Introduction to probability and statistics

Prof. Dr. Alexander Drewitz
Universität zu Köln

Preliminary version of February 1, 2019

Contents

1 Probability theory **5**
- 1.1 A short history of classical probability theory . . . 5
- 1.2 Mathematical modelling of random experiments . . . 6
- 1.3 σ-algebras, probability measures, and probability spaces . . . 8
- 1.4 Examples of probability spaces . . . 11
  - 1.4.1 Discrete probability spaces . . . 11
  - 1.4.2 Continuous probability spaces . . . 13
- 1.5 Conditional probabilities . . . 14
- 1.6 Independence . . . 15
  - 1.6.1 Finite products of probability spaces . . . 17
- 1.7 Random variables . . . 18
- 1.8 Specific distributions . . . 23
  - 1.8.1 Discrete distributions . . . 23
  - 1.8.2 Distribution functions . . . 26
  - 1.8.3 Distributions with densities . . . 27
- 1.9 Expectation . . . 30
  - 1.9.1 Second (and higher) moments . . . 35
- 1.10 Generating functions . . . 39
- 1.11 Convergence of random variables . . . 41
  - 1.11.1 Almost sure convergence . . . 41
  - 1.11.2 Convergence in $\mathcal{L}^p$ . . . 42
  - 1.11.3 Convergence in probability . . . 43
  - 1.11.4 Convergence in distribution . . . 43
- 1.12 Some fundamental tools and inequalities . . . 45
  - 1.12.1 Markov's and Chebyshev's inequalities . . . 45
  - 1.12.2 The Borel-Cantelli lemmas . . . 45
  - 1.12.3 Jensen's inequality . . . 47
- 1.13 Interdependence of types of convergence of random variables . . . 48
- 1.14 Laws of large numbers . . . 50
  - 1.14.1 Weak law of large numbers . . . 50
  - 1.14.2 Strong law of large numbers . . . 52
  - 1.14.3 Are we investigating unicorns? . . . 54
- 1.15 Central limit theorem . . . 54
- 1.16 Markov chains . . . 57
  - 1.16.1 Stationary distribution . . . 59
  - 1.16.2 Classification of states . . . 65
  - 1.16.3 The Perron-Frobenius theorem . . . 66
  - 1.16.4 Quantitative convergence of Markov chains to equilibrium . . . 67

2 Statistics **73**
- 2.1 Estimators . . . 75
  - 2.1.1 Properties of estimators . . . 76
  - 2.1.2 Maximum likelihood estimators . . . 79
  - 2.1.3 Fisher information and the Cramér-Rao inequality . . . 80
  - 2.1.4 Consistency of estimators . . . 83
- 2.2 Confidence regions . . . 87
  - 2.2.1 One recipe for constructing confidence regions . . . 87

2.2.2 Order statistics 88
2.3 Tests 89
2.4 Testing for alternatives (‘Alternativtests’) 93
2.4.1 Chi-square test of goodness of fit (‘Chiquadrat-Anpassungstest’) 95

Chapter 1

Probability theory

1.1 A short history of classical probability theory

Arguably the (first) historically acknowledged hour of birth of basic probability theory was a series of exchanges of letters between Pascal and de Fermat in 1654. Most prominently, they were investigating the so-called 'problem of points', in which the setting is as follows: Two players, A and B, are consecutively playing a game in which each player has the same chance of winning. The first player who wins a fixed number of times (say, three) wins the prize money. Now imagine that due to unforeseen circumstances the tournament has to be stopped prematurely at a time when player A has won twice and player B has won once. How should the prize money be distributed?
In fact, this problem had been around for some time already when Pascal and de Fermat found the right answer: Attempts to solve it had been given by Pacioli in 1494, Cardano in 1539 and Forestani in 1603, just to name a few. The bulk of those solutions either comes out of (mathematically) thin air and looks naive or at least implausible from today's point of view. Moreover, it had also been suggested that the problem is in fact a jurisdictional problem, not a mathematical one.
Either way, in the exchange of letters mentioned above, de Fermat gave a solution to the problem by essentially enumerating all possible outcomes for the remaining games and then distributing the money in the same ratio as these outcomes make player A or B win, respectively.[1] Pascal, on the other hand, realized that this combinatorial approach is getting prohibitively expensive in the case of not three wins, but an arbitrary number of wins. He solved the problem recursively, thereby reducing the computational complexity and at the same time already introducing the fundamental concept of an 'expectation' (without naming it this way), see Definition 1.9.1.
Further motivation came from the desire to gauge insurance premia; it is not completely clear, however, why it was exactly during the above centuries that probability theory came into being. In fact, gambling had been a favorite pastime for millenia,[2] and also insurance and law (see e.g. Noams slowdown paper?) would have proved useful areas of application of probability theory. It may be argued that the means to 'generate' randomness had not been sufficiently sophisticated before, when for example in the place of dice people used so-called 'Astragali', which is a type of bone. Indeed, according to today's knowledge, it was only in the 10th century that all possible results of a series of throwing a die several times had been completely enumerated.
A further impediment might have been that (excessive) gambling had been frowned upon by the Catholic Church. Cardano, who had already been mentioned above, and who also was a medic[3], found a neat way around this moral difficulty when explaining why he dealt with gambling: 'Even if gambling were altogheter an evil, still, on account of the very large number of people who play, it would seem to be a natural evil. For that very reason it ought to be discussed by a medical doctor like one of the incurable diseases.'[4]

[1] Inherent to this approach is the assumption that all those possible outcomes of the remaining games are equally likely (i.e., they have the same probability, although the definition of probability in this setting is not clear either; there are at least two different interpretations, the 'objective' (i.e., derived through symmetries such as in dice or coin tosses – keep in mind, though, that coins and dice do usually have some asymmetries; for example, for a die one usually has a number of cavities on each side specifying the value of the die roll – or by investigating frequencies with which certain events happen) and the 'subjective' (epistemic or Bayesian, where a probability is a measure of the intensity of belief). We will soon have a closer look at the frequentist motivation.

[2] in China, and also around 500 BC people were gambling on the streets of Rome

[3] and who also has the Cardan joint named after him

[4] If you are interested in the history of this so-called 'classical probability theory', you might want to have a look

1.2 Mathematical modelling of random experiments

So while the intuition of probability has been around for a long time from a practical point of view, it was only in 1933[5] that A. N. Kolmogorov started developing an axiomatic mathematical theory, see [Kol33]. While it is worthwhile to have a look at this original article, other texts which are suitable for accompanying this lecture are the books by A. Renyi, [Ren69], I. Schneider [Sch88], G. Gigerenzer and C. Krüger [GK99], as well as U. Krengel [Kre05] and R. Durrett [Dur10].
One of the principal goals of probability theory is to describe experiments which are either random or for which one does not have sufficient information or sufficient computing power to describe them deterministically. For the latter, consider e.g. the rules of playing roulette in a casino. Even after the croupier has spun the wheel as well as the ball, a player is still allowed to make a bet until 'rien ne va plus'. However, once ball and wheel have been spun essentially all information necessary to compute the (now essentially deterministic apart from possible air currents, percussions, etc.) outcome of this particular instance of the game. Hence, if a player was fast enough to use all this information to compute the outcome of this game, she would be able to beat the house. However, without technical support (see e.g. so-called 'roulette computers'), a random modeling of this situation is still very close to reality. In the same spirit, a sequence of die rolls or coin tosses would be considered random experiments.
We will not go into details of discussing the more philosophical question of whether randomness actually does exist or not (see also 'Bell's theorem' for a more philosophical contribution on the existence of randomness).
A possible outcome of an experiment such as the one described above is usually referred to as an *elementary event* and denoted by ω. The set of elementary events is most often denoted by Ω (so $\omega \in \Omega$). An *event* is a set $A \subset \Omega$ of elementary events to which we want to be able to associate a probability $\mathbb{P}(A)$ via a suitable function $\mathbb{P} : 2^\Omega \to [0,1]$ defined on the power set 2^Ω of Ω (i.e., on the set of all subsets of Ω). Ideally, one would like to ba able to make sense of $\mathbb{P}(A)$ for *all* $A \subset \Omega$. However, we will later on see that this is generally asking too much, and for such a function $\mathbb{P}$ to have 'reasonable' properties, it will have to be restricted to a subset of 2^Ω.[6] To a small extent, we will investigate the theoretical basics for this in Section 1.3. However, as an anticipation of our findings below, it will turn out below that as long as Ω is either finite or countable, such problems will not arise.
Given an elementary event $\omega \in \Omega$ (i.e., the outcome of a single realization of a random experiment) and an event $A \subset \Omega$, the interpretation is that *the event A has occurred* if $\omega \in A$ and *A has not occurred* if $\omega \notin A$.
The first goal in describing experiments as outlined above is to find Ω and $\mathbb{P}$ in a way that is suitable for the respective experiment.

Example 1.2.1. *Consider a sequence of three consecutive rolls of a fair die. In this setting, the most obvious choice would arguably be $\Omega = \{1,2,3,4,5,6\}^3$, and $\mathbb{P}$ would be the uniform probability measure on Ω characterized by $\mathbb{P}(\{(\omega_1,\omega_2,\omega_3)\}) = 6^{-3}$ for any $\omega = (\omega_1,\omega_2,\omega_3) \in \Omega$ (since the die is fair, each outcome should have the same probability).*
We could e.g. consider the event A that the second and third dice both show 1, i.e. $A = \{\omega \in \Omega : \omega_2 = \omega_3 = 1\}$. Its probability would be

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(\{\omega \in \Omega : \omega_2 = \omega_3 = 1\}) \\ &= \sum_{\omega\in\Omega\,:\,\omega_2=\omega_3=1} \mathbb{P}(\omega) = |\{\omega \in \Omega : \omega_2 = \omega_3 = 1\}| \cdot 6^{-3} = 6 \cdot 6^{-3} = 1/36\end{aligned}$$

(in the second equality we already used in anticipation, see (1.3.7) below, the fact that we want the probability of an event to be equal to the sum of the probabilities of the one-element events constituted by its elementary events)

Remark 1.2.2. *There are lots of other choices for Ω and $\mathbb{P}$ to describe the above experiment of three consecutive die rolls; however, the above is arguably the most intuitive.*

While the above example was probably quite intuitive, we will now describe how everyday experience leads us to the axiomatic foundations of probability theory, and this is also most likely to be the reason you might have arrived at the solution to the above example on your own.

---

at [Hal90] or also [Das88], where the first one has a more mathematical flavour, and the latter a more philosophical one.

[5] This is the other generally acknowledged hour of birth of probability theory, and one might argue that before, probability theory was more the application of mathematical tools from other areas to problems that somehow involved probability theory.

[6] See also the Banach-Tarski paradox to get a first impression on what could go wrong if one admits arbitrary subsets of Ω.

Imagine performing a certain repeatable experiment n times, and denote the elementary events that you observe by $\omega_1, \dots, \omega_n \in \Omega$. For any event $A \subset \Omega$ you can now record the relative frequency with which this event occurred during those n repetitions of the experiment. Indeed, introducing for $A \subset \Omega$ the *indicator function of* A as

$$\begin{aligned} \mathbb{1}_A : \Omega &\to \{0,1\} \\ \omega &\mapsto \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases} \end{aligned}$$

we can define the relative frequency of A via

$$h(A, \omega_1, \dots, \omega_n) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(\omega_i).$$

Now if we performed the experiment another n times we would probably observe realizations $(\widetilde{\omega}_1, \dots, \widetilde{\omega}_n) \neq (\omega_1, \dots, \omega_n)$, and quite possibly also different relative frequencies

$$h(A, \widetilde{\omega}_1, \dots, \widetilde{\omega}_n) \neq h(A, \omega_1, \dots, \omega_n).$$

However, experience tells us that as $n \to \infty$, the relative quantities should converge to a 'limiting relative frequency', and this gives rise to what we will interpret as the (frequentist) probability of the event A:

$$\mathbb{P}(A) := \lim_{n \to \infty} h(A, \widetilde{\omega}_1, \dots, \widetilde{\omega}_n). \tag{1.2.1}$$

So far, this is no rigorous mathematical definition of $\mathbb{P}(A)$ since it was experience which told us that the limit on the RHS ('right-hand side') of (1.2.1) exists, not a mathematical theorem. Instead, the properties that we observe for relative frequencies will lead us to the axioms which will be the foundations of probability theory. Indeed, we do have the following properties of relative frequencies (independent of the actual realization $\omega_1, \dots, \omega_n$ of the sequence of experiments):

- $$h(A, \omega_1, \dots, \omega_n) \in [0,1]$$
  for all $A \subset \Omega$;
- $$h(\varnothing, \omega_1, \dots, \omega_n) = 0, \quad h(\Omega, \omega_1, \dots, \omega_n) = 1;$$
- if $A, B \subset \Omega$ with $A \cap B = \varnothing$, then
  $$h(A \cup B, \omega_1, \dots, \omega_n) = h(A, \omega_1, \dots, \omega_n) + h(B, \omega_1, \dots, \omega_n).$$
- if $A \subset \Omega$, then
  $$h(A^c, \omega_1, \dots, \omega_n) = 1 - h(A, \omega_1, \dots, \omega_n),$$
  where $A^c := \Omega \backslash A$ is the *complement of A in Ω*.

Motivated by our heuristic guiding identity (1.2.1), we postulate the same properties to hold true for the function $\mathbb{P}$, which we aim to use for prescribing probabilities to subsets of Ω.

- $$\mathbb{P}(A) \in [0,1] \quad \forall A \subset \Omega;$$
- $$\mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(\Omega) = 1; \tag{1.2.2}$$
- If $A, B \subset \Omega$ and $A \cap B = \varnothing$, then
  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \tag{1.2.3}$$
- If $A \subset \Omega$, then
  $$\mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{1.2.4}$$

Remark 1.2.3. *The above is called the 'frequentist' approach to probability. However, sometimes people also want to attribute probabilities to 'experiments' that cannot be repeated easily. E.g., one could ask what the probability of a 'Grexit' (i.e., the event that Greece leaves the Euro zone) within the next five years, that Germany wins the 2022 World Cup, or that you will win the Fields' medal at some point in your career. This then leads to the 'Bayesian' approach to probability (named after 18th century mathematician Thomas Bayes).*

1.3 σ-algebras, probability measures, and probability spaces

As alluded to above, we will in general not be able to assign probabilities to all subsets of Ω. Hence, we will have to restrict to certain subsets $\mathcal{F}$ of its power set 2^Ω. From (1.2.2) to (1.2.4) we readily deduce that the following properties should be fulfilled for such subsets.

Definition 1.3.1. *Let Ω be a non-empty set. A subset $\mathcal{F}$ of 2^Ω is called an* algebra over Ω *if the following properties are fulfilled:*

(a)
$$\Omega \in \mathcal{F}; \tag{1.3.1}$$

(b)
$$A \in \mathcal{F} \text{ implies } A^c \in \mathcal{F}; \tag{1.3.2}$$

(c) For all $n \in \mathbb{N}$,
$$A_1, A_2, \dots, A_n \in \mathcal{F} \text{ implies } \bigcup_{j\in\{1,\dots,n\}} A_j \in \mathcal{F}. \tag{1.3.3}$$

It will, however, turn out that in order to treat the case of infinite Ω and to do interesting things such as asymptotic analysis, the properties of an algebra are not sufficient.
Instead, we will require that $\mathcal{F}$ is also closed under *countable* unions.

Definition 1.3.2. *A subset $\mathcal{F}$ of 2^Ω is called a* σ-algebra over Ω *if the following properties are fulfilled:*

(a)
$$\Omega \in \mathcal{F}; \tag{1.3.4}$$

(b)
$$A \in \mathcal{F} \text{ implies } A^c \in \mathcal{F}; \tag{1.3.5}$$

(c)
$$A_1, A_2, \dots \in \mathcal{F} \text{ implies } \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}. \tag{1.3.6}$$

The following properties can be derived using the definition of a σ-algebra.

Lemma 1.3.3. *Let $\mathcal{F}$ be a σ-algebra over Ω. Then:*

(a) $\varnothing \in \mathcal{F}$;

(b) $A, B \in \mathcal{F}$ implies $A \cup B$, $A \cap B$, $A\triangle B$, and $A\backslash B \in \mathcal{F}$; here, $A\triangle B := (A\backslash B) \cup (B\backslash A)$ denotes the symmetric difference of A and B.

(c) $A_1, A_2, \dots \in \mathcal{F}$ implies that
$$\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{F}.$$

Proof. (a) Note that (1.3.4) in combination with (1.3.5) implies $\varnothing \in \mathcal{F}$.

(b) Choosing $A_1 := A$, $A_2 := B$, $A_i := \varnothing$ for all $i \geqslant 3$, the property $A \cup B \in \mathcal{F}$ follows immediately from part (*a*) and property (1.3.6).

The remaining parts of the proof are left as an exercise (De Morgan's laws might prove useful here). □

Exercise 1.3.4. *If $\mathcal{F}$ is a σ-algebra over Ω and $F \in \mathcal{F}$, then*
$$F \cap \mathcal{F} := \{F \cap G : G \in \mathcal{F}\}$$
is a σ-algebra over F (it is called the trace σ-algebra of F in $\mathcal{F}$*).*

Definition 1.3.5. *If $\mathcal{F}$ is a σ-algebra over Ω, then the pair $(\Omega, \mathcal{F})$ is called a* measurable space.

For a set Ω, a nice σ-algebra $\mathcal{F}$ over Ω will be the set of events that we will be able to gauge in terms of probability. This will usually be done using *probability measures,* which are function with specific properties defined on a σ-algebra contained in 2^Ω and mapping to $[0,1]$. Their exact definition is motivated by and extends the properties of (1.2.2) and (1.2.3).

