) = Y^{-1}(h^{-1}(B))$. Thus, $A \in \sigma(Y)$ as $h^{-1}(B) \in \mathfrak{B}(\mathbb{R})$.

On the other hand, let $A$ be in $\sigma(Y)$. Then there exists $B \in \mathfrak{B}(\mathbb{R})$ such that

$$A = Y^{-1}(B) = Y^{-1}(uh(B)) = \big(h(Y)\big)^{-1}(h(B)).$$

Thus, $A \in \sigma(Y^2)$ if we can show that $h(B) \in \mathfrak{B}(\mathbb{R})$.

For this purpose we define

$$\mathscr{D} := \{C \subseteq [0,1] : h(C) \in \mathfrak{B}(\mathbb{R})\}.$$

It is easy to show that $\mathscr{D}$ is a $\sigma$-algebra. Moreover, each interval $[a,b]$ with $0 \leqslant a < b \leqslant 1$ is in $\mathscr{D}$ since $h([a,b]) = [a^2, b^2] \in \mathfrak{B}(\mathbb{R})$. Since these intervals generate $\mathfrak{B}(\mathbb{R}) \cap [0,1]$ we have $\mathfrak{B}(\mathbb{R}) \cap [0,1] \subseteq \mathscr{D}$.

6. The density $f\colon \mathbb{R}^2 \to \mathbb{R}$ of $(X,Y)$ and the marginal density $f_Y\colon \mathbb{R} \to \mathbb{R}$ of $Y$ is given in Example 5.2.5. For every $y \in \mathbb{R}$ we obtain

$$\begin{aligned}
g(y) :&= \frac{1}{f_Y(y)} \int_{\mathbb{R}} x f(x,y)\,dx \\
&= \sqrt{2\pi\sigma_2^2}\ \exp\left(\frac{1}{2\sigma_2^2} y^2\right) \\
&\qquad \int_{\mathbb{R}} x \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left(-\frac{1}{2(1-\varrho^2)\sigma_1^2\sigma_2^2}\left(\sigma_2^2x^2 - 2\varrho\sigma_1\sigma_2 xy + \sigma_1^2 y^2\right)\right) dx \\
&= \frac{1}{\sqrt{2\pi(1-\varrho^2)\sigma_1^2}} \int_{\mathbb{R}} x \exp\left(-\frac{1}{2(1-\varrho^2)\sigma_1^2}\left(x - \tfrac{\sigma_1}{\sigma_2}\varrho y\right)^2\right) dx \\
&= \frac{\sigma_1}{\sigma_2}\varrho y.
\end{aligned}$$

Theorem 8.5.2 implies that

$$E[X|Y] = \frac{\sigma_1}{\sigma_2}\varrho Y \quad P\text{-a.s}$$

7. (a) Theorem 5.1.5 implies that the density $f_X\colon \mathbb{R} \to \mathbb{R}_+$ of $X$ is given by

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = xe^{-x}\mathbb{1}_{[0,\infty)}(x)\int_0^\infty e^{-xy}\,dy = e^{-x}\mathbb{1}_{[0,\infty)}(x),$$

and that the density $f_Y\colon \mathbb{R} \to \mathbb{R}_+$ of $Y$ is given by

$$f_Y(y) = \int_{\mathbb{R}} (x,y)\,dx = \mathbb{1}_{[0,\infty)}(y)\int_0^\infty xe^{-x(y+1)}\,dx = \frac{1}{(1+y)^2}\mathbb{1}_{[0,\infty)}(y).$$

(b) For every $x > 0$ we obtain

$$\begin{aligned}
g(x) &= \frac{1}{f_X(x)}\int_{\mathbb{R}} y f(x,y)\,\lambda(dy) \\
&= e^x \int_0^\infty yxe^{-x(y+1)}\,dy \\
&= x\int_0^\infty ye^{-xy}\,dy = \frac{1}{x}.
\end{aligned}$$