Definition 1.3.6. *A* probability measure $\mathbb{P}$ *on the measurable space* $(\Omega, \mathcal{F})$ *is a function* $\mathbb{P} : \mathcal{F} \to [0,1]$, *such that the following properties are fulfilled:*

(a)
$$\mathbb{P}(\Omega) = 1;$$

(b) for every sequence $A_1, A_2, \ldots \in \mathcal{F}$ *of pairwise disjoint sets (i.e.,* $A_i \cap A_j = \varnothing$ *for all* $i, j \in \mathbb{N}$ *with* $i \neq j$*), one has*
$$\mathbb{P}\big(\cup_{n\in\mathbb{N}} A_n\big) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n). \tag{1.3.7}$$
This property is usually referred to as σ-additivity.

In this context, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ *is called a* probability space, *and* $\mathcal{F}$ *is often referred to as the* σ-algebra of events. *The properties required for* $\mathbb{P}$ *here are the so-called* Kolmogorov axioms.

Example 1.3.7. • *Let* Ω *be a finite non-empty set and* $\mathcal{F} := 2^\Omega$. *Then*
$$\mathbb{P} : \quad \mathcal{F} \ni F \mapsto \frac{|F|}{|\Omega|}$$
defines a probability measure on $(\Omega, \mathcal{F})$.

Indeed, the RHS is well-defined since $0 < |\Omega|$, *and is an element of* $[0,1]$ *since* $F \subset \Omega$. *In addition,* $\mathbb{P}(\Omega) = 1$. *Now for* (F_n) *a sequence of pairwise disjoint subsets of* Ω *we get*
$$\mathbb{P}(\cup_{n\in\mathbb{N}} F_n) = \frac{|\cup_{n\in\mathbb{N}} F_n|}{|\Omega|} = \sum_{n\in\mathbb{N}} \frac{|F_n|}{|\Omega|} = \sum_{n\in\mathbb{N}} \mathbb{P}(F_n),$$
which establishes the σ*-additivity and hence finishes the proof.*

$\mathbb{P}$ *as defined above is also referred to as the* uniform probability measure on $(\Omega, \mathcal{F})$, *and the corresponding probability space / experiment is oftentimes called* Laplace space *or* Laplace experiment.

•

Definition 1.3.8. *Let* $(\Omega, \mathcal{F})$ *be a measurable space. For* $\omega \in \Omega$, *the* Dirac measure in ω *is defined via*
$$\begin{aligned} \delta_\omega : \mathcal{F} &\to [0,1]. \\ F &\mapsto \mathbb{1}_F(\omega). \end{aligned}$$

It is easy to show that δ_ω *as defined above indeed is a probability measure: By definition,* $\delta_\omega(F) \in \{0,1\} \subset [0,1]$ *for all* $F \in \mathcal{F}$. *In addition,* $\delta_\omega(\Omega) = \mathbb{1}_\Omega(\omega) = 1$, *and for a pairwise disjoint sequence* (F_n) *with* $F_n \in \mathcal{F}$ *for all* $n \in \mathbb{N}$, *we have that*
$$\delta_\omega(\cup_n F_n) = \mathbb{1}_{\cdot\cup_n F_n}(\omega),$$
which equals 1 *if and only if* $\omega \in F_n$ *for some* $n \in \mathbb{N}$ *and* 0 *otherwise. Similarly,*
$$\sum_n \delta_\omega(F_n) = \sum_n \mathbb{1}_{F_n}(\omega),$$
and since the F_n *are pairwise disjoint, again we get that the RHS equals* 1 *if and only if* $\omega \in F_n$ *for some* $n \in \mathbb{N}$ *and* 0 *otherwise. Thus,* δ_ω *defines a probability measure on* $(\Omega, \mathcal{F})$.

We now collect further properties of probability spaces.

Proposition 1.3.9. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space.*

(a)
$$\mathbb{P}(\varnothing) = 0;$$

(b) For all $A, B \in \mathcal{F}$ with $A \cap B = \varnothing$ we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \quad \textit{(additivity).} \tag{1.3.8}$$

(c) For all $A \in \mathcal{F}$,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

(d) For all $A, B \in \mathcal{F}$ with $A \subset B$, one has

$$\mathbb{P}(A) \leqslant \mathbb{P}(B) \quad \textit{(monotonicity).} \tag{1.3.9}$$

(e) For every sequence $(A_n)_{n\in\mathbb{N}}$ with $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$,

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) \leqslant \sum_{n\in\mathbb{N}} \mathbb{P}(A_n) \quad \textit{(σ-subadditivity).}$$

(f) Let $(A_n)_{n\in\mathbb{N}}$ be a sequence with $A_n \in \mathcal{F}$ and $A_n \subset A_{n+1}$ for all $n \in \mathbb{N}$. Then, with $A := \cup_{n\in\mathbb{N}} A_n$, one has

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) \quad \textit{(continuity of probability measures from below)}$$

(g) Let $(A_n)_{n\in\mathbb{N}}$ be a sequence with $A_n \in \mathcal{F}$ and $A_n \supset A_{n+1}$ for all $n \in \mathbb{N}$. Then, with $A := \cap_{n\in\mathbb{N}} A_n$, one has

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) \quad \textit{(continuity of probability measures from above)}$$

Proof. (a) We have $\varnothing = \dot\cup_{n=1}^{\infty} \varnothing$ (where we write $\dot\cup$ in order to emphasise that it is a union over disjoint sets), hence the σ-additivity supplies us with

$$\mathbb{P}(\varnothing) = \sum_{n\in\mathbb{N}} \mathbb{P}(\varnothing).$$

Since $\mathbb{P}(\varnothing) \in [0, 1]$, the only value of $\mathbb{P}(\varnothing)$ for which this equality can hold true is $\mathbb{P}(\varnothing) = 0$.

(b) Setting $A_1 := A$, $A_2 := B$, and $A_i := \varnothing$ for all $i \geqslant 3$, this follows from the σ-additivity (1.3.7) of probability measures.

(c) We have $A \dot\cup A^c = \Omega$, hence the additivity (1.3.8) provides us with

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1,$$

from which the result follows.

(d) We have $B = A \dot\cup (B \backslash A)$, and hence additivity gives

$$\mathbb{P}(A) + \mathbb{P}(B \backslash A) = \mathbb{P}(B).$$

Since $\mathbb{P}$ takes non-negative values only, the claim follows.

(e) We define $B_n := A_n \backslash \cup_{j=1}^{n-1} A_j$. Then $B_n \in \mathcal{F}$ for all n, they form a sequence of pairwise disjoint sets, and $\cup_{n\in\mathbb{N}} A_n = \dot\cup_{n\in\mathbb{N}} B_n$. Thus,

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \mathbb{P}(\dot\cup_{n\in\mathbb{N}} B_n) \overset{\sigma\text{-additivity}}{=} \sum_{n\in\mathbb{N}} \mathbb{P}(B_n) \leqslant \sum_{n\in\mathbb{N}} \mathbb{P}(A_n),$$

where the last inequality follows from (1.3.9).

(f) As before, set $B_n := A_n \backslash \cup_{j=1}^{n-1} A_j$. We get $A = \cup_{n\in\mathbb{N}} B_n$, and using the monotonicity of the sequence $(A_n)_{n\in\mathbb{N}}$ also $A_m = \cup_{n=1}^{m} B_n$. Thus,

$$\mathbb{P}(A_m) = \mathbb{P}(\cup_{n=1}^{m} B_n) = \sum_{n=1}^{m} \mathbb{P}(B_n).$$

Taking limit on both sides we obtain

$$\lim_{m\to\infty} \mathbb{P}(A_m) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \mathbb{P}(\cup_{n\in\mathbb{N}} B_n) = \mathbb{P}(A).$$

(g) Left as an exercise.

□

The following result helps computing the probabilities of unions of events which are not necessarily pairwise disjoint.

Lemma 1.3.10. *(a) For $A, B \in \mathcal{F}$,*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

(b) Let $n \geqslant 2$, $A_1, \ldots, A_n \in \mathcal{F}$. Then

$$\mathbb{P}\Big(\bigcup_{k=1}^{n} A_k\Big) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}). \quad \textit{(Inclusion-exclusion formula)}$$

Proof. (a) We have $A \cup B = A \dot\cup (B \backslash A) \in \mathcal{F}$ and thus

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A).$$

In addition, $B = (B \backslash A) \dot\cup (A \cap B)$, which in combination with the above yields

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

and finishes the proof.

(b) We proceed by induction. For $n = 2$, this is part (a). Now assume the statement holds for some $n \geqslant 2$. Then by part (a),

$$\mathbb{P}\Big(\bigcup_{k=1}^{n+1} A_k\Big) = \mathbb{P}\Big(\bigcup_{k=1}^{n} A_k\Big) + \mathbb{P}(A_{n+1}) - \mathbb{P}\Big(\bigcup_{k=1}^{n} (A_{n+1} \cap A_k)\Big).$$

Applying the induction assumption to the first and third summand we can continue to get

$$\begin{aligned}
&\sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) + \mathbb{P}(A_{n+1}) \\
&\quad - \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}(A_{n+1} \cap A_{i_1} \cap \ldots \cap A_{i_k}) \\
&= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) + \mathbb{P}(A_{n+1}) \\
&\quad + \sum_{k=2}^{n+1} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_{k-1} < i_k = n+1} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) \\
&= \sum_{k=1}^{n+1} (-1)^{k-1} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant n+1} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}),
\end{aligned}$$

which proves the induction step.

□

1.4 Examples of probability spaces

1.4.1 Discrete probability spaces

As already mentioned before, a principal part of this lecture is concerned with those instances of Ω where we don't run into technical difficulties concerning the measurability of subsets of Ω, i.e., with so-called 'discrete probability' spaces.

Definition 1.4.1. *If Ω is either finite or countably infinite, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}$ some probability measure on $(\Omega, \mathcal{F})$, then $(\Omega, \mathcal{F}, \mathbb{P})$ is called a* discrete probability space.

In the setting of discrete probability spaces, we can always associate a probability to *any* subset $A \in 2^\Omega$ of Ω (which is not true for general probability space, see Remark 1.4.8 below). Also, note that in this context, $\mathbb{P}$ is uniquely defined by the values $\mathbb{P}(\{\omega\})$ for one point sets, for which we often write $\mathbb{P}(\omega)$ – both these properties are not necessarily true anymore for uncountable Ω. We have already seen a first example for a probability space in Example 1.2.1.

Claim 1.4.2. *If $(\Omega, 2^\Omega, \mathbb{P})$ is a discrete probability space, then for all $A \in \mathcal{F}$,*

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega). \tag{1.4.1}$$

In particular, $\mathbb{P}$ is uniquely determined by its values on one-point sets.

Proof. Since $A \in 2^\Omega$, and $\{\omega\} \in 2^\Omega$ for all $\omega \in \Omega$, the σ-additivity of $\mathbb{P}$ immediately implies (1.4.1). □

Example 1.4.3. *Consider the situation where you toss two fair and indistinguishable coins. We can choose $\Omega := \{(T,T), (H,H), (H,T)\}$ where the elementary event $\omega = (T,T)$ corresponds to both coins showing tails, $\omega = (H,H)$ corresponds to both coins showing heads and (H,T) corresponds to one coin showing heads and and the other coin showing tails. As common in the case of a discrete space, we endow Ω with the power set σ-algebra 2^Ω. Due to Claim 1.4.2 it is now sufficient to give all the values of $\mathbb{P}(\omega)$. The cases (T,T) and (H,H) can be handled along the same punchline as in Example 1.2.1 yielding $\mathbb{P}((H,H)) = \mathbb{P}((T,T)) = \frac{1}{4}$. Thus, by Proposition 1.3.9 (c), we infer that $\mathbb{P}((H,T)) = 1 - \mathbb{P}(\{(H,H)\} \cup \{(T,T)\}) = 1 - \frac{1}{2} = \frac{1}{2}$. While this apparent asymmetry might look slightly puzzling at a first glance, it is caused by the fact that there is only one possible outcome of the experiment leading to both coins showing head or tails; on the other hand, since the coins cannot be distinguished, there are two outcomes leading to seeing one coin showing heads and the other showing tails. This explains the above choice of probabilities.*

Exercise 1.4.4. *Assuming a uniform distribution on all orderings that a deck of 52 cards numbered from 1 to 52 can be in (i.e., all orderings have the same probability), what is the probability that at least one of the cards is in the right place (where by 'right place' we mean that the card with number m is in the m-th position)?*
Solution: Let

$$\Omega := \Big\{(x_1, \ldots, x_{52}) \ : \ x_i \in \{1, 2, \ldots, 52\} \ \forall i \in \{1, 2, \ldots, 52\} \text{ and } x_i \neq x_j \ \forall i \neq j\Big\}$$

be the space of all orderings, $\mathcal{F} := 2^\Omega$, and for any $A \in \mathcal{F}$ we set $\mathbb{P}(A) := \frac{|A|}{|\Omega|} = \frac{|A|}{52!}$.
Then according to Example 1.3.7, $(\Omega, \mathcal{F}, \mathbb{P})$ actually defines a probability space.
Writing $A_i := \{\omega \in \Omega \ : \ \omega_i = i\}$ for the event that the i-th card is in the right place, we get $\mathbb{P}(A_1) = \frac{51!}{52!} = \frac{1}{52}$, $\mathbb{P}(A_1 \cap A_2) = \frac{50!}{52!} = \frac{1}{52 \cdot 51}$, or more generally, for $1 \leqslant i_1 < \ldots < i_k \leqslant 52$,

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \frac{1}{52 \cdot 51 \cdot \ldots \cdot (52-k+1)!} = \frac{(52-k)!}{52!}.$$

The event that at least one card is in the right position can then be written as

$$\begin{aligned}\{\omega \in \Omega \ : \ \exists i \in \{1, \ldots, 52\} \text{ such that } \omega_i = i\} &= \{\omega \in \Omega \ : \ \exists i \in \{1, \ldots, 52\} \text{ such that } \omega \in A_i\} \\ &= \cup_{i=1}^{52} A_i.\end{aligned}$$

Hence, using the inclusion-exclusion formula we compute

$$\mathbb{P}(\cup_{i=1}^{52} A_i) = \sum_{k=1}^{52} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant 52} (-1)^{k-1} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}).$$

Using that

$$\big|\{(i_1, i_2, \ldots i_k) \in \{1, 2, \ldots, 52\}^k \ : \ 1 \leqslant i_1 < \ldots < i_k \leqslant 52\}\big| = \binom{52}{k},$$

we can continue the above as

$$= \sum_{k=1}^{52} (-1)^{k-1} \binom{52}{k} \frac{(52-k)!}{52!} = \sum_{k=1}^{52} (-1)^{k-1} \frac{1}{k!}.$$

Note that if we replace 52 *by* n *in the above and take* $n \to \infty$*, we see that the probability we are interested in converges to* $1 - e^{-1}$.

As the above examples already suggest, the first goal in describing random experiments is to find a corresponding probability space. For this purpose, it is essential to have a precise set of rules according to which the random experiments are performed. To get an understanding of what kind of ambiguities might occur, have a look at the so-called *Bertrand's paradox* or the Monty-Hall problem ('Ziegenproblem'), the latter of which is also planned to be investigated in the exercise classes.
A commonly used concept that has already appeared before is the following (see also Example 1.3.7).

Definition 1.4.5. *A discrete probability space* $(\Omega, \mathcal{F}, \mathcal{P})$ *with* $|\Omega| < \infty$ *is called* Laplace Probability Space *if* $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ *for all* $A \in \mathcal{F}$.

As in the case of rolling dice, the Laplace probability space (or 'Laplace model') is reasonable if none of the possible outcomes of the random experiments seems favourable over any other.

Remark 1.4.6. *For computing probabilities in Laplace experiments (such as Example 1.2.1), 'counting' is essentially enough to determine the respective probabilities. Probability spaces on the other hand provide an elegant tool for coping with more complex situations: Assume for example that we had reason to suspect that the die in Example 1.2.1 is three time as likely to show* 6 *than any other number. We can then define another probability measure* $\mathbb{P}^*$ *on the same measurable space* $(\Omega, \mathcal{F})$*, which is characterized via*

$$\mathbb{P}^*(\omega) := \prod_{i=1}^{3} \Big(\frac{1}{8} \mathbb{1}_{\{\omega \in \Omega \,:\, \omega_i \neq 6\}}(\omega) + \frac{3}{8} \mathbb{1}_{\{\omega \in \Omega \,:\, \omega_i = 6\}}(\omega) \Big), \quad \forall \omega \in \Omega.$$

Convince yourself that this indeed defines a probability measure! Note that in this notation, the previous probability measure $\mathbb{P}$ *corresponding to the Laplace experiment would just be characterized via* $\mathbb{P}(\omega) = \frac{1}{6^3}$ *for all* $\omega \in \Omega$.

1.4.2 Continuous probability spaces

So far have we have only investigated discrete probability spaces. However, it turns out to be necessary to consider more general setups.