Consequently, Theorem 8.5.2 implies

$$E[Y|X] = \frac{1}{X} \quad P\text{-a.s.}$$

## 10.9. Chapter 9

1. (a) By the triangular inequality (9.1.2) it follows for each $n \in \mathbb{N}$ that

$$\|X\|_p \leqslant \|X - X_n\|_p + \|X_n\|_p\,.$$

By taking the limit inferior and using $X_n \to X$ in $\mathscr{L}^p(\Omega, P)$ one obtains

$$\|X\|_p \leqslant \liminf_{n\to\infty} \|X_n\|_p . \tag{10.9.12}$$

Analogously, the triangular inequality implies for each $n \in \mathbb{N}$ that

$$\|X_n\|_p \leqslant \|X_n - X\|_{p)} + \|X\|_p .$$

By taking the limit superior and using $X_n \to X$ in $\mathscr{L}^p(\Omega, P)$ one obtains

$$\limsup_{n\to\infty} \|X_n\|_{p)} \leqslant \|X\|_p . \tag{10.9.13}$$

It follows from the inequalities (10.9.12) and (10.9.13) that $E[|X_n|^p] \to E[|X|^p]$.

(b) Part (c) in Theorem 4.1.6 implies for each $n \in \mathbb{N}$ that

$$|E[X_n] - E[X]| \leqslant E[|X_n - X|] = \|X_n - X\|_p .$$

Consequently, one obtains $E[X_n] \to E[X]$.

The converse implication is incorrect as illustrated by taking random variables $X_n \colon \Omega \to \mathbb{R}$ with

$$P(X_n = 1) = P(X_n = -1) = \frac{1}{2} \qquad \text{for all } n \in \mathbb{N} .$$

It follows that $E[X_n] = 0$ but $E[|X_n - 0|] = E[|X_n|] = 1$ for all $n \in \mathbb{N}$.

(c) Part (a) implies $E[|X_n|^2] \to E[|X|^2]$. Inequality 9.1.3 yields $\|X_n - X\|_{1)} \to 0$, and thus part (b) implies $E[X_n] \to E[X]$. Both conclusions together result in

$$\mathrm{Var}[X_n] = E[X_n^2] - \big(E[X_n]\big)^2 \to E[X^2] - \big(E[X]\big)^2 = \mathrm{Var}[X],$$

which completes the solution.

2. (a) For each $n \in \mathbb{N}$ one computes

$$\int_{-c}^{c} \frac{n\,|u|}{\pi(1+n^2u^2)}\,\lambda(du) = 2\int_0^c \frac{nu}{\pi(1+n^2u^2)}\,du = \frac{1}{n}\ln(1+n^2c^2) \to \infty \quad \text{for } c \to \infty.$$

It follows that $E[|X_n|] = \infty$ and inequality 4.2.10 implies that $E[|X_n|^2] = \infty$. (In fact, using a more general inequality, one can show that $E[|X_n|^p] = \infty$ for all $p \geqslant 1$.

(b) For every $n \in \mathbb{N}$ and $\varepsilon > 0$ one obtains

$$P(|X_n| > \varepsilon) = 2\int_\varepsilon^\infty \frac{n}{\pi(1+n^2u^2)}\,\lambda(du) = \frac{2}{\pi}\Big(\frac{\pi}{2} - \arctan(\varepsilon n)\Big) \to 0 \quad \text{for } n \to \infty.$$

Consequently, $X_n \to 0$ in probability.

(c) By an application of L'Hôpital's rule it follows that

$$\lim_{n\to\infty} \frac{\frac{\pi}{2} - \arctan(\varepsilon n)}{1/n} = \frac{1}{\varepsilon}.$$

Consequently, it follows that

$$\sum_{k=1}^{\infty} P(|X_k| > \varepsilon) = \infty.$$

Since the sequence $(X_n)_{n\in\mathbb{N}}$ is independent, the Borel-Cantelli Lemma 6.2.8 implies that

$$P\left(\limsup_{n\to\infty}\{X_n > \varepsilon\}\right) = 1,$$

and thus $(X_n)_{n\in\mathbb{N}}$ does not converge to 0 $P$-a.s.