Example 1.4.7. *Consider a needle thrown in the plane* $\mathbb{R}^2$ *and denote by* $\varphi \in [0, 2\pi)$ *the angle that the line obtained by continuing the needle to infinity in both directions encloses with the* x*-axis.*
One intuitive choice would then be to set $\Omega = [0, 2\pi)$*, and we would also like to have that*

$$\mathbb{P}([a, b)) = \frac{b-a}{2\pi} \quad \text{for any } 0 \leqslant a < b < 2\pi. \tag{1.4.2}$$

As a consequence, we would like the corresponding σ*-algebra to contain all intervals* $[a, b) \subset [0, 2\pi)$ *at least. The smallest such* σ*-algebra is denoted by* $\mathcal{B}([0, 2\pi))$*, and called the 'Borel-*σ*-algebra on* $[0, 2\pi)$*'.*

Remark 1.4.8. *In the subsequent course 'Probability Theory I' we will show that there exists a unique probability measure* $\mathbb{P}$ *on* $\mathcal{B}([0, 2\pi))$ *such that* (1.4.2) *holds true,* $\mathcal{B}([0, 2\pi))$ *contains all intervals* $[a, b) \subset [0, 2\pi)$*, and such that* $\mathbb{P}$ *cannot be extended to a probability measure on the power set* $2^{[0,2\pi)}$.

The following exercise shows that it does indeed make sense to talk about the smallest σ-algebra containing some subset of 2^Ω as done in the previous example

Exercise 1.4.9. *Let* $(\mathcal{F}_\lambda)_{\lambda \in \Lambda}$ *be a family of* σ*-algebras over a non-empty set* Ω*. Show that*

$$\cap_{\lambda \in \Lambda} \mathcal{F}$$

again is a σ*-algebra over* Ω.

Definition 1.4.10. *Let* Ω *be a non-empty set and* $\mathcal{D} \subset 2^\Omega$*. Then the smallest* σ*-algebra containing* $\mathcal{D}$ *is denoted by*

$$\sigma(\mathcal{D}) := \bigcap_{\mathcal{E}} \mathcal{E},$$

where the intersection is over all σ*-algebras* $\mathcal{E}$ *on* Ω *and containing* $\mathcal{D}$*. It is well-defined due to Exercise 1.4.9.*

1.5 Conditional probabilities

Above we had introduced the concept of a probability measure. Sometimes we already have some partial information available on the outcome of an event. It is then immediate to ask how this additional information 'changes' the probabilities of the outcome. E.g., again in the above Example 1.2.1 of rolling dice, imagine there is an oracle that tells you the numbers shown by the first two dice. Assume that via this oracle you now know that the first and second die both show 6. While originally the probability of three dice showing 6 is 6^{-3}, with this additional information at hand you might rather say it should be 1/6. Formally:

Definition 1.5.1. *Let $F, G \in \mathcal{F}$ be such that $\mathbb{P}(G) > 0$. Then we define the* conditional probability of F given G *as*

$$\mathbb{P}(F \,|\, G) := \frac{\mathbb{P}(F \cap G)}{\mathbb{P}(G)}. \tag{1.5.1}$$

Recalling our guiding motivation for properties of probability measures via (limiting) relative frequencies, the RHS of (1.5.1) would correspond to the (limiting) relative frequency of experiments for which F occurs among those experiments for which G occurs:

$$\frac{h(F \cap G, \omega_1, \ldots, \omega_n)}{h(G, \omega_1, \ldots, \omega_n)} = \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{F\cap G}(\omega_i)}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{G}(\omega_i)} = \frac{\sum_{i=1}^n \mathbb{1}_{F\cap G}(\omega_i)}{\sum_{i=1}^n \mathbb{1}_{G}(\omega_i)}.$$

Exercise 1.5.2. *Check that with $F := \{\omega_1 = \omega_2 = 6\}$ and $G := \{\omega_1 = \omega_2 = \omega_3 = 6\}$ we do indeed get that $\mathbb{P}(G \,|\, F) = 1/6$.*

Proposition 1.5.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $G \in \mathcal{F}$ such that $\mathbb{P}(G) > 0$ the function*

$$\mathcal{F} \ni A \mapsto \mathbb{P}(A \,|\, G) \tag{1.5.2}$$

defines a probability measure on $(\Omega, \mathcal{F})$ with $\mathbb{P}(G) = 1$.

Proof. Since $\mathbb{P}(G) > 0$ by assumption, the function in (1.5.2) is well-defined. Since for any $A \in \mathcal{F}$ one has $\mathbb{P}(A \cap G) \leqslant \mathbb{P}(G)$, the function maps from $\mathcal{F}$ to $[0, 1]$. Furthermore, $\mathbb{P}(G \,|\, G) = \frac{\mathbb{P}(G\cap G)}{\mathbb{P}(G)} = 1$ and $\mathbb{P}(\varnothing \,|\, G) = \frac{\mathbb{P}(\varnothing\cap G)}{\mathbb{P}(G)} = 0$, so it only remains to check the σ-additivity of the function. For this purpose, let (A_n) be a sequence of pairwise disjoint elements of $\mathcal{F}$. Then

$$\mathbb{P}(\cup_n A_n \mid G) = \frac{\mathbb{P}((\cup_n A_n) \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(\dot\cup_n (A_n \cap G))}{\mathbb{P}(G)} \overset{\sigma\text{-additivity of } \mathbb{P}}{=} \sum_{n\in\mathbb{N}} \frac{\mathbb{P}(A_n \cap G)}{\mathbb{P}(G)} = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n \,|\, G),$$

which proves the σ-additivity of the function in (1.5.2), which thus defines a probability measure. □

The following result is interesting in its own right and will also be used the proof of Bayes' formula below.

Theorem 1.5.4 (Law of Total Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(B_n)_{n\in I}$ be an at most countable partition (i.e., I is at most countable) of Ω into events; i.e., one has $B_i \cap B_j = \varnothing$ for all $i, j \in I$ with $i \neq j$, as well as $\cup_{n\in I} B_n = \Omega$, and $B_n \in \mathcal{F}$ for all $n \in I$. Then for all $A \in \mathcal{F}$,*

$$\mathbb{P}(A) = \sum_{n\in I} \mathbb{P}(A \,|\, B_n) \cdot \mathbb{P}(B_n),$$

where for simplicity of notation we set $\mathbb{P}(A \,|\, B_n) = 0$ if $\mathbb{P}(B_n) = 0$.

Proof. Since the $(B_n)_{n\in I}$ form a partition with I at most countable, we get using the (σ)-additivity of $\mathbb{P}$ that

$$\mathbb{P}(A) = \mathbb{P}(A\cap\dot\cup_{n\in I} B_n) = \mathbb{P}(\dot\cup_{n\in I}(A\cap B_n)) = \mathbb{P}\big(\dot\cup_{n\in I}(A\cap B_n)\big) = \sum_{\substack{n\in I\\ \mathbb{P}(B_n)>0}} \mathbb{P}(A\cap B_n) = \sum_{n\in I} \mathbb{P}(A \,|\, B_n)\cdot\mathbb{P}(B_n),$$

where as before for simplicity of notation we set $\mathbb{P}(A \,|\, B_n) = 0$ if $\mathbb{P}(B_n) = 0$. □

This will prove useful in particular in the statistics part of the lecture, where we will go into further detail regarding this result.

Theorem 1.5.5 (Bayes' Formula)**.** *Let the assumptions of Theorem 1.5.4 be fulfilled and let* $\mathbb{P}(A) > 0$*. Then for any* $n \in I$*,*

$$\mathbb{P}(B_n \,|\, A) = \frac{\mathbb{P}(A \,|\, B_n) \cdot \mathbb{P}(B_n)}{\sum_{j \in I} \mathbb{P}(A \,|\, B_j) \cdot \mathbb{P}(B_j)}.$$

In particular, for any $B \in \mathcal{F}$

$$\mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(A \,|\, B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)},$$

where again we interpret $\mathbb{P}(A \,|\, B) = 0$ *if* $\mathbb{P}(B) = 0$*.*

Proof. Using the definition of conditional probabilities, we get

$$\mathbb{P}(B_n \,|\, A) = \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \,|\, B_n) \cdot \mathbb{P}(B_n)}{\mathbb{P}(A)},$$

and applying the Law of Total Probability to the denominator, the result follows. □

The above theorem can lead to surprising results.

Example 1.5.6. *Assume the setting of testing people on a certain illness. In this setting, a quantity that is of principal interest in applications is the conditional probability that a person is sick given that the test has shown a positive result.*
Assume an appropriate probability space to be given, denote by B *the event that a person is sick and in order to be specific let us say that* $\mathbb{P}(B) = 10^{-5}$*. Assume that we know the test gives the correct result in 99% of the cases, no matter whether a person is sick or not. In particular, denoting by* A *the event that the test shows a positive result, we infer*

$$\mathbb{P}(A \,|\, B) = .99 \quad \text{and} \quad \mathbb{P}(A \,|\, B^c) = .01, \tag{1.5.3}$$

As alluded to before, the quantity of major interest is $\mathbb{P}(B \,|\, A)$*. Now Bayes' Theorem tells us that we can compute it once we know* $\mathbb{P}(A \,|\, B)$*,* $\mathbb{P}(A)$*, and* $\mathbb{P}(B)$*.*
But using (1.5.3) *and the Law of Total Probability, we infer that*

$$\mathbb{P}(A) = \mathbb{P}(A \,|\, B)\mathbb{P}(B) + \mathbb{P}(A \,|\, B^c)\mathbb{P}(B^c) = .99 \cdot 10^{-5} + 0.01 * (1 - 10^{-5}) \approx 0.01.$$

Then Bayes' Theorem supplies us with

$$\mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(A \,|\, B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)} = \frac{.99 \cdot 10^{-5}}{.01} \approx 0.001.$$

It may seem very surprising at first glance that in the above example the probability that a person is sick, conditionally on the test showing a positive result, is so small. The reason for this is, as one may infer from the above computations, that the occurrence of this certain illness in the population is significantly smaller than the probability with which the test fails.

1.6 Independence

We have already used the concept of independence in Example 1.2.1 above, when we tacitly assumed that the outcome of one of the die does not have any influence on the probabilities of the occurrence of the outcomes of the remaining dice (or, equivalently, to justify the use of a Laplace probability space also for the threefold die roll). In essence, two events are independent if the occurrence of one event does not influence the probability of occurrence of the other event, and vice versa. Put in mathematical terms, events $F, G \in \mathcal{F}$ with $\mathbb{P}(G) > 0$ would thus be considered independent if

$$\mathbb{P}(F \,|\, G) = \mathbb{P}(F). \tag{1.6.1}$$

In terms of our interpretation of relative frequencies, this means that the (limiting) relative frequency of F is not changed if we restrict to those experiments for which G occurs.
Multiplying (1.6.1) on both sides by $\mathbb{P}(G)$, this can be rewritten as

$$\mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G),$$

which is the guiding identity for the following definition.

Definition 1.6.1. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a family of events $(A_\lambda)_{\lambda\in\Lambda} \subset \mathcal{F}$, $\Lambda \neq \varnothing$, is called* independent *if for any $J \subset \Lambda$ finite one has*

$$\mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbb{P}(A_j). \tag{1.6.2}$$

In this context we usually just say that the A_λ are independent.

If two events $A, B \in \mathcal{F}$ are independent, then we sometimes also use the notation $A \perp B$.
So far we have mostly been dealing with a finite number of events. It might not be apparent yet, but we will later on deal with infinite numbers of independent events and hence give the above definition in full generality already.

Example 1.6.2. *(a) In any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the events $\Omega, \varnothing \in \mathcal{F}$ are independent. Indeed, we only have to check* (1.6.2) *for $n = 2$, for which we get*

$$\mathbb{P}(\Omega \cap \varnothing) = \mathbb{P}(\varnothing) = 0$$

and

$$\mathbb{P}(\Omega) \cdot \mathbb{P}(\varnothing) = 1 \cdot 0 = 0.$$

However, if $F \in \mathcal{F}$ with $\mathbb{P}(F) \in (0, 1)$, then F and F^c are not independent. Indeed, one has $\mathbb{P}(F \cap F^c) = \mathbb{P}(\varnothing) = 0$, but $\mathbb{P}(F)\mathbb{P}(F^c) \in (0, 1)$.

(b) Two events A, B with $\mathbb{P}(B) > 0$ are independent if and only if $\mathbb{P}(A) = \mathbb{P}(A \,|\, B)$.

Indeed, multiplying both sides of this equation by $\mathbb{P}(B) > 0$ we immediately get $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B)$, which is the equality characterizing the independence of two events.

(c) Consider again the Laplace probability space of three consecutive dice rolls from Example 1.2.1. Our guiding intuition in using the Laplace assumption, i.e., assigning to any possible outcome of these three dice rolls the same probability, was hinging on the assumption that the result of one of the dice rolls should have no influence on the probabilities of the others. For example, if we introduce the events

$$A_1 := \{\omega \in \Omega \,:\, \omega_1 \leqslant 3\}, \tag{1.6.3}$$

$$A_3 := \{\omega \in \Omega \,:\, \omega_3 \text{ is even}\}, \tag{1.6.4}$$

we see that

$$\mathbb{P}(A_1 \cap A_3) = \frac{|A_1 \cap A_3|}{|\Omega|} = \frac{3 \cdot 3 \cdot 6}{6^3} = 1/4.$$

On the other hand,

$$\mathbb{P}(A_1)\mathbb{P}(A_3) = \frac{3}{6}\frac{3}{6} = 1/4,$$

and thus A_1 and A_3 are independent.

Remark 1.6.3. *It is important here to notice that two events can be independent, although the occurrence of one event does have an influence on how the other event can be realized. For this purpose, consider in Example 1.2.1 the events*

$$A_1 := \{\omega \in \Omega \,:\, \omega_1 \text{ is even}\}$$

and

$$A_2 := \{\omega \in \Omega \,:\, \omega_1 + \omega_2 \text{ is even}\}.$$

Now if we know that A_1 occurs, then A_2 will be realized if and only if the of the second die shows an even number, whereas if A_1 does not occur, then A_2 can be realized if and only if the second die shows an odd number. Hence, there is a causal dependence between A_1 and A_2, and in common language we would not call the events A_1 and A_2 independent. However, in the sense of Definition 1.6.1, they are independent, since we have

$$\mathbb{P}(A_1)\mathbb{P}(A_2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4},$$

and

$$\mathbb{P}(A_1 \cap A_2) = \frac{3 \cdot 3 \cdot 6}{6^3} = \frac{1}{4}.$$

This is because behind Definition 1.6.1 was the equality of (1.6.1)*, which means that the occurrence of* A_1 *should not have any influence on the* probability *with which* A_2 *occurs, not on* how *the event* A_2 *is actually realized; and this is indeed the case as the last two displays show us.*

Exercise 1.6.4. *(a) Show that if the dice are biased (e.g., they slightly favour even over odd numbers – this would require a corresponding adaptation of the probability measure* $\mathbb{P}$*) in this example, then* A_1 *and* A_2 *are not independent anymore.*

(b) Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $(A_\lambda)_{\lambda \in \Lambda} \subset \mathcal{F}$ *a family of events. Show that this family is independent if and only if for any subset* $J \subset \Lambda$*, the family* $(A_\lambda)_{\lambda \in J} \subset \mathcal{F}$ *is an independent family of events.*

Example 1.6.5. *Give an example which shows that for events* $A_1, \ldots, A_n \in \mathcal{F}$*, the validity of*

$$\mathbb{P}(A_k \cap A_l) = \mathbb{P}(A_k)\mathbb{P}(A_l) \quad \forall 1 \leqslant k < l \leqslant n, \tag{1.6.5}$$

does not yet imply the independence of the events $A_1, \ldots, A_n \in \mathcal{F}$*.*
In particular, this means that if the events $A_1, \ldots, A_n$ *are pairwise independent, (i.e., if* $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ *for all* $1 \leqslant i < j \leqslant n$*), it does not necessarily imply that the events* $A_1, \ldots, A_n$ *are independent.*
Solution: *In our favourite Example 1.2.1 define*

$$A_1 := \{\omega \,:\, \omega_1 \text{ is even }\},$$

$$A_2 := \{\omega \,:\, \omega_2 \text{ is even }\},$$

and

$$A_3 := \{\omega \,:\, \omega_1 + \omega_2 \text{ is even }\}.$$

It is then not hard to check that (1.6.5) *holds for* $n = 3$*, i.e., the events are pairwise independent; however, we have*

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{3 \cdot 3 \cdot 6}{6^3} = \frac{1}{4},$$

whereas

$$\mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}.$$

Thus, the events A_1*,* A_2*, and* A_3 *are not independent.*

While above we learned one way how to check independence of events, we will now provide a short interlude for the reverse problem, i.e., how to model independent events.

1.6.1 Finite products of probability spaces

In our guiding Example 1.2.1, intuition led us to a reasonable probability space for modeling the experiment. How do we treat more complicated situations of modeling (possibly different) 'independent' experiments?
If $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1), \ldots, (\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ are discrete probability spaces used for modeling n possibly different experiments, then if the result of one a subset of those experiments does not influence the probabilities of the outcomes of the remaining experiments, they can be modeled as a single experiment modeled by the following product space

$$(\Omega_1 \times \ldots \times \Omega_n, \mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n, \mathbb{P}_1 \otimes \ldots \otimes \mathbb{P}_n), \tag{1.6.6}$$

with

$$\mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n = 2^{\Omega_1 \times \ldots \times \Omega_n},$$

and where the product measure is completely defined via

$$\mathbb{P}_1 \otimes \ldots \otimes \mathbb{P}_n((\omega_1, \ldots, \omega_n)) := \prod_{k=1}^{n} \mathbb{P}_k(\omega_k), \quad \forall (\omega_1, \ldots, \omega_n) \in \Omega_1 \times \ldots \times \Omega_n. \tag{1.6.7}$$

Exercise 1.6.6. *(a) Show that* (1.6.7) *does define a probability measure on* $(\Omega_1 \times \ldots \times \Omega_n, 2^{\Omega_1 \times \ldots \times \Omega_n})$.

(b) If $A_k \subset \Omega_k$ *describes a possible outcome of the* k*-th experiment in the original setup, then in the product space setup this corresponds to*

$$B_k := \Omega_1 \times \Omega_{k-1} \times A_k \times \Omega_{k+1} \times \Omega_n,$$

and $A_1 \times \ldots \times A_n = \cap_{k=1}^n B_k$ *describes the event that the* k*-th experiment has output* A_k*, for all* $1 \leqslant k \leqslant n$*, in the product space setup. Show that the events* $B_1, \ldots, B_n$ *are independent.*

Example 1.6.7. *If, for example, we consider the roll of a loaded die as in Remark 1.4.6 and the toss of one of the two fair coins of Example ex:indistCoins, we can take*

$$\Omega = \{1,2,3,4,5,6\} \times \{H,T\}, \quad \mathcal{F} = 2^\Omega,$$

and the probability measure $\mathbb{P}$ *on* $(\Omega, \mathcal{F})$ *characterized through*

$$\mathbb{P}((\omega_1,\omega_2)) = \Big(\frac{1}{8}\mathbb{1}_{\{\omega\in\Omega\,:\,\omega_1\neq 6\}}(\omega) + \frac{3}{8}\mathbb{1}_{\{\omega\in\Omega\,:\,\omega_1=6\}}(\omega)\Big) \times \frac{1}{2}, \quad \forall \omega \in \Omega.$$

Again, check that $\mathbb{P}$ *defines a probability measure* $\mathbb{P}$ *on* $(\Omega, 2^\Omega)$*, so in particular* $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega} \mathbb{P}(\omega) = 1$.

1.7 Random variables

So far, we have described the outcomes of random experiments we were interested in by corresponding events of the underlying and suitably chosen probability space. E.g., in the basic Example 1.2.1 of rolling three dice we could describe the event that the first die shows a number not larger than three and at the same time the third die shows an even number via the event $A_1 \cap A_3$, see (1.6.3) and (1.6.4).
Now on one hand, in more complex situations this approach becomes more and more cumbersome. On the other hand, in probability theory one is oftentimes not interested so much in the very structure of the underlying probability space (for which there might be many choices as we shortly outlined below Example 1.2.1), but rather in observables of certain experiments (such as e.g. the sum of the three dice rolls in Example 1.2.1. Thus, it turns out to be useful to be able to describe outcomes of experiments without possibly knowing the specific structure of the underlying probability space. The following definition will serve as a first step into that direction.

Definition 1.7.1. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $(E, \mathcal{E})$ *a measurable space. Then a function* $X : \Omega \to E$ *is called a* random variable ('Zufallsvariable') *if for all* $A \in \mathcal{E}$ *its preimage under* X *is contained in* $\mathcal{F}$*, i.e., if* $X^{-1}(A) := \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$.
More generally, for measurable spaces $(E, \mathcal{E})$ *and* $(G, \mathcal{G})$*, a mapping* $f : E \to G$ *is called* $\mathcal{E}-\mathcal{G}$*-measurable if* $f^{-1}(A) \in \mathcal{E}$ *for all* $A \in \mathcal{G}$*. Hence, a random variable from* $(\Omega, \mathcal{F})$ *to* $(E, \mathcal{E})$ *is nothing else than an* $\mathcal{F} - \mathcal{E}$*-measurable function. For a random variable* X *the values* $X(\omega)$*,* $\omega \in \Omega$*, are called* realizations *of the random variable* X.

Oftentimes in this introductory class we will be interested in the case where E is just a finite set and $\mathcal{E} = 2^E$. Also, by convention, random variables are usually denoted using upper case letters such as X and Y, and the shorthand $\mathbb{P}(X \in A) := \mathbb{P}(X^{-1}(A))$ for $A \in \mathcal{E}$ is very common.

Example 1.7.2. *(a) Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a discrete probability space and* $(E, \mathcal{E})$ *an arbitrary measurable space. Then any function* $X : \Omega \to E$ *is a random variable.*

Indeed, for any $A \in \mathcal{E}$ *we get that* $X^{-1}(A) \in 2^\Omega = \mathcal{F}$*, and hence* X *is a random variable.*

(b) In the setting of Example 1.2.1 we can e.g. define the mappings X_n*,* $1 \leqslant n \leqslant 3$*, via* $X_n(\omega) := \omega_n$*, which due to part (a) are thus random variables, and* X_n *describes the outcome of the* n*-th die.*

The observable 'sum of the three dice rolls' can then be written as

$$\sum_{i=1}^3 X_i,$$

which, since the underlying probability space is discrete, is a random variable again.

(c) We can also use random variables to describe events: The event F that the second and third die show both 1 *can be written in terms of random variables as*

$$F = \{\omega \in \Omega : X_2(\omega) = X_3(\omega) = 1\},$$

or, using the shorthand notation,

$$F = \{X_2 = X_3 = 1\}.$$

Since we are dealing with a Laplace experiment, its probability is given by

$$\begin{aligned}\mathbb{P}(F) &= \mathbb{P}(\{\omega \in \Omega : X_2(\omega) = X_3(\omega) = 1\}) = \mathbb{P}(\{\omega \in \Omega : \omega_2 = \omega_3 = 1\}) \\ &= |\{\omega \in \Omega : \omega_2 = \omega_3 = 1\}| \cdot 6^{-3} = 6 \cdot 6^{-3} = 1/36.\end{aligned}$$

The slightly more complicated example

$$\mathbb{P}(X_1 \text{ is odd}, X_1 + X_2 + X_3 \text{ is even}) = 1/4$$

is left as an exercise. Here and in the following we implicitly use the notation $\mathbb{P}(F, G) := \mathbb{P}(F \cap G)$.

In our context, there are two principal classes of random variables:

Definition 1.7.3. *(a) A random variable X from an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called a* real random variable ('(reelle) Zufallsgröße').

Here, $\mathcal{B}(\mathbb{R})$ is the Borel-σ-algebra of $\mathbb{R}$*, i.e., it is the smallest σ-algebra over $\mathbb{R}$ that contains all open subsets of $\mathbb{R}$ ($\mathcal{B}(\mathbb{R})$ also contains all singletons $\{x\}$ for $x \in \mathbb{R}$, all closed sets, and all intervals).*

(b) A random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and mapping to $(E, \mathcal{E})$ as in the definition is called a discrete random variable ('diskrete Zufallsvariable') *if $X(\Omega)$ is at most countable, and if $2^{X(\Omega)} \subset \mathcal{E}$.*

Remark 1.7.4. *Note that according to this definition, a random variable can be a discrete random variable and a real random variable at once.*

Lemma 1.7.5. *Let X and Y be real random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function (or let f just be measurable from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which admits significantly more functions). Then $X + Y$, $X - Y$, $X \cdot Y$, X/Y (if the latter is well-defined for all $\omega \in \Omega$) and $f \circ X$ are also real random variables.*
If X is a discrete random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(E, \mathcal{E})$ and $g : E \to \mathbb{R}$ is an arbitrary function, then $g \circ X$ is a discrete real random variable.

We will only prove this lemma in the lecture 'probability theory I', where we will have derived all the tools required for doing so. As already mentioned above, instead of trying to necessarily give probability spaces explicitly, we want to understand the probabilities with which random variables take values in certain 'nice' sets, namely elements of $\mathcal{E}$. I.e., we are interested in probabilities of the form $\mathbb{P}(X^{-1}(A))$, with $A \in \mathcal{E}$.
The following theorem tells us that if X is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and mapping into the measurable space $(E, \mathcal{E})$, then $(E, \mathcal{E}, \mathbb{P} \circ X^{-1})$ is a probability space as well.

Theorem 1.7.6 (Image measure ('Bildmaß') induced by a random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ be a random variable. Then*

$$\begin{aligned}\mathbb{P} \circ X^{-1} : \mathcal{E} &\to [0, 1] \\ A &\mapsto \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A)\end{aligned} \tag{1.7.1}$$

defines a probability measure on $(E, \mathcal{E})$.

Remark 1.7.7. *As is obvious from* (1.7.1)*, the notation $\mathbb{P} \circ X^{-1}$ comes from the fact that one can consider it as the concatenation of the pre-image operator*

$$\begin{aligned}X^{-1} : \mathcal{E} &\to \mathcal{F} \\ A &\mapsto X^{-1}(A),\end{aligned}$$

with the map induced by the probability measure

$$\begin{aligned}\mathbb{P} : \mathcal{F} &\to [0, 1] \\ F &\mapsto \mathbb{P}(F).\end{aligned}$$

Proof of Theorem 1.7.6. Since X is a random variable, we have that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{E}$. Thus, $\mathbb{P} \circ X^{-1}$ is well-defined on $\mathcal{E}$, and it takes values in the interval $[0,1]$. In addition,

$$\mathbb{P} \circ X^{-1}(E) = \mathbb{P}(X^{-1}(E)) = \mathbb{P}(\Omega) = 1.$$

It remains to check the σ-additivity. For this purpose, let (A_n) be a sequence of pairwise disjoint sets with $A_n \in \mathcal{E}$ for all $n \in \mathbb{N}$. We will need the observation that

$$X^{-1}(\dot\cup_n A_n) = \dot\cup_n X^{-1}(A_n).$$

Using this identity, we obtain

$$\mathbb{P} \circ X^{-1}(\cup_n A_n) = \mathbb{P}\big(X^{-1}(\dot\cup_n A_n)\big) = \mathbb{P}(\dot\cup_n X^{-1}(A_n)) = \sum_{n\in\mathbb{N}} \mathbb{P}(X^{-1}(A_n)) = \sum_{n\in\mathbb{N}} \mathbb{P} \circ X^{-1}(A_n),$$

where in the last but one inequality we used the σ-additivity of $\mathbb{P}$ itself. This finishes the proof. □

The above result gives rise to one of the fundamental concepts of probability theory.

Definition 1.7.8. *Two (or more) random variables will be called* identically distributed*, if their distributions coincide*

Definition 1.7.9. $\mathbb{P} \circ X^{-1}$ *as defined in Theorem 1.7.6 is called the* distribution ('Verteilung') of X *or the* law of X*. It is sometimes also written as* $\mathbb{P}_X$*.*

The distribution of X describes the probabilities with which the random variable X takes values in the sets $A \in \mathcal{E}$. Thus, from a probabilistic point of view it contains all the information we need in order to understand the random experiment X is describing (without knowing the actual probability space $(\Omega, \mathcal{F}, \mathbb{P})$). As a consequence, random variables are usually characterized by their distributions, and a plethora of them is so important that they get their own names. We will just have a look at a few of them here. For this purpose we will assume that there is an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which is rich enough to have such random variables defined on it (which you are invited to check as an exercise), but we will not need its specific structure here.
As we have seen before, the concept of independence of events plays an important role in probability theory. Now since we use random variables to describe random experiments, and since in particular we want to have a notion of independent experiments as well, we want to introduce the concept of independent random variables. The easiest way to do so arguably is to reduce it to the notion of independence of σ-algebras. For this purpose, we start with the following.

Definition 1.7.10. *Let $(E, \mathcal{E})$ be a measurable space. Then a subset $\mathcal{D}$ of 2^E is called a* sub-σ-algebra of $\mathcal{E}$ *if*

(a) $\mathcal{D}$ is a σ-algebra, and

(b) $\mathcal{D} \subset \mathcal{E}$.

Having introduced the notion of families of independent events above, it will turn out useful to define the independence of σ-algebras as well. This will prove particularly useful in the context of random variables.

Definition 1.7.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A family of sub-σ-algebras $(\mathcal{F}_\lambda)_{\lambda\in\Lambda}$ of $\mathcal{F}$ is called* independent*, if any family of subsets $(A_\lambda)_{\lambda\in\Lambda}$ with $A_\lambda \in \mathcal{F}_\lambda$ is independent.*

Note that it is essential here that each $\mathcal{F}_\lambda$ is a sub-σ-algebra of $\mathcal{F}$ in order to be able to apply $\mathbb{P}$ to its elements.
The following is a direct consequence of the definition of independence of events from Definition 1.6.1 as well as Definition 1.7.10.

Remark 1.7.12. *A family of sub-σ-algebras $(\mathcal{F}_\lambda)$, $\lambda \in \Lambda$, of $\mathcal{F}$, is an independent family of σ-algebras if and only if for all $n \in \mathbb{N}$, $\lambda_1, \dots, \lambda_n \in \Lambda$ with $\lambda_i \neq \lambda_j$ for all $1 \leqslant i < j \leqslant n$, and $F_{\lambda_k} \in \mathcal{F}_{\lambda_k}$ one has*

$$\mathbb{P}(\cap_{1\leqslant k\leqslant n} F_{\lambda_k}) = \prod_{k=1}^{n} \mathbb{P}(F_{\lambda_k}).$$

The following result is the next step on the way to defining independence of random variables.

Lemma 1.7.13. *Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ be a random variable. Then $X^{-1}(\mathcal{E}) := \{X^{-1}(F) : F \in \mathcal{E}\}$ is a sub-σ-algebra of $\mathcal{F}$.*

Proof. Since X is a random variable we have $X^{-1}(\mathcal{E}) \subset \mathcal{F}$. It remains to show that $X^{-1}(\mathcal{E})$ is a σ-algebra. For this purpose, since $E \in \mathcal{E}$ we clearly have $\Omega = X^{-1}(E) \in X^{-1}(\mathcal{E})$.
Regarding the stability under taking complements, if $A \in X^{-1}(\mathcal{E})$, then there exists $F \in \mathcal{E}$ with $A = X^{-1}(F)$. Now since $\mathcal{E}$ is a σ-algebra, it follows that $F^c \in \mathcal{E}$, and we have $A^c = X^{-1}(F^c)$, thus $A^c \in X^{-1}(\mathcal{E})$.
It remains to show the stability under countable unions. For this purpose let $(A_n)_n$ be a sequence of sets with $A_n \in X^{-1}(\mathcal{E})$ for all $n \in \mathbb{N}$. Thus, there exists a sequence $(F_n)_n$ of sets with $F_n \in \mathcal{E}$ for all $n \in \mathbb{N}$ and $A_n = X^{-1}(F_n)$. Since $\mathcal{E}$ is a σ-algebra we get that $\cup_n F_n \in \mathcal{E}$. Now since $\cup_n A_n = X^{-1}(\cup_n F_n)$ this supplies us with the fact that $\cup_n A_n \in X^{-1}(\mathcal{E})$.
Thus, $X^{-1}(\mathcal{E})$ is a sub-σ-algebra of $\mathcal{F}$. □

Definition 1.7.14. *Let $(E_\lambda, \mathcal{E}_\lambda)_{\lambda\in\Lambda}$ be a family of measurable spaces and let $(X_\lambda)_{\lambda\in\Lambda}$ be a family of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $X_\lambda : \Omega \to E_\lambda$ for all $\lambda \in \Lambda$. The family $(X_\lambda)_{\lambda\in\Lambda}$ is called* independent *if the family of sub-σ-algebras $(X_\lambda^{-1}(\mathcal{E}_\lambda))_{\lambda\in\Lambda}$ of $\mathcal{F}$ is independent.*

Remark 1.7.15.

- *Similarly to Remark 1.7.12, it follows from the definition that such a family of random variables is independent if and only if one has*

$$\mathbb{P}\big(X_{\lambda_k} \in G_{\lambda_k}\, \forall 1 \leqslant k \leqslant n\big) = \prod_{k=1}^{n} \mathbb{P}(X_{\lambda_k} \in G_{\lambda_k}) \tag{1.7.2}$$

for all $n \in \mathbb{N}$, $\lambda_j \in \Lambda$ for all $1 \leqslant j \leqslant n$, $\lambda_i \neq \lambda_j$ for all $1 \leqslant i \neq j \leqslant n$, and $G_{\lambda_j} \in \mathcal{E}_{\lambda_j}$ for all $1 \leqslant j \leqslant n$.

Here, following standard jargon we write $\mathbb{P}\big(X_{\lambda_k} \in G_{\lambda_k}\, \forall 1 \leqslant k \leqslant n\big)$ for

$$\mathbb{P}\big(\cap_{k=1}^n \{X_{\lambda_k} \in G_{\lambda_k}\}\big) = \mathbb{P}\big(\cap_{k=1}^n \underbrace{\{\omega \in \Omega : X_{\lambda_k}(\omega) \in G_{\lambda_k}\}}_{\in\mathcal{F}}\big).$$

- *As in the case of independent events, we also use the notation $X \perp Y$ to write that the random variables X and Y are independent. The reason for this will only become apparent in the lecture 'probability theory I' when we introduce $\mathcal{L}^2$ spaces – in fact, in this case independence in some sense corresponds to orthogonality.*

In the case of discrete random variables, the above definition can be simplified.