3. (a) For every $n \in \mathbb{N}$ it follows that

$$P\left(\frac{X_n}{\ln n} \geqslant \frac{1}{\alpha}\right) = P\left(X_n \geqslant \frac{1}{\alpha}\ln n\right) = e^{-\frac{\alpha}{\alpha}\ln n} = \frac{1}{n}.$$

Part (c) of the Borel-Cantelli Lemma 6.2.8 implies

$$P\left(\frac{X_n}{\ln n} \geqslant \frac{1}{\alpha} \text{ for infinitely many } n\right) = 1.$$

(b) For every $\beta < \alpha$ and $n \in \mathbb{N}$ it follows that

$$P\left(\frac{X_n}{\ln n} \geqslant \frac{1}{\beta}\right) = P\left(X_n \geqslant \frac{1}{\beta}\ln n\right) = e^{-\frac{\alpha}{\beta}\ln n} = n^{-\frac{\alpha}{\beta}}.$$

Since $\sum_{n=1}^{\infty} n^{-\frac{\alpha}{\beta}} < \infty$ for $\beta < \alpha$, part (a) of the Borel-Cantelli Lemma 6.2.8 implies

$$P\left(\frac{X_n}{\ln n} \geqslant \frac{1}{\beta} \text{ for infinitely many } n\right) = 0,$$

which yields

$$P\left(\frac{X_n}{\ln n} \leqslant \frac{1}{\beta} \text{ for eventually all } n\right) = 1,$$

By choosing a sequence $(\beta_k)_{k\in\mathbb{N}}$ with $\beta_k \nearrow \alpha$ for $k \to \infty$ one deduces that

$$\begin{aligned} P\left(\liminf_{n\to\infty} \frac{X_n}{\ln n} \leqslant \frac{1}{\alpha}\right) &= P\left(\bigcap_{k=1}^{\infty}\left\{\liminf_{n\to\infty} \frac{X_n}{\ln n} \leqslant \frac{1}{\beta_k}\right\}\right) \\ &= \lim_{k\to\infty} P\left(\liminf_{n\to\infty} \frac{X_n}{\ln n} \leqslant \frac{1}{\beta_k}\right) = 1, \end{aligned}$$

which completes the solution.

(c) Obviously, part (a) and (b) imply part (c).

4. Assume that $(X_n)_{n\in\mathbb{N}}$ converges in probability to $X$. For each $n \in \mathbb{N}$ one obtains

$$\sup_{m\geqslant n} P\left(|X_m - X_n| \geqslant \varepsilon\right) \leqslant P\left(|X - X_n| \geqslant \frac{\varepsilon}{2}\right) + \sup_{m\geqslant n} P\left(|X_m - X| \geqslant \frac{\varepsilon}{2}\right).$$

Both terms on the right hand side converge to 0 for $n \to \infty$ since $X_n \xrightarrow{P} X$.

Assume that

$$\lim_{n\to\infty} \sup_{m\geqslant n} P\left(|X_m - X_n| \geqslant \varepsilon\right) = 0 \qquad \text{for all } \varepsilon > 0.$$

It follows that there exists an increasing sequence $(n_k)_{k\in\mathbb{N}} \subseteq \mathbb{N}$ such that

$$P\left(|X_m - X_n| \geqslant 2^{-k}\right) \leqslant 2^{-k} \quad \text{for all } m, n \geqslant n_k.$$

Since $\sum_{k=1}^{\infty} 2^{-k} < \infty$, part (a) of the Borel-Cantelli Lemma 6.2.8 implies that there exists $\Omega_0 \in \mathscr{A}$ with $P(\Omega_0) = 1$ such that for all $\omega \in \Omega_0$ we have

$$\left|X_{n_{k+1}}(\omega) - X_{n_k}(\omega)\right| \geqslant 2^{-k} \text{only for finitely many } k \in \mathbb{N}.$$

Consequently, for each $\omega \in \Omega_0$ the sum $\sum_{\ell=1}^{\infty} X_{n_{\ell+1}}(\omega) - X_{n_\ell}(\omega)$ converges. It follows that

$$\lim_{k\to\infty} X_{n_k} = X_{n_1} + \lim_{k\to\infty} \sum_{\ell=1}^{k} X_{n_{\ell+1}} - X_{n_\ell} = X_{n_1} + \sum_{\ell=1}^{\infty} X_{n_{\ell+1}} - X_{n_\ell} \quad P\text{-a.s.}.$$

Thus, the subsequence $(X_{n_k})_{k\in\mathbb{N}}$ converges $P$-a.s. Since this argument can be applied to each subsequence of $(X_n)_{n\in\mathbb{N}}$, an application of Theorem 9.1.9 completes the solution.

5. (1)⇒(2): Assume for a contradiction that (2) is not true, i.e.

$$\limsup_{n\to\infty} E\big[\,|X_n - X| \wedge 1\big] > 0.$$

It follows that there exists an $\varepsilon > 0$ and a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ such that

$$E\big[\,|X_{n_k} - X| \wedge 1\big] > \varepsilon \quad \text{for all } k \in \mathbb{N}\,. \tag{10.9.14}$$

However, since $X_n \xrightarrow{P} X$, Theorem 9.1.9 implies that there exists a subsequence $(X_{n_{k_\ell}})_{\ell\in\mathbb{N}}$ such that $X_{n_{k_\ell}} \to X$ $P$-a.s. for $\ell \to \infty$. It follows that

$$Y_\ell := \left|X_{n_{k_\ell}} - X\right| \wedge 1 \to 0 \quad P\text{-a.s. for } \ell \to \infty.$$

Since $E[|Y_\ell|] \leqslant 1$ for all $\ell \in \mathbb{N}$, Lebesgue's dominated convergence Theorem 4.1.10 implies $E[Y_\ell] \to 0$ for $\ell \to \infty$, which contradicts (10.9.14).

(2)⇒(1): for each $\varepsilon > 0$ and $n \in \mathbb{N}$ one obtains

$$P(|X_n - X| \geqslant \varepsilon) = E[\mathbb{1}_{\{|X_n - X|\geqslant\varepsilon\}}] \leqslant E\left[\tfrac{1}{\varepsilon}\,|X_n - X| \wedge 1\right] \leqslant \frac{1}{\varepsilon} E\left[|X_n - X| \wedge 1\right],$$

which shows (a).

6. (a) Since for all $\varepsilon \in (0,1)$ we have $P(|X_n| \geqslant \varepsilon) = p_n$ for all $n \in \mathbb{N}$, we have:

$$X_n \xrightarrow{P} 0 \;\Leftrightarrow\; P(|X_n| \geqslant \varepsilon) \to 0 \quad \text{for all } \varepsilon > 0 \;\Leftrightarrow\; p_n \to 0.$$

(b) Since $E[|X_n|^p] = p_n^p$ for all $n \in \mathbb{N}$, we have

$$X_n \to 0 \quad \text{in } L_P^p(\Omega) \;\Leftrightarrow\; E[|X_n|^p] \to 0 \;\Leftrightarrow\; p_n \to 0.$$

(c) Note, since $X_n$ only attains the values 0 and 1 it follows that

$$\begin{aligned} X_n \to 0 \quad P\text{-a.s.} &\Leftrightarrow P(\lim X_n = 0) = 1 \\ &\Leftrightarrow P(X_n = 0 \text{ for eventually all } n \in \mathbb{N}) = 1. \end{aligned}$$

Recall that

$$\{X_n = 0 \text{ for eventually all } n \in \mathbb{N}\} = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{X_n = 0\} = \liminf_{n\to\infty} \{X_n = 0\}.$$

From Exercise 1.3.1 it follows that

$$\Big(\liminf_{n\to\infty} \{X_n = 0\}\Big)^c = \limsup_{n\to\infty} \{X_n = 1\}.$$