Lemma 1.7.16. *In the setting of Definition 1.7.14, assume that the random variables X_λ are all discrete random variables. Then they form an independent family if and only if for all $n \in \mathbb{N}$, all pairwise distinct $\lambda_1, \dots, \lambda_n \in \Lambda$, and all $x_{\lambda_1} \in E_{\lambda_1}$, $x_{\lambda_2} \in E_{\lambda_2}, \dots, x_{\lambda_n} \in E_{\lambda_n}$, one has that*

$$\mathbb{P}(X_{\lambda_1} = x_{\lambda_1}, \dots, X_{\lambda_n} = x_{\lambda_n}) = \prod_{j=1}^{n} \mathbb{P}(X_{\lambda_j} = x_j). \tag{1.7.3}$$

Proof. If the family is independent, choosing $G_{\lambda_j} := \{x_{\lambda_j}\}$ for all $1 \leqslant j \leqslant n$ (note that $\{x_{\lambda_j}\} \in \mathcal{E}_{\lambda_j}$ since by assumption X_{λ_j} is a discrete random variable) and using that the λ_i are pairwise distinct, it follows from (1.7.2) that (1.7.3) holds true.
To prove the reverse direction assume (1.7.3) to hold for all choices of the corresponding parameters, and let arbitrary G_{λ_j}, $1 \leqslant j \leqslant n$, as in Remark 1.7.15 be given. Since each $G_{\lambda_j} \cap X_{\lambda_j}(\Omega)$ is countable, we get that

$$\begin{aligned}
\mathbb{P}\big(X_{\lambda_k} \in G_{\lambda_k}\, \forall 1 \leqslant k \leqslant n\big) &= \mathbb{P}\big(X_{\lambda_k} \in G_{\lambda_k} \cap X_{\lambda_k}(\Omega)\, \forall 1 \leqslant k \leqslant n\big) \\
&\overset{\sigma\text{-additivity}}{=} \sum_{\substack{x_{\lambda_1}\in G_{\lambda_1}\cap X_{\lambda_1}(\Omega),\dots,\\ x_{\lambda_n}\in G_{\lambda_n}\cap X_{\lambda_n}(\Omega)}} \mathbb{P}\big(X_{\lambda_1} = x_{\lambda_1}, \dots, X_{\lambda_n} = x_{\lambda_n}\big) \\
&\overset{(1.7.3)}{=} \sum_{\substack{x_{\lambda_1}\in G_{\lambda_1}\cap X_{\lambda_1}(\Omega),\dots,\\ x_{\lambda_n}\in G_{\lambda_n}\cap X_{\lambda_n}(\Omega)}} \prod_{j=1}^{n} \mathbb{P}\big(X_{\lambda_j} = x_{\lambda_j}\big) = \prod_{j=1}^{n} \Big(\sum_{x_{\lambda_j}\in G_{\lambda_j}\cap X_{\lambda_j}(\Omega)} \mathbb{P}\big(X_{\lambda_j} = x_{\lambda_j}\big)\Big) \\
&= \prod_{j=1}^{n} \mathbb{P}\big(X_{\lambda_j} \in G_{\lambda_j} \cap X_{\lambda_j}(\Omega)\big) = \prod_{j=1}^{n} \mathbb{P}\big(X_{\lambda_j} \in G_{\lambda_j}\big)
\end{aligned}$$

(make sure you also understand the fourth equality sign, which essentially is a version of the distributive law for infinitely many summands). This establishes (1.7.2) and hence finishes the proof. □

Example 1.7.17. *In the context of Examples 1.2.1 and 1.7.2, the random variables X_1 and $X_2 + X_3$ are independent.*
The first thing we verify is that $X_2 + X_3$ is again a random variable. In fact, in Example 1.7.2 we had observed that any function defined on a discrete probability space is a random variable; thus, so is $X_2 + X_3$, and it maps to $\{1, \dots, 12\}$ endowed with the σ-algebra $2^{\{1,\dots,12\}}$. We now have to check the independence. In this setting, see Remark 1.7.15 and observe that the family of random variable for which we want to show independence consists of two elements, X_1 and $X_2 + X_3$, this boils down to showing that for any $A_1 \in 2^{\{1,\dots,6\}}$ and any $A_2 \in 2^{\{1,\dots,12\}}$ we have

$$\mathbb{P}(X_1 \in A_1, X_2 + X_3 \in A_2) = \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 + X_3 \in A_2). \tag{1.7.4}$$

Now we observe that

$$\{X_1 \in A_1\} = F_1 \times \{1, \dots, 6\}^2$$

for $F_1 := A_1 \subset \{1, \dots, 6\}$. Similarly,

$$\{X_2 + X_3 \in A_2\} = \{1, \dots, 6\} \times F_2$$

for some *$F_2 \subset \{1, \dots, 6\}^2$. As a consequence, using the fact that $\mathbb{P}$ was defined as the uniform distribution on Ω, we get*

$$\mathbb{P}(X_1 \in A_1, X_2 + X_3 \in A_2) = \mathbb{P}(F_1 \times F_2) = \frac{|F_1 \times F_2|}{6^3} = \frac{|F_1| \cdot |F_2|}{6^3}.$$

On the other hand we obtain

$$\mathbb{P}(X_1 \in A_1) = \mathbb{P}(F_1 \times \{1, \dots, 6\}^2) = \frac{|F_1|}{6}$$

and similarly

$$\mathbb{P}(X_2 + X_3 \in A_2) = \mathbb{P}(\{1, \dots, 6\} \times F_2) = \frac{|F_2|}{6^2}.$$

This establishes (1.7.4).

Remark 1.7.18. *Sums of independent random variables and convolutions: Oftentimes one is interested in functionals of several, possibly independent, random variables (as exemplified by the above example again, where we considered the sum of the second and third die roll). It immediately arises the question of what we can say about the distribution of the corresponding functional in this case.*
For simplicity we start with considering the case of independent $\mathbb{Z}$-valued random variables and want to investigate the distribution of their sum. For this purpose, let X and Y be independent random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and for notational convenience set $p_X(x) := \mathbb{P}(X = x)$ as well as $p_Y(x) := \mathbb{P}(Y = x)$. Then we get for $n \in \mathbb{Z}$ that

$$\begin{aligned} p_{X+Y}(z) = \mathbb{P}(X + Y = z) &= \sum_{x \in \mathbb{Z}} \mathbb{P}(X = x, Y = z - x) = \sum_{x \in \mathbb{Z}} \mathbb{P}(X = x)\mathbb{P}(Y = z - x) \\ &= \sum_{x \in \mathbb{Z}} p_X(x) p_Y(z - x), \end{aligned}$$

where we used the independence of X and Y (in the form of Lemma 1.7.16) to obtain the penultimate equality. The right-hand side interpreted as a function in z turns out to be so important that it is granted its own name. It is called the (discrete) convolution $p_X * p_Y$ of p_X and p_Y. *Thus, from the above we deduce the equality of functions*

$$p_{X+Y} = p_X * p_Y.$$

Thus, in this case of independent random variables X and Y, the distribution of $X + Y$ only depends on the distributions of X and Y. Furthermore, from the above we also infer that the convolution is a commutative operation:

$$p_X * p_Y = p_Y * p_X.$$

The convolution will play an important role in establishing functional connections between the distributions of sums of independent random variables. We will get back to this identity at a later point in this class.
One reason the convolution is helpful here is the following fact: The Fourier transform of a convolution of two functions equals the product of the convolutions of those functions.
We will get back to this in the exercise classes.

1.8 Specific distributions

The following remark will prove useful in the next sections when we introduce various different distributions: On the one hand, it tells us that for any given probability measure on a measurable space, we can always find a random variable that has this probability measure as its distribution. On the other hand, this also suggests why we will mostly be concerned with distributional properties of random variables, rather than the details of the underlying probability space.

Remark 1.8.1. *Given any distribution μ (i.e., a probability measure on $(E, \mathcal{E})$) one can construct a random variable X with law μ as follows. Indeed, ake $(E, \mathcal{E}, \mu)$ as the underlying probability space and choose X to be the identity on E. Then X defines a random variable from $(E, \mathcal{E}, \mu)$ to $(E, \mathcal{E})$ with law μ.*

1.8.1 Discrete distributions

In order to rephrase the above examples in terms of distributions, we recall the notion of a Dirac measure from Definition 1.3.8. We will call any distribution on a measurable space $(E, \mathcal{E})$ which is of the form

$$\sum_{n\in\mathbb{N}} \alpha_n \delta_{x_n},$$

where $x_n \in E$ and $\alpha_n \geqslant 0$ with $\sum_{n\in\mathbb{N}} \alpha_n = 1$ a *discrete distribution.*

Example 1.8.2. *A random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is called* Bernoulli distributed with parameter $p \in [0, 1]$ *(named after the Swiss mathematician Jacob Bernoulli (1655–1705)) if $X : \Omega \to \{0, 1\}$ as well as*

$$\mathbb{P}(X = 1) = p, \quad \textit{and} \quad \mathbb{P}(X = 0) = 1 - p.$$

In this case one writes $X \sim \mathrm{Ber}_p$ and the law / distribution $\mathbb{P} \circ X^{-1}$ is referred to as the Bernoulli distribution Ber_p *which, using Definition 1.3.8, can be written as*

$$\mathrm{Ber}_p = p\delta_1 + (1 - p)\delta_0.$$

A random variable that is Bernoulli distributed describes a coin flip (biased if $p \neq 1/2$), for example. Assume w.l.o.g. ('o.B.d.A.'; without loss of generality) that the coin shows heads with probability p and tails with probability $1 - p$. If a gambler plays n independent trials with this coin and wins if the coin shows heads, whereas she loses if the coin shows tails, then this can be modeled on the probability space $\Omega = \{0, 1\}^n$, where 1 is identified with heads (or winning, for that matter), and 0 with tails (i.e., losing). We choose $\mathcal{F} := 2^\Omega$. For $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ we define

$$\mathbb{P}(\omega) = p^{\sum_{j=1}^n \omega_j} (1 - p)^{n - \sum_{j=1}^n \omega_j}.$$

Then $X_i : \Omega \to \{0, 1\}$ defined via $X_i(\omega) := \omega_i$ describes the result of the i-th coin flip, and it is not hard to check that the $X_i \sim \mathrm{Ber}_p$, and that the $X_1, \ldots, X_n$ form an independent family.
Assume now that the gambler is interested in the number of coin flips that she has won out of the n trials. For this purpose, introduce the random variable $S_n(\omega) := \sum_{i=1}^n \omega_i$. For $k \in \{0, 1, \ldots, n\}$ there are $\binom{n}{k}$ elements $\omega \in \Omega$ for which one has

$$S_n(\omega) = k,$$

and hence we see that the probability that the gambler has won exactly k out of n coin flips, is given by

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

The distribution obtained in this way directly leads us to the next example.

Example 1.8.3. *A random variable X is called* Binomially distributed with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$, *if*

$$\textit{for all } k \in \{0, 1, \ldots, n\} \textit{ one has } \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{1.8.1}$$

In this case, one writes $X \sim \mathrm{Bin}_{n,p}$ and its distribution is referred to as the Binomial distribution $\mathrm{Bin}_{n,p}$, *which can be written as*

$$\mathrm{Bin}_{n,p} = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} \delta_k.$$

More generally and not necessarily in the context of coin flips, it can be interpreted as describing the number of successes in n independent trials, when the random variables $X_1, \dots, X_n$ describing the trials form an independent family, and each trial has a probability of p to be successful.
We can summarize the above observations in the following claim.

Claim 1.8.4. *The sum*

$$S_n := \sum_{j=1}^{n} X_j$$

of independent random variables $X_1, \dots, X_n$, each distributed according to Ber_p*, is distributed according to* $\mathrm{Bin}_{n,p}$.

The concept of a family of random variables that are independent and all have the same distribution is so important that it has its own name.

Definition 1.8.5. *A family (X_λ), $\lambda \in \Lambda$, is called* independent identically distributed (i.i.d.) ('unabhängig identisch verteilt' (u.i.v.))*, if*

(a) the family (X_λ), $\lambda \in \Lambda$, is an independent family of random variables, and

(b) if the X_λ, $\lambda \in \Lambda$, all have the same distribution.

In addition, the binomial distribution is intimately connected to so-called 'urn models'. For this purpose, assume that you are given an urn containing $N \in \mathbb{N}$ balls, $K \in \{0, 1, \dots, N\}$ of which are white and $N-K$ of which are black. Then the probability that when drawing $n \in \mathbb{N}$ times from this urn in a uniform i.i.d. fashion with replacement ('Ziehen mit Zurücklegen') you have seen exactly $k \in \{0, 1, \dots, n\}$ white balls, is given by

$$\mathrm{Bin}_{n,N/K}(k).$$

Indeed, this follows immediately from Claim 1.8.4, since the draws can be interpreted as an i.i.d. family $X_1, \dots, X_n$ of $\mathrm{Ber}(K/N)$ *distributed variables, where $X_i = 1$ if the i-th draw results in a white ball and $X_i = 0$ if the i-th draw results in a black ball.*

Example 1.8.6. *A random variable X is called* geometrically distributed with success parameter $p \in (0, 1)$*, if*

$$\text{for all } k \in \mathbb{N} \text{ one has } \mathbb{P}(X = k) = p(1-p)^{k-1}. \tag{1.8.2}$$

In this case we write $X \sim \mathrm{Geo}_p$, and its distribution is referred to as the Geometric distribution Geo_p*, which can be written as*

$$\mathrm{Geo}_p = \sum_{k=1}^{\infty} p(1-p)^{k-1} \delta_k.$$

Remark 1.8.7. *Some authors call X geometrically distributed if instead of* (1.8.2)*,*

$$\text{for all } k \in \mathbb{N}_0 \text{ one has } \mathbb{P}(X = k) = p(1-p)^k.$$

Example 1.8.8. *A random variable X is called* Poisson distributed *with parameter $\nu > 0$ if $X : \Omega \to \mathbb{N}_0$ and*

$$\mathbb{P}(X = k) = e^{-\nu} \frac{\nu^k}{k!} \quad \forall k \in \mathbb{N}_0.$$

In this case we write $X \sim \mathrm{Poi}_\nu$, and its distribution is referred to as the Poisson distribution Poi_ν *(named after the French mathematician Siméon Denis Poisson (1781 – 1840)), which can be written as*

$$\mathrm{Poi}_\nu = e^{-\nu} \sum_{k=0}^{\infty} \frac{\nu^k}{k!} \delta_k.$$

Poisson distributed random variables are e.g. used to describe the number of customers that have called a customer service center in a certain time interval. The reason for such a description being feasible is given by the following theorem.

Theorem 1.8.9 (Poisson limit theorem)**.** *Let (p_n) be a sequence of numbers from $[0, 1]$ such that the limit $\nu := \lim_{n\to\infty} np_n$ exists. Then for each $k \in \mathbb{N}_0$,*

$$\lim_{n\to\infty} \mathrm{Bin}_{n,p_n}(k) = \mathrm{Poi}_\nu(k).$$

Proof. For $k \in \mathbb{N}_0$ fixed we have

$$\mathrm{Bin}_{n,p_n}(k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n!}{k!(n-k)!} \frac{(p_n n)^k}{n^k} \Big(1 - \frac{p_n n}{n}\Big)^{n-k} \xrightarrow{n\to\infty} \frac{\nu^k}{k!} e^{-\nu} = \mathrm{Poi}_\nu(k).$$

□

This result gives rise to the fact that the Poisson distribution is used for modelling e.g. the number of customers that contact a call center during a certain time interval. We partition the time interval into n subintervals of equal width, and as we take n to infinity, it is reasonable to assume that in any of the subintervals either zero or one customers are calling. Due to symmetry and independence, it furthermore seems reasonable to assume that the probability of a customer calling in a subinterval has a probability decaying like p/n some $p \in (0,\infty)$ (as $n \to \infty$), and that the fact that a customer has called during one subinterval does not influence the probabilities that a customer is calling during another time interval.[7] Thus, the probability of k customers calling during the original time interval should be reasonably approximated by $\mathrm{Bin}_{n,p/n}(k)$ if n is large. The above Theorem 1.8.9 now shows that the Binomial distribution is the right candidate for this.

Example 1.8.10. *Let $N \in \mathbb{N}$, and $K, n \in \{0,1,\dots,N\}$. A random variable X is called* hypergeometrically distributed with parameters N, tK, n *if $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \{0,1,\dots,N\}$ with*

$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \text{for } k \in \{0 \vee n + K - N, \dots, n \wedge K\}, \tag{1.8.3}$$

and $\mathbb{P}(X = k) = 0$ otherwise.[8]
In this case we write $X \sim \mathrm{Hyp}_{N,K,n}$, and its distribution is referred to as the Hypergeometric distribution $\mathrm{Hyp}_{N,K,n}$ with parameters N, K, and n. *The hypergeometric distribution can be interpreted as an urn model ('Ziehen ohne Zurücklegen'). Assume in total there are N balls in an urn, K of them are white and $N-K$ of them are black. For $k \in \{0,1,\dots,N\}$, the quantity $\mathrm{Hyp}_{N,K,n}(k)$ then gives the probability that in a uniformly random draw of n balls without replacement ('Ziehen ohne Zurücklegen') out of the urn, there are exactly k white balls and $n-k$ black balls.*[9] *Indeed, there are altogether $\binom{N}{n}$ possibilities to draw n balls (without replacement) out of an urn that contains N balls. Since there are $\binom{K}{k}$ possibilities to choose k white balls out of the K white balls present in the urn and $\binom{N-K}{n-k}$ possibilities to choose $n-k$ black balls out of the all in all $N-K$ black balls in the urn, there are in total $\binom{K}{k} \cdot \binom{N-K}{n-k}$ favourable outcomes to our drawing procedure. Since we were supposing that the drawing mechanism was uniformly random, we infer that the probability we are after is given by*

$$\frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}.$$

Example 1.8.11. *We give some further distributions which occur frequently:*

(a) A random variable X is called Rademacher distributed *(named after the German-American mathematician Hans Rademacher) if $X : \Omega \to \{0,1\}$ as well as*

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}.$$

The corresponding distribution $\mathbb{P} \circ X$ is called Rademacher distribution*, named after the German-American mathematician Hans Rademacher. To us, it will mostly be important as the distribution of the increments of simple random walk to be introduced below.*

(b) As a generalization to the Binomial distribution introduced in Example 1.8.3, we consider the multinomial distribution. *For $r, n \in \mathbb{N}$ and $p_i \in (0,1)$, $1 \leqslant i \leqslant r$, such that $\sum_{i=1}^r p_i = 1$, we call an $\mathbb{N}_0^r$-valued random variable X* multinomially distributed with parameters $n, p_1, \dots, p_r$ *if*

$$\mathbb{P}(X = (x_1,\dots,x_r)) = n! \frac{\prod_{i=1}^r p_i^{x_i}}{\prod_{i=1}^r x_i!}, \quad \forall (x_1,\dots,x_r) \in \mathbb{N}_0^r \text{ with } \sum_{i=1}^r x_i = n, \tag{1.8.4}$$

[7] These are slightly delicate issues; in fact, if the customer center in question is e.g. that of an energy retailer and there is a power outage during some part of the time interval we consider, then these assumptions will generally not be met. However, they seem reasonable to assume during normal operation.

[8] We use the notation $a \vee b := \max(a,b)$ and $a \wedge b := \min(a,b)$ which is quite common in probability theory.

[9] This also explains the restriction on k given in (1.8.3): In a sample of size n there cannot be more than $n \wedge K$ white balls; at the same time, there have to be at least $n - (N-K) = n + K - N$ white balls in the sample, since at most $N-K$ balls are black.

and $\mathbb{P}(X = (x_1, \dots, x_r)) = 0$ *otherwise. We write* $\mathrm{Mult}_{n,p_1,\dots,p_r}$ *for the respective distribution.*

The multinomial distribution is named this way due to the multinomial coefficient, which for $n \in \mathbb{N}$ *as well as* $x_1, \dots, x_r \in \mathbb{N}_0$ *with* $\sum_{i=1}^r x_i = n$*, is defined as*

$$\binom{n}{x_1, x_2, \dots, x_r} := \frac{n!}{x_1! \cdot x_2! \cdot \ldots \cdot x_r!}.$$

Thus, the probability in (1.8.4) *can alternatively be expressed as*

$$\binom{n}{x_1, x_2, \dots, x_r} \prod_{i=1}^r p_i^{x_i}.$$

In the case $r = 2$*, the multinomial distribution boils down to the binomial distribution, in the sense that the first coordinate of a* $\mathrm{Mult}_{n,p,1-p}$*-distributed random variable is a* $\mathrm{Bin}_{n,p}$*-distributed random variable. In general, it can also be interpreted as the* n*-fold independent repetition of an experiment which has* r *different possible outcomes, each of them realized with probability* p_i*,* $1 \leqslant i \leqslant r$*, at each trial.*

In terms of an urn model, if we have an urn with K_i *balls of color* i*,* $1 \leqslant i \leqslant r$*, then with* $N := \sum_{i=1}^r K_i$ *and* $n \in \mathbb{N}$*, a* $\mathrm{Mult}_{n,\frac{K_1}{N},\dots,\frac{K_r}{N}}$*-distributed random variable* X *describes the number accumulated outcome of* n *independent draws with replacement ('Ziehen mit Zurücklegen'):* X_i *is the number of times a ball of color* i*,* $1 \leqslant i \leqslant r$*, has been drawn among the* n *trials.*

1.8.2 Distribution functions

For real-valued random variables (which we sometimes also refer to as 'real random variables') the concept of its (cumulative) distribution function plays a prominent role.

Definition 1.8.12. *Let* X *be a real random variable defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. Then the function*

$$\begin{aligned} F_X : \mathbb{R} &\to [0,1], \\ t &\mapsto \mathbb{P}(X \leqslant t), \end{aligned}$$

is called the (cumulative) distribution function (or cdf) of X ('Verteilungsfunktion von X').

It is apparent from the definition that the distribution function of X depends on X only through its distribution $\mathbb{P}_X$. The main reason why distribution functions are important in this context is that they characterize probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, cf. Theorem 1.8.18 below.

Remark 1.8.13. *Note that the cdf is well defined since* $\{X \leqslant x\} = \{\omega \in \Omega : X(\omega) \in (-\infty, x]\} \in \mathcal{F}$ *due to the fact that* X *is a real-valued random variable and for all* $x \in \mathbb{R}$*,* $(-\infty, x] \in \mathcal{B}(\mathbb{R})$*.*

Example 1.8.14. *Let* $X \sim \mathrm{Geo}_p$*. Then the distribution function of* X *is given by*

$$F_X(t) = \begin{cases} 0, & \text{if } t < 1, \\ \sum_{j=1}^{\lfloor t \rfloor} p(1-p)^{j-1} = p\frac{1-(1-p)^{\lfloor t \rfloor}}{1-(1-p)} = 1 - (1-p)^{\lfloor t \rfloor}, & \text{if } t \geqslant 1. \end{cases}$$

Exercise 1.8.15. *If* X *is a discrete real random variable, then* F_X *has jumps exactly at the points in* $\{x \in X(\Omega) : \mathbb{P}(X^{-1}\{x\}) > 0\}$ *and is piecewise constant otherwise.*

Theorem 1.8.16. *If* X *is a real random variable, then its distribution function* F_X *has the following properties:*

(a) F_X *is non-decreasing;*

(b)

$$\lim_{t \to -\infty} F_X(t) = 0, \quad \lim_{t \to \infty} F_X(t) = 1;$$

(c) F_X *is right-continuous (i.e., for all* $t_0 \in \mathbb{R}$ *one has* $F_X(t_0) = \lim_{t \downarrow t_0} F_X(t)$*);*

Proof. If X is a real random variable, denote the underlying probability space by $(\Omega, \mathcal{F}, \mathbb{P})$ and the distribution function of X by F_X. Then $F_X(t) = \mathbb{P}(X \leqslant t) \in [0,1]$. In addition, using the monotonicity property of Proposition 1.3.9, we get for $h > 0$ that

$$F_X(t) = \mathbb{P}(X \leqslant t) \leqslant \mathbb{P}(X \leqslant t + h) = F_X(t + h),$$

hence F_X is non-decreasing.
To establish the second property, we observe that since $X(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, we get $\{X \leqslant t\} \downarrow \varnothing$ as $t \to -\infty$. Thus, the continuity from above for $\mathbb{P}$ (see Proposition 1.3.9) implies

$$F_X(t) = \mathbb{P}(X \leqslant t) \overset{t \to -\infty}{\longrightarrow} \mathbb{P}(\varnothing) = 0.$$

Similarly, we get that $\{X \leqslant t\} \uparrow \Omega$ as $t \to \infty$ and hence the continuity of $\mathbb{P}$ from below (see Proposition 1.3.9) implies

$$F_X(t) = \mathbb{P}(\{X \leqslant t\}) \overset{t \to \infty}{\longrightarrow} \mathbb{P}(\Omega) = 1.$$

Altogether, this proves the second point.
It remains to prove the right-continuity. Thus, let $t_0 \in \mathbb{R}$ be given. Note that $\{X \leqslant t + h\} \downarrow \{X \leqslant t\}$ as $h \downarrow 0$, so in particular for a any sequence $(h_n)_{n \in \mathbb{N}}$ with $h_n > 0$ and $h_n \to 0$ as $n \to \infty$. Combining this with continuity from above for $\mathbb{P}$, we deduce

$$F_X(t + h_n) = \mathbb{P}(X \leqslant t + h_n) \overset{n \downarrow 0}{\longrightarrow} \mathbb{P}(X \leqslant t) = F_X(t).$$

Since $t_0 \in \mathbb{R}$ was chosen arbitrarily, this proves the right-continuity of F_X. □

This leads us to the following definition.

Definition 1.8.17. *Any function $F : \mathbb{R} \to [0,1]$ that satisfies the three properties given in Theorem 1.8.16 is called a* distribution function (on $\mathbb{R}$; 'Verteilungsfunktion').

The following result complements Theorem 1.8.16, and combined they establish that there is a correspondence between random variables and distribution functions.

Theorem 1.8.18. *There exists a one-to-one correspondence between probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and distribution functions on $\mathbb{R}$ in the following sense: For $\mathbb{P}$ a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we consider the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ and the random variable id_R, which is the identity function on $\mathbb{R}$; this gives rise to the distribution function F_{id}. Then the mapping which maps from the set of probability measures on $(\mathbb{R}, \mathcal{B}()\mathbb{R}))$ to the set of distribution functions, and which is defined via taking $\mathbb{P}$ to the distribution function F_{id}, is bijective.*

We will only be able to prove this result in 'Probability theory I', but it will already be helpful in dealing with distributions with densities.
As an immediate consequence, we obtain the following corollary.

Corollary 1.8.19. *For each distribution function F there exists a random variable X, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and mapping to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $F = F_X$.*

Proof. Let a distribution function F be given. By Theorem 1.8.18 there exists a probability measure $\mathbb{P}$ such that considering the identity id on $\mathbb{R}$, we have $F_{\mathrm{id}} = F$. □

Given a distribution function with positive density, you will be asked to explicitly construct a random variable X with the properties named in the previous corollary in Exercise sheet number five.

1.8.3 Distributions with densities

We have already seen in Example 1.4.7 that it can be useful to have a concept of distributions which are not discrete but take values in a continuum such as $\mathbb{R}$. Theorem 1.8.16 suggests that it might indeed be possible to have such distributions.

Definition 1.8.20. *A function $f : \mathbb{R} \to [0, \infty)$ which is Riemann-integrable on every bounded interval, and for which*

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \quad \textit{(Riemann integral)},$$

is called a probability density ('Wahrscheinlichkeitsdichte').

The motivation for suggesting distributions induced by integrals over probability densities is the following.

Exercise 1.8.21. *(a) Let f be a probability density. Show that the function*

$$F:\ \mathbb{R}\to\mathbb{R}$$
$$t\mapsto\int_{-\infty}^{t} f(x)\,\mathrm{d}x$$

defines a distribution function. In this case, we refer to f as the density of F.

(b) If F has a density, then F is a continuous function. However, not every distribution function F which is continuous has a density.

The result of this exercise in combination with Theorem 1.8.16 is good news since it tells us that there is a whole new class of probability distributions obtained by integrating over probability densities. Indeed, if X is a random variable the distribution function of which is given by

$$F_X(t)=\int_{-\infty}^{t} f(x)\,\mathrm{d}x,$$

then by definition we have

$$\mathbb{P}(X\in(-\infty,t])=\mathbb{P}(X\leqslant t)=\int_{-\infty}^{t} f(x)\,\mathrm{d}x.$$

From this we get for $a\leqslant b$ that

$$\mathbb{P}(X\in(a,b])=\mathbb{P}(X\in(-\infty,b]))-\mathbb{P}(X\in(-\infty,a]))=\int_{a}^{b} f(x)\,\mathrm{d}x.$$

Extending this identity, we can deduce that for any finite union of (say left open right closed) intervals $I\subset\mathbb{R}$ we have

$$\mathbb{P}(X\in I)=\int_{I} f(x)\,\mathrm{d}x,$$

with I the domain of integration. Motivated by this equality we would like to define the distribution on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ by having this identity not only for finite unions of intervals; in fact, we would like that for all $B\in\mathcal{B}(\mathbb{R})$ one has

$$\mathbb{P}(X\in B)=\int_{B} f(x)\,\mathrm{d}x,$$

i.e., the function

$$\mathcal{B}(\mathbb{R})\ni B\mapsto\int_{B} f(x)\,\mathrm{d}x$$

should define a probability measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$.[10]
There is, however, a significant problem inherent to this approach: Our notion of integral, i.e., the Riemann integral, is not powerful enough to develop a fully-fledged theory of probability measures on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ induced by probability densities (which in some sense we already swept under the rug by omitting the more difficult part of the proof of Theorem 1.8.16). Indeed, we already get into trouble by requiring the σ-additivity. For this purpose, consider the probability density $\frac{1}{2\pi}\mathbb{1}_{[0,2\pi)}$ from Example 1.4.7 where a needle was thrown, and assume we are interested in the probability of the event that the angle the needle encloses with the x-axis is rational. We have for any $q\in\mathbb{Q}\cap[0,2\pi)$ (note that $\{q\}\in\mathcal{B}(\mathbb{R})$) we would certainly want to be able to associate a probability with the set $\{q\}$) that

$$\int_{q}^{q}\mathbb{1}_{[0,2\pi)}(x)\,\mathrm{d}x=0, \tag{1.8.5}$$

i.e., the probability of seeing a specific (rational) angle should be 0. Since $\mathcal{B}(\mathbb{R})$ is stable under countable unions, we would also want to be able to associate a probability to $\cup_{q\in\mathbb{Q}\cap[0,2\pi)}\in\mathcal{B}([0,2\pi))$. From (1.8.5)

[10] It might be useful to note here that in the case of discrete random variables we never had any issues when dealing with the distribution on the range of the random variable; this was the case since the distribution was completely determined by the probabilities of the type $\mathbb{P}(X=x)$, $x\in E$.

in combination with the postulated σ-additivity for probability measures we would want to get that this probability is 0. However, using the Riemann integral we cannot even integrate $\mathbb{1}_{[0,2\pi)}$ over the set $\cup_{q\in\mathbb{Q}\cap[0,2\pi)}$ since the upper and lower sums converge to different limits, namely 2π and 0.
As a consequence, in this introductory course we will not be able to give a theory unifying discrete random variables and random variables whose distribution has a density; instead, the concepts we develop will usually have to be developed separately for both types of random variables separately.
It will only be the case in the sequel class 'Probability Theory I' when we introduce the concept of the 'Lebesgue integral' that we will have a comprehensive theory at our disposal for treating not only the above two types of random variables, but even more general ones at once.[11]
We will now fix some notation.

Definition 1.8.22. *(a) If μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with*

$$\mu((-\infty, t]) = \int_{\infty}^{t} f(x)\,\mathrm{d}x \quad \forall t \in \mathbb{R}, \tag{1.8.6}$$

for some probability density f, then we say that μ has density *f.*

Also, in passing we note that a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which fulfills (1.8.6) *is already characterized this way. We will, however, not prove this result in this introductory class.*

(b) If X is a real random variable such that F_X is a distribution function with density f, then X is called a continuous random variable with density *f, or shortly, a* continuous random variable *or a* random variable with a density*.*

Remark 1.8.23. *The above terminology is slightly unfortunate since a continuous random variables X with a density is not necessarily continuous as a function from Ω to $\mathbb{R}$. In fact, the notion of continuity is not even well-defined is Ω has no metric or topology.*

Remark 1.8.24. *(a) If f is a probability density that defining a probability measure μ via the identity* (1.8.6)*, and if $\widetilde{f}$ is a function such that $\{x \in \mathbb{R} : f(x) \neq \widetilde{f}(x)\}$ is finite, then $\widetilde{f}$ is a probability density as well, and it gives rise to the same measure μ via*

$$\mu((-\infty, t]) = \int_{-\infty}^{t} \widetilde{f}(x)\,\mathrm{d}x.$$

(b) If f is a probability density which is continuous in $x_0 \in \mathbb{R}$, then the corresponding distribution function

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

is differentiable in x_0 with $F'(x_0) = f(x_0)$ according to the Fundamental Theorem of Calculus ('Hauptsatz der Differential- und Integralrechnung').

Example 1.8.25. *(a) For $a, b \in \mathbb{R}$ with $a < b$ the* uniform distribution ('Gleichverteilung') on the interval $[a, b]$ *has the density*

$$\mathbb{R} \ni x \mapsto \frac{1}{b-a}\mathbb{1}_{[a,b]}(x).$$

We write $\mathrm{Uni}([a, b])$ *for the uniform distribution on the interval $[a, b]$, and the corresponding distribution function is given by*

$$F(t) = \begin{cases} 0, & \text{if } t \leqslant a, \\ \frac{t-a}{b-a}, & \text{if } t \in (a, b), \\ 1, & \text{if } t \geqslant b. \end{cases}$$

(b) Let $\kappa > 0$. The exponential distribution ('Exponentialverteilung') with parameter κ *has density*

$$\mathbb{R} \ni x \mapsto \begin{cases} \kappa e^{-\kappa x}, & \text{if } x \geqslant 0, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \mathrm{Exp}(\kappa)$ if X is a random variable that is exponentially distributed with parameter $\kappa > 0$.

[11] Most of the concepts such as expectation, variance, etc. that we will develop below will only be introduced for these two types of random variables. However, as we will see in the sequel 'Probability Theory I', these concepts can be introduced in a fairly general manner using the theory of Lebesgue integration.

(c) The normal or Gaussian distribution ('Normalverteilung' or 'Gaußverteilung', named after the German mathematician Carl Friedrich Gauss (1777–1855)) with parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ *has the density*

$$\mathbb{R} \ni x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$ if X is a random variable that is normally distributed with parameters μ and σ^2.

It should also be noted here that the cumulative distribution function of the standard Normal distribution $\mathcal{N}(0,1)$ is usually denoted by

$$\Phi(t) := \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x, \tag{1.8.7}$$

and that there is no closed expression for general values of t for the right-hand side. There are, however, tables to look up those values for a variety of different values for t.

We will get back to those distributions after having introduced the concept of expectation.

Remark 1.8.26. *As in some way a generalization to Section 1.6.1, one can show that to an arbitrary finite family (F_λ), $\lambda \in \Lambda$, of distribution functions, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a finite family (X_λ), $\lambda \in \Lambda$, of random variables defined on it such that the family is independent and for each $\lambda \in \Lambda$, we have $X_\lambda \sim F_\lambda$. We will not prove this result in this class, since it requires stronger tools than we have at our disposal, but you may use the result without further mentioning for the rest of this class.*

1.9 Expectation

There are certain quantities associated to random variables (or their distributions for that matter) that play a key role. Arguably the most important one is their expectation, if it exists.