An application of the Borel-Cantelli Lemma 6.2.8 yields:

$$\begin{aligned} \sum p_n < \infty &\Rightarrow P\big(\limsup_{n\to\infty} \{X_n = 1\}\big) = 0, \\ \sum p_n = \infty &\Rightarrow P\big(\limsup_{n\to\infty} \{X_n = 1\}\big) = 1. \end{aligned}$$

Since we have

$$P\big(\liminf_{n\to\infty} \{X_n = 1\}\big) = 1 - P\big(\limsup_{n\to\infty} \{X_n = 1\}\big),$$

we obtain that $X_n \to 0$ $P$-a.s. if and only if $\sum p_n < \infty$.

7. For every $\varepsilon > 0$ and $n \in \mathbb{N}$ with $n \geqslant \frac{1}{\varepsilon}$ one obtains

$$\begin{aligned} P\left(|X_n| \geqslant \varepsilon\right) &= 1 - P\left(-\varepsilon < X_n < \varepsilon\right) \\ &= 1 - \int_{-\varepsilon \wedge \frac{1}{n}}^{\varepsilon \wedge \frac{1}{n}} \frac{n}{2}\, \lambda(du) \\ &= 1 - \int_{-\frac{1}{n}}^{\frac{1}{n}} \frac{n}{2}\, \lambda(du) \\ &= 1 - 1 = 0, \end{aligned}$$

which shows that $X_n \to 0$ in probability. Alternatively, one can show that

$$E[|X|] = \int_{-\frac{1}{n}}^{\frac{1}{n}} |x| \, \frac{n}{2} \, \lambda(du) = \frac{1}{2n} \to 0,$$

which implies $X_n \to 0$ in probability by Theorem 9.1.11.

8. By the definition of the random variables $X_n$ one can expect that $X_n \to 1_{[1,2]}(U) = 1$ for $n \to \infty$. In order to show this let $n \in \mathbb{N}$ and compute

$$E\left[(X_n - 1)^2\right] = 0^2 P(X_n = 1) + 1^2 P(X_n = 0) = P\big(U \in (2 - \tfrac{1}{n}, 2]\big) = \frac{1}{n} \to 0,$$

which shows that $X_n \to 1$ in $\mathscr{L}^2(\Omega, P)$.

9. Note that each $Y \in \mathscr{L}^1_P(\Omega)$ satisfies $|E[Y|\mathscr{D}]| \leqslant E\big[\,|Y|\,|\mathscr{D}\big]$ $P$-a.s. by part (d) in Theorem 8.4.1 since $|Y| - Y \geqslant 0$. Consequently, we obtain for each $n \in \mathbb{N}$:

$$E\big[\,|E[X_n|\mathscr{D}] - E[X|\mathscr{D}]|\,\big] = E\big[\,|E[X_n - X|\mathscr{D}]|\,\big] = E\big[E[|X_n - X|\,|\mathscr{D}]\big] = E[|X_n - X|],$$

which shows $E[X_n|\mathscr{D}] \to E[X|\mathscr{D}]$ in $L^1_P(\Omega)$.

# List of Figures

# Bibliography


[1] P. Billingsley. *Probability and measure. 3rd ed.* Chichester: John Wiley & Sons Ltd., 1995.

[2] L. Breiman. *Probability.* Philadelphia, PA: SIAM, 1992.

[3] M. Capiński and P. E. Kopp. *Measure, integral and probability. 2nd ed.* London: Springer, 2004.

[4] K. L. Chung and F. AitShalia. *Elementary probability theory. With stochastic processes and an introduction to mathematical finance. 4th rev. ed.* New York, NY: Springer, 2003.

[5] R. Durrett. *Probability. Theory and examples. 4th ed.* Cambridge: Cambridge University Press, 2010.

[6] J. Jacod and P. Protter. *Probability essentials. 2nd revised ed.* Berlin: Springer, 2003.

[7] A. F. Karr. *Probability.* New York, NY: Springer-Verlag, 1993.

[8] D. Williams. *Probability with martingales.* Cambridge: Cambridge University Press, 1991.

# Index