Definition 1.9.1. *Let X be a discrete real random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(E, \mathcal{E})$ with $E \subset \mathbb{R}$. We say that it has* finite expectation *if*

$$\sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}(X = x) < \infty. \tag{1.9.1}$$

In this case its expectation ('Erwartungswert') *(or sometimes called* mean*) is defined as*

$$\mathbb{E}[X] := \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x). = \sum_{x \in E} x \cdot \mathbb{P}_X(x). \tag{1.9.2}$$

If, on the other hand, X is a continuous real random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with density ϱ, we say that it has finite expectation *if*

$$\int_{-\infty}^{\infty} |x| \cdot \varrho(x) \, \mathrm{d}x < \infty. \tag{1.9.3}$$

In this case the quantity

$$\int_{-\infty}^{\infty} x \cdot \varrho(x) \, \mathrm{d}x$$

is well-defined as a (finite) real number, and it is called the expectation ('Erwartungswert') *(or sometimes called* mean*) of X, and also denoted by $\mathbb{E}[X]$. Indeed, in this setting, by the improper Riemann integral can be defined via*

$$\int_{-\infty}^{\infty} x \cdot \varrho(x) \, \mathrm{d}x = \int_{-\infty}^{c} x \cdot \varrho(x) \, \mathrm{d}x + \int_{c}^{\infty} x \cdot \varrho(x) \, \mathrm{d}x,$$

where $c \in \mathbb{R}$ is arbitrary.
Random variables with finite expectation are also called integrable random variables ('integrierbare Zufallsvariablen'). *A random variable whose expectation equals 0 is called* centred ('zentriert').[12]

[12] The reason for this is clear for continuous random variables, and it will become clear for general random variables once one has the notion of Lebesgue integration.

Remark 1.9.2. *(a) Given the definition of expectation in* (1.9.2)*, at a first glance it might seem slightly unnatural to require the summability of the absolute value* $|x|$ *in* (1.9.1) *instead of just summability of* x *in order to talk about 'finite expectation'. The reason for this is that this absolute summability condition ensures that one will always obtain the same expectation in* (1.9.2)*, independent of the order in which the summation is performed.*

Note also that if X *has finite expectation, then the left-hand side of* (1.9.1) *equals* $\mathbb{E}[|X|]$.

(b) The rightmost sum of (1.9.2) *provides us with an interpretation of the expectation in terms of physics: If we consider the distribution of a unit mass on* $\mathbb{R}$ *induced by* $\mathbb{P} \circ X^{-1}$ *(in the sense that any point* $x \in X(\Omega) \subset \mathbb{R}$ *is given mass* $\mathbb{P} \circ X^{-1}(x)$*), then the expectation* $\mathbb{E}[X]$ *is the center of gravity of this distribution of mass.*

(c) Just to be on the safe side we will always consider random variables to be either discrete or continuous in the following.

(d) In what follows, if e.g. we write $\mathbb{E}[X]$ *or* $\mathbb{E}[f(X)]$ *for some function* f*, then we always tacitly assume that* X *and* $f(X)$ *are either discrete random variables or continuous random variables with a density, so that we are at least in the position to check whether the respective expectations make sense.*

Remark 1.9.3. *(a) From this definition it follows that the expectation of a random variable* X *only depends on the random variable* X *through its distribution* $\mathbb{P} \circ X^{-1}$*. For the case of discrete random variables this is immediate from the definition, while in the case of continuous random variables it is most convenient to retreat to Lebesgue integration, which we will not do in this course.*

(b) The above distinction in the definition of expectations for discrete random variables on the one hand and random variables with a density on the other hand seems slightly artificial. It will in fact turn out that using the Lebesgue integral we can define the expectation for general real random variables, in particular comprising the two special cases of discrete random variables and random variables with a density.

Definition 1.9.4. *(a) For real random variables we usually denote by*

$$X^+ := X \vee 0$$

its positive part and by

$$X^- := -(X \wedge 0)$$

its negative part, which then implies the equality

$$X = X^+ - X^-.$$

(b) For a random variable X *on* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an event* $A \subset \mathcal{F}$ *we define*

$$\mathbb{E}[X;\, A] := \mathbb{E}[X \cdot \mathbb{1}_A],$$

if the expectation on the right-hand side exists.

Definition 1.9.5. *We denote by* $\mathcal{L}^1 := \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ *the set of all discrete or continuous real random variables defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *with finite expectation.*
If at most one of the expectations $\mathbb{E}[X^+]$ *and* $\mathbb{E}[X^-]$ *is infinite, then* X *is called 'quasi-integrable' and one can still define its (infinite) expectation as* $\mathbb{E}[X] \in \{-\infty, \infty\}$ *as before. It will follow from Proposition 1.9.7 below that if* X *has finite expectation, then*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

This still holds true if X *is only quasi-integrable, and in this case exactly one of the expectations on the right-hand side is infinite.*

In particular, when considering sums we will still call them well-defined as long as no expression of the type $\infty - \infty$ occurs.

Example 1.9.6. *(a) In the setting of Examples 1.2.1 and 1.7.2, we get*

$$\mathbb{E}[X_1] = \sum_{x \in X_1(\Omega)} x \cdot \mathbb{P}(X_1 = x) = \sum_{j=1}^{6} j \frac{1}{6} = \frac{21}{6} = 3.5$$

This is of course what we would have expected for a fair die.

(b) Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} &\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x \overset{x \mapsto x+\mu}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x+\mu) e^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x \\ &= \mu + \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x}_{=0 \text{ due to symmetry}} = \mu, \end{aligned}$$

where we took advantage of the fact that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \,\mathrm{d}x = 1$$

to take μ out of the integral. Thus, X has finite expectation, namely μ. Thus, we observe that the parameter μ is in fact the expectation of an $\mathcal{N}(\mu, \sigma^2)$-distributed random variable. We will soon investigate the role of σ^2.

(c) Let $X \sim \mathrm{Exp}(\kappa)$ some $\kappa \in (0, \infty)$. Using integration by parts we get

$$\int_{-\infty}^{\infty} \mathbb{1}_{x \in [0,\infty)}(x) x \varrho(x) \,\mathrm{d}x = \int_0^{\infty} x \kappa e^{-\kappa x} \,\mathrm{d}x = -x e^{-\kappa x}|_{x=0}^{\infty} - \int_0^{\infty} -e^{-\kappa x} \,\mathrm{d}x = \frac{1}{\kappa}$$

(d) Let $X \sim \mathrm{Poi}(\nu)$ some $\nu > 0$. Then

$$\mathbb{E}[X] = \sum_{n \in X(\Omega)} n \cdot \mathbb{P}(X = n) = \sum_{n \in \mathbb{N}_0} n \cdot e^{-\nu} \frac{\nu^n}{n!} = \nu \sum_{n \in \mathbb{N}} e^{-\nu} \frac{\nu^{n-1}}{(n-1)!} = \nu.$$

Proposition 1.9.7 (Properties of expectations)**.** *Let $X, Y \in \mathcal{L}^1$ and $c \in \mathbb{R}$.*

(a) If $X \leqslant Y$ (i.e., $X(\omega) \leqslant Y(\omega)$ for all $\omega \in \Omega$), then

$$\mathbb{E}[X] \leqslant \mathbb{E}[Y] \quad \textit{(monotonicity of expectation)};$$

(b) $cX + Y \in \mathcal{L}^1$, and

$$\mathbb{E}[cX + Y] = c\mathbb{E}[X] + \mathbb{E}[Y] \quad \textit{(linearity of expectation)}; \tag{1.9.4}$$

(c) if in addition X and Y are independent, then $XY \in \mathcal{L}^1$ and

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{1.9.5}$$

Remark 1.9.8. *It is sometimes helpful to have the above result not only when $X, Y \in \mathcal{L}^1$, but also if some of the respective random variables are only quasi-integrable. Thus, you might want to check that*

- *(a) holds true if X and Y are only quasi-integrable;*
- (1.9.4) *holds true also if either at most one of the random variables is quasi-integrable, or otherwise if $c \geqslant 0$ and both random variables are quasi-integrable with*
  - *(a) $\mathbb{E}[X^+] = \mathbb{E}[Y^+] = \infty$, or*
  - *(b) $\mathbb{E}[X^-] = \mathbb{E}[Y^-] = \infty$.*
- (1.9.5) *does not generally hold true if we require one of the random variables X and Y to be quasi-integrable only.*

Proof. We will give the proof for discrete random variables. The proof in the case when at least one of the two random variables is continuous can e.g. be done by approximating continuous random variables by discrete ones.

(a) If $X, Y \in \mathcal{L}^1$ we get

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}(X = x) = \sum_{x \in X(\Omega), y \in Y(\Omega)} x \cdot \mathbb{P}(X = x, Y = y),$$

where the last equality takes advantage of the fact that X has finite expectation. Now since $X \leqslant Y$ we get that $\{X = x, Y = y\}$ can only be non-empty if $y \geqslant x$. Thus, we can upper bound the right-hand side of the previous display by

$$\sum_{x \in X(\Omega), y \in Y(\Omega)} y \cdot \mathbb{P}(X = x, Y = y) = \sum_{y \in X(\Omega)} y \cdot \mathbb{P}(X = y) = \mathbb{E}[Y],$$

where now in the first equality we used the fact that Y has finite expectation.

(b) For $c = 0$ we have $cX = 0$ which has finite expectation 0. Thus, without loss of generality assume $c \neq 0$. We have

$$\sum_{x \in (cX)(\Omega)} |x| \cdot \mathbb{P}(cX = x) = \sum_{x \in X(\Omega)} |cx| \cdot \mathbb{P}(cX = cx) = |c| \cdot \sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}(X = x) < \infty,$$

where the last inequality follows since X has finite expectation. Now since the above sums converge absolutely, we get that the equations hold true without the absolute value signs as well, the left-hand side then equals $\mathbb{E}[cX]$ and the right-hand side equals $c\mathbb{E}[X]$.

Thus, it remains to show that $X + Y$ has finite expectation and that $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. Using that X and Y have finite expectations we get

$$\begin{aligned}
&\sum_{z \in (X+Y)(\Omega)} |z| \cdot \mathbb{P}(X + Y = z) = \sum_{x \in X(\Omega)} \sum_{z - x \in Y(\Omega)} |z| \cdot \mathbb{P}(X = x, Y = z - x) \\
&\leqslant \sum_{x \in X(\Omega)} \sum_{z - x \in Y(\Omega)} (|x| + |z - x|) \cdot \mathbb{P}(X = x, Y = z - x) \\
&= \sum_{x \in X(\Omega)} \sum_{\widetilde{y} \in Y(\Omega)} |x| \cdot \mathbb{P}(X = x, Y = \widetilde{y}) + \sum_{x \in X(\Omega)} \sum_{\widetilde{y} \in Y(\Omega)} |\widetilde{y}| \cdot \mathbb{P}(X = x, Y = \widetilde{y}) \\
&= \sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}(X = x) + \sum_{\widetilde{y} \in Y(\Omega)} |\widetilde{y}| \cdot \mathbb{P}(Y = \widetilde{y}) < \infty.
\end{aligned}$$

In particular, all sums in the above chain are absolutely convergent. Hence, we can omit the absolute values in those computations, in which case the inequality turns into an equality and we deduce

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

from the equations.

(c) We compute similarly to the previous part that

$$\begin{aligned}
&\sum_{z \in (XY)(\Omega)} |z| \cdot \mathbb{P}(XY = z) = \sum_{y \in Y(\Omega), y \neq 0} \sum_{\frac{z}{y} \in X(\Omega)} |y| \cdot |z/y| \underbrace{\mathbb{P}(X = z/y, Y = y)}_{= \mathbb{P}(X = z/y)\mathbb{P}(Y = y) \text{ (independence)}} \\
&\overset{x := z/y}{=} \sum_{x \in X(\Omega)} |x| \cdot \mathbb{P}(X = x) \sum_{y \in Y(\Omega)} |y| \cdot \mathbb{P}(Y = y),
\end{aligned}$$

and the right-hand side is finite by assumption. In particular, this implies $XY \in \mathcal{L}^1$. Again, due to the absolute summability of all occurring sequences, we can redo the above without absolute value signs to get

$$\sum_{z \in (XY)(\Omega)} z \cdot \mathbb{P}(XY = z) = \sum_{z \in X(\Omega)} z \cdot \mathbb{P}(X = z) \sum_{y \in Y(\Omega)} y \cdot \mathbb{P}(Y = y),$$

which due to $X, Y, XY \in \mathcal{L}^1$ amounts to (1.9.5).

□

Example 1.9.9. *We had already seen in Claim 1.8.4 that the sum of n independent random variables $X_1, \dots, X_n$, each distributed according to Ber_p, $p \in [0,1]$, has distribution $\mathrm{Bin}_{n,p}$. Thus, if $Y \sim \mathrm{Bin}_{n,p}$, we get, using the fact that the expectation of a random variable depends only on its distribution,*

$$\mathbb{E}[Y] = \mathbb{E}\Big[\sum_{j=1}^{n} X_j\Big] = \sum_{j=1}^{n} \mathbb{E}[X_j] = np,$$

where in the second equality we used the linearity of expectation.
Thus, the expectation of a $\mathrm{Bin}_{n,p/n}$-distributed random variable equals p. This is in accordance with the Poisson limit theorem, i.e., Theorem 1.8.9, which states that the $\mathrm{Bin}_{n,p/n}$ distribution converges to Poi_p in some *sense, since according to Example 1.9.6 (d) the expectation of a Poi_p distributed random variable is also p.*

It will often be the case that we will want to compute the expectation not only of a real random variable X, but of certain functionals such as $f(X) = X^2$. Lemma 1.7.5 tells us that if either X is a real discrete random variable or otherwise if f is real-valued continuous, then $f(X)$ is a real random variable again. Thus, at least under the assumption that $f(X)$ is discrete or that is again has a density, it makes sense to check whether $f(X)$ has finite expectation and if so, to compute it.
Now oftentimes the distribution of X, is well known, i.e., the values $\mathbb{P}(X = x)$ are easy to obtain, whereas the ones for $\mathbb{P}(f(X) = x)$ are harder to get. On the other hand, they are a priori needed to compute

$$\mathbb{E}[f(X)] = \sum_{x \in f(X(\Omega))} x \cdot \mathbb{P}(f(X) = x),$$

if it exists. The following result tells us how to compute $\mathbb{E}[f(X)]$ using the original distribution of X instead of that of $f(X)$.

Proposition 1.9.10. *[Change of variable formula ('Transformationssatz')]*

(a) Let X be a discrete random variable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(E, \mathcal{E})$ and let $f : E \to \mathbb{R}$ be an arbitrary function. Then, $f(X)$ has finite expectation if and only if

$$\sum_{x \in X(\Omega)} |f(x)| \cdot \mathbb{P}(X = x) < \infty,$$

and in this case

$$\mathbb{E}[f(X)] = \sum_{x \in X(\Omega)} f(x) \cdot \mathbb{P}(X = x). \tag{1.9.6}$$

(b) If X is a continuous random variable with density ϱ such that $f(X)$ also is a continuous random variable, then $f(X) \in \mathcal{L}^1$ if and only if

$$\int_{-\infty}^{\infty} |f(x)| \cdot \varrho(x)\,\mathrm{dx} < \infty, \tag{1.9.7}$$

and in this case the expectation is finite and given by

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x) \cdot \varrho(x)\,\mathrm{dx}.$$

Proof. We will give the proof of the first part only.

$$\begin{aligned}\mathbb{E}[|f(X)|] &= \sum_{x \in f(X(\Omega))} |x| \cdot \mathbb{P}(f(X) = x) = \sum_{x \in f(X(\Omega))} \sum_{y \in X(\Omega)\,:\, f(y) = x} |f(y)| \cdot \mathbb{P}(X = y) \\ &= \sum_{y \in X(\Omega)} |f(y)| \cdot \mathbb{P}(X = y).\end{aligned}$$

Thus, $f(X)$ has finite expectation if and only if $\sum_{y \in X(\Omega)} |f(y)| \cdot \mathbb{P}(X = y) < \infty$, and since in this case all sums are absolutely convergent, we can omit the absolute value signs and obtain (1.9.6). □

Example 1.9.11. *Find an example of a continuous random variable X with a density and a continuous function $f : \mathbb{R} \circlearrowleft$, such that the random variable $f \circ X$ does not have a density (cf. assumptions in Proposition 1.9.10).*

We will take advantage of Proposition 1.9.10 heavily from now on and therefore not give an explicit example for it right now.

1.9.1 Second (and higher) moments

Whereas for a random variable with finite expectation the quantity $\mathbb{E}[X]$ is referred to as the first moment, it turns out that higher moments will play a crucial role, and sometimes they even completely determine a distribution (see e.g. [Bil95, Section 30]).

Definition 1.9.12. *Let $p \in \mathbb{N}$ and let X be a discrete or continuous real random variable such that the random variable X^p is quasi-integrable. Then the* p-th moment *of X is defined as $\mathbb{E}[X^p] \in [-\infty, \infty]$*
For $p > 0$, the space of all real random variables X on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[|X|^p] < \infty$ is denoted by $\mathcal{L}^p := \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$.

Since for $0 < p \leqslant q$ we have $|x|^p \leqslant 1 + |x|^q$ for all $x \in \mathbb{R}$ we immediately obtain the inclusion

$$\mathcal{L}^q \subset \mathcal{L}^p. \tag{1.9.8}$$

Example 1.9.13. *Let X be a continuous real random variable with probability density*

$$\varrho(x) = \frac{\mathbb{1}_{[1,\infty)}(x)\frac{1}{x^3}}{\int_{-\infty}^{\infty} \mathbb{1}_{[1,\infty)}(x)\frac{1}{x^3}\,\mathrm{d}x}, \quad x \in \mathbb{R}.$$

Then

$$\mathbb{E}[|X|^p] = \int_{-\infty}^{\infty} |x|^p \varrho(x)\,\mathrm{d}x$$

is finite for $p \in (0, 2)$ and infinite for $p \in [2, \infty)$. Thus, $X \in \mathcal{L}^p$ for $p \in (0, 2)$ but $X \notin \mathcal{L}^p$ for $p \in [2, \infty)$.

In order to prove the fundamental Hölder inequality below we will need the following auxiliary result.

Lemma 1.9.14 (Young's inequality (English mathematician William Henry Young (1863–1942)))**.** *Let $a, b \in [0, \infty)$ and $p, q \in (1, \infty)$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}. \tag{1.9.9}$$

Definition 1.9.15. *Let $I \subset \mathbb{R}$ be an interval and $\varphi : I \to \mathbb{R}$ a function. Then φ is called* convex *if for all $x, y \in I$ and $\lambda \in (0, 1)$ the inequality*

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda\varphi(x) + (1-\lambda)\varphi(y) (= \varphi(x) + (1-\lambda)(\varphi(y) - \varphi(x)) \tag{1.9.10}$$

holds true.
We call φ strictly convex *if the above inequality is strict whenever $x \neq y$.*

Proof of Lemma 1.9.14.

$$ab = \exp\{\ln(ab)\} = \exp\{\ln a + \ln b\} = \exp\Big\{\frac{1}{p}\ln(a^p) + \frac{1}{q}\ln(b^q)\Big\}$$
$$\overset{\text{convexity of the exponential function}}{\leqslant} \frac{1}{p}a^p + \frac{1}{q}b^q,$$

which finishes the proof. □

Theorem 1.9.16 (Hölder inequality (German mathematician Otto Ludwig Hölder (1859–1937)))**.** *Let $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then, for random variables X, Y one has*

$$\mathbb{E}[|XY|] \leqslant \big(\mathbb{E}[|X|^p]\big)^{\frac{1}{p}} \big(\mathbb{E}[|Y|^q]\big)^{\frac{1}{q}}. \tag{1.9.11}$$

Remark 1.9.17. *(a) In particular, this implies that if $X \in \mathcal{L}^p$ and $Y \in \mathcal{L}^q$, then $XY \in \mathcal{L}^1$.*

(b) The special case of $p = q = 1/2$ gives a version of the Cauchy-Schwarz (Augustin-Louis Cauchy (1789–1857), Hermann Schwarz (1843–1921) inequality you might know from linear algebra (or might get to know in functional analysis) for inner products.

(c) Hölder's inequality not only holds for expectations (which will be interpreted as integration against probability measures in 'Probability Theory I') but also for more general integrals.

(d) We will tacitly use the convention $0 \cdot \infty = 0$ from now on.

Proof. If one of the factors on the right-hand side of (1.9.11) is infinite or 0, then the statement is trivial. Indeed, if we have e.g. $\mathbb{E}[|X|^p] = 0$, then this implies that $\mathbb{P}(X = 0) = 1$ (exercise). Thus, $\mathbb{E}[XY] = 0$ as well, and the desired inequality holds true again. If, on the other hand, none of the factors of the right-hand side is zero, but one of them is infinity, then the inequality is obvious as well.
Thus, assume without loss of generality that both factors are finite and positive, and, again without loss of generality, we can even assume that both factors are 1, by considering

$$\frac{X}{\big(\mathbb{E}[|X|^p]\big)^{\frac{1}{p}}}$$

instead of X, and similarly for Y. Using Lemma 1.9.14 we deduce

$$\mathbb{E}[|XY|] \leqslant \mathbb{E}\Big[\frac{|X|^p}{p} + \frac{|Y|^q}{q}\Big] = \frac{1}{p}\mathbb{E}\big[|X|^p\big] + \frac{1}{q}\mathbb{E}\big[|Y|^q\big] = 1,$$

which finishes the proof. □

Definition 1.9.18. *For $X \in \mathcal{L}^1$, the quantity*

$$\mathrm{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \in [0, \infty]$$

is called the variance of X.

From the expression on the left-hand side it is clear that the variance is always non-negative, since the random variable in the expectation on the left-hand side is non-negative. Furthermore, this expression shows that the variance gauges the expected quadratic deviation of X from its expectation $\mathbb{E}[X]$. It is a simple measure for how strongly the random variable X fluctuates around its mean.
Using the generalized linearity of expectation given in Remark 1.9.8, we can rewrite the variance as

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

which holds true in the case $\mathbb{E}[X^2] = \infty$ as well. Thus, we immediately obtain the following corollary.

Corollary 1.9.19. *For $X \in \mathcal{L}^1$, we have $\mathrm{Var}(X) < \infty$ if and only if $\mathbb{E}[X^2] < \infty$.*

Definition 1.9.20. *The* covariance of two random variables $X, Y \in \mathcal{L}^1$ *is defined as*

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{1.9.12}$$

if the right-hand side is well-defined in $[-\infty, \infty]$.
The two random variables are called uncorrelated *if $\mathrm{Cov}(X, Y) = 0$.*

Again we note that variance and covariance only depend on the random variables involved through their corresponding distributions.
In some sense the covariance $\mathrm{Cov}(X, Y)$ tells us how strongly X and Y are correlated, i.e., how strongly they change together. If both X and Y tend to take values above their expectation on the same subset of Ω, and also tend to take values below their expectations on similar sets, then according to (1.9.12) this should imply that their covariance is positive; on the other hand, if X tends to take values above its expectation on subsets of Ω where Y tends to take values below its expectation, and vice versa, then this would suggest that their covariance is negative. Therefore, if X and Y are independent one might possibly guess that $\mathrm{Cov}(X, Y)$ vanishes. This is indeed the case as part (c) of Proposition 1.9.21 below shows. Note, however, that the converse is not generally true as will be asked to show in Exercise 1.9.22.
We now collect some properties of covariances and variances in the following result.

Proposition 1.9.21. *Let X and Y be random variables in $\mathcal{L}^2$ and let $a, b, c, d \in \mathbb{R}$. Then*

(a)

$$\operatorname{Cov}(aX + b, cY + d) = ac \operatorname{Cov}(X, Y);$$

in particular,

$$\operatorname{Var}(a(X + b)) = a^2 \operatorname{Var}(X); \tag{1.9.13}$$

(b)

$$|\operatorname{Cov}(X, Y)| \leqslant \sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)};$$

(c) if X and Y are independent, then they are uncorrelated;

Proof. (a) Using the linearity of expectation we get

$$\begin{aligned}\operatorname{Cov}(aX + b, cY + d) &= \mathbb{E}\big[(aX + b - \mathbb{E}[aX + b])(cY + d - \mathbb{E}[cY + d])\big] \\ &= ac\mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = ac \operatorname{Cov}(X, Y).\end{aligned}$$

(b)

$$\begin{aligned}|\operatorname{Cov}(X, Y)| &= \mathbb{E}\big[|X - \mathbb{E}[X]| \cdot |Y - \mathbb{E}[Y]|\big] \leqslant \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]^{\frac{1}{2}} \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big]^{\frac{1}{2}} \\ &= \sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)},\end{aligned}$$

where the inequality is a consequence of the Cauchy Schwarz inequality.

(c) Since $X, Y \in \mathcal{L}^2 \subset \mathcal{L}^1$, we use Proposition 1.9.7 to get $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ from which the statement follows immediately.

□

Exercise 1.9.22. *Find an example of real random variables X, Y which are uncorrelated but not independent.*

We now compute some variances of distributions we got to know earlier in this course.

Example 1.9.23.

- *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. Then we get using Proposition 1.9.10 that*

$$\begin{aligned}\operatorname{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] &= \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ \overset{x \mapsto \sigma x + \mu}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma x)^2 e^{-\frac{x^2}{2}} \mathrm{d}x &= \frac{\sigma^2}{\sqrt{2\pi}} \Big(\underbrace{-x e^{-\frac{x^2}{2}} \Big|_{x=-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \mathrm{d}x}_{=\sqrt{2\pi}} \Big) = \sigma^2,\end{aligned}$$

  where we used integration by parts for the penultimate equality. Hence, we observe that the second parameter in $\mathcal{N}(\mu, \sigma^2)$ denotes the variance of the random variable. In particular, this means that the normal distribution is completely distributed by its expectation and its variance.

  Furthermore, we deduce that the standard normal distribution from Example 1.8.25 (c) has mean 0 *and variance* 1.

- *Let $X \sim \operatorname{Geo}_p$ for $p \in (0, 1)$. We first compute $\mathbb{E}[X]$ and for this purpose we take advantage of the following useful trick. For $q \in (-1, 1)$, the formula for the geometric series supplies us with*

$$\sum_{j=1}^{\infty} q^j = \frac{q}{1 - q}.$$

  Since the left-hand side defines a power series that is absolutely convergent for $q \in (-1, 1)$, we recall from basic calculus lectures that its derivative can be computed term by term. Thus, differentiating both sides of the equation gives

$$\sum_{j=1}^{\infty} j q^{j-1} = \frac{(1 - q) - q(-1)}{(1 - q)^2} = \frac{1}{(1 - q)^2}. \tag{1.9.14}$$

Using this identity for $q = 1 - p$ we can compute

$$\mathbb{E}[X] = \sum_{j=1}^{\infty} j\mathbb{P}(X = j) = \sum_{j=1}^{\infty} jp(1-p)^{j-1} = \frac{p}{p^2} = \frac{1}{p}. \tag{1.9.15}$$

We now have to compute $\mathbb{E}[X^2]$. For this purpose we differentiate (1.9.14) *once again (and again, the left-hand side can be differentiated term by term on $(-1, 1)$ due to its absolute convergence) to obtain*

$$\sum_{j=2}^{\infty} j(j-1)q^{j-2} = \frac{2}{(1-q)^3}. \tag{1.9.16}$$

Thus, we get using the change of variable formula that

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{j=1}^{\infty} j^2\mathbb{P}(X = j) = \sum_{j=1}^{\infty} j^2 p(1-p)^{j-1} \\ &= p(1-p)\sum_{j=1}^{\infty} j(j-1)(1-p)^{j-2} + p\sum_{j=1}^{\infty} j(1-p)^{j-1} = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2},\end{aligned}$$

where we took advantage of (1.9.15) *and* (1.9.16) *to get the third equality. As a consequence, we can compute*

$$\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

If we want to compute the variance of the sum of random variables, the following result turns out to be useful by decomposing it into a sum of variances and corresponding covariances.

Proposition 1.9.24. *Let $X_1, \dots, X_n$ be random variables in $\mathcal{L}^2$. Then*

$$\operatorname{Var}\Big(\sum_{j=1}^{n} X_j\Big) = \sum_{j=1}^{n} \operatorname{Var}(X_j) + \sum_{1 \leqslant i,j \leqslant n, i \neq j} \operatorname{Cov}(X_i, X_j).$$

Proof. Due to Proposition 1.9.21 (a), without loss of generality, we can assume $\mathbb{E}[X_i] = 0$ for all $1 \leqslant i \leqslant n$. Using the linearity of expectation we get

$$\begin{aligned}\operatorname{Var}\Big(\sum_{j=1}^{n} X_j\Big) &= \mathbb{E}\Big[\Big(\sum_{j=1}^{n} X_j\Big)^2\Big] - \underbrace{\Big(\mathbb{E}\Big[\sum_{j=1}^{n} X_j\Big]\Big)^2}_{=0 \text{ by assumption}} = \sum_{i,j=1}^{n} \mathbb{E}[X_i X_j] \\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{1 \leqslant i,j \leqslant n, i \neq j} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

Note that

$$\mathbb{E}[|X_i X_j|] \leqslant \big(\mathbb{E}[X_i^2]\big)^{\frac{1}{2}} \big(\mathbb{E}[X_j^2]\big)^{\frac{1}{2}} < \infty$$

due to Hölder's inequality with $p = q = \frac{1}{2}$, hence all expectations in the above equations are well-defined, and so are all the sums. □

If the random variables in the above result turn out to be uncorrelated, all covariances in the above result vanish and the computation of the variance becomes significantly simpler. The corresponding result is used so often that it deserves its own name.

Corollary 1.9.25 (Bienaymé formula (Irénée-Jules Bienaymé (1796–1878), French probabilist and statistician))**.** *Let $X_1, \dots, X_n$ be uncorrelated random variables in $\mathcal{L}^2$. Then*

$$\operatorname{Var}\Big(\sum_{j=1}^{n} X_j\Big) = \sum_{j=1}^{n} \operatorname{Var}(X_j).$$

Example 1.9.26. *(a) Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent random variables. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Note that it follows from the linearity of expectations and Bienaymé's lemma that $X+Y$ has mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$. However, it is not clear that $X+Y$ is normally distributed. To check the latter, assume without loss of generality that $\mu_1=\mu_2=0$ and use the fact that $X+Y$ is a continuous real random variable with density given by the convolution $\varrho_{X+Y}=\varrho_X*\varrho_Y$ (we use this identity without proof, since the proof requires some measure theory). Thus,

$$\begin{aligned}
\varrho_{X+Y}(z) &= \int_{-\infty}^{\infty} \varrho_X(x)\varrho_Y(z-x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma_1^2}} e^{-\frac{(z-x)^2}{2\sigma_2^2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{\infty} \exp\Big\{-\frac{\sigma_2^2x^2+\sigma_1^2(z-x)^2}{2\sigma_1^2\sigma_2^2}\Big\}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \int_{-\infty}^{\infty} \exp\Big\{-\frac{\big(\sqrt{\sigma_1^2+\sigma_2^2}x-\frac{\sigma_1^2 z}{\sqrt{\sigma_1^2+\sigma_2^2}}\big)^2-\frac{\sigma_1^4z^2}{\sigma_1^2+\sigma_2^2}+\sigma_1^2z^2}{2\sigma_1^2\sigma_2^2}\Big\}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\Big\{-\frac{(\sigma_1^4z^2-\sigma_1^4z^2-\sigma_2^2z^2\sigma_1^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}\Big\} \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\Big\{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}\Big\},
\end{aligned}$$

where we completed the square for x in the numerator ('quadratische Ergänzung'). This proves that $X+Y$ is normally distributed with the required variance (and mean), and hence proves the claim.

(b) Let $X\sim \mathrm{Bin}_{n,p}$ for some $n\in\mathbb{N}$ and $p\in[0,1]$. In Claim 1.8.4 we had seen that X has the same distribution as $\sum_{j=1}^n Y_j$, where the Y_j are independent random variables distributed according to Bin_p. Now $\mathrm{Var}(Y_j)$ is easy to compute since $\mathbb{E}[Y_j]=p$ and $\mathbb{E}[Y_j^2]=p$. Thus, $\mathrm{Var}(Y_j)=p(1-p)$. Now since $\mathrm{Var}(X)$ depends on X only through its distribution, we get the first equality of

$$\mathrm{Var}(X)=\mathrm{Var}\Big(\sum_{j=1}^n Y_j\Big)=\sum_{j=1}^n \mathrm{Var}(Y_j)=np(1-p),$$

where in the second equality we used Corollary 1.9.25.

The following lemma is interesting in its own right, but a generalization of it will play an important role when we introduce the concept of conditional expectations (which heuristically will amount to averaging over partial information of $\mathcal{F}$ only) in 'Probability Theory I'. It can be interpreted in the sense that the best approximation to a random variable X by a constant c is via its expectation $c=\mathbb{E}[X]$ (if distance is measured in terms of the second moment of $X-c$).

Lemma 1.9.27. *Let $X\in\mathcal{L}^2$ be a random variable. Then the function*

$$\mathbb{R}\ni s\mapsto \mathbb{E}[(X-s)^2]$$

is minimized at $s=\mathbb{E}[X]$, which is the strict global minimu. In particular, we have $\mathbb{E}[(X-s)^2]\geqslant \mathrm{Var}(X)$ for all $s\in\mathbb{R}$.

Proof. We compute using the linearity of expectation of Proposition 1.9.7 that

$$\mathbb{E}[(X-s)^2]=\mathbb{E}[X^2]-2s\mathbb{E}[X]+s^2=(\mathbb{E}[X^2]-\mathbb{E}[X]^2)+(\mathbb{E}[X]-s)^2.$$

From this it is obvious that the function attains its minimum for $s=\mathbb{E}[X]$, in which case it equals $\mathrm{Var}(X)$. This finishes the proof. □

1.10 Generating functions

We have seen in Example 1.9.26 (a) that we can investigate sums of independent random variables; however, this resulted in slightly tedious computations and we give here a tool for simplifying this procedure, at least in certain important cases of $\mathbb{N}_0$-valued random variables. For this purpose we introduce the so-called 'generating functions' of $\mathbb{N}_0$-valued random variables completely characterise their distributions, see Theorem 1.10.2 below.

Definition 1.10.1. *Let X be an $\mathbb{N}_0$-valued random variable. The function*

$$\begin{aligned} G_X : [-1,1] &\to \mathbb{R} \\ s &\mapsto \sum_{n\in\mathbb{N}_0} \mathbb{P}(X=n)\cdot s^n (= \mathbb{E}[s^X] \text{ according to Proposition 1.9.10}) \end{aligned} \tag{1.10.1}$$

is called the generating function or moment generating function of X ('Erzeugendenfunktion', 'momentenerzeugende Funktion von X').

Another reason for why generating functions play an important role is that they, as well as their derivatives, can often be explicitly computed and give information about the moments of the random variable.

Theorem 1.10.2. *Let G_X be the generating function of the $\mathbb{N}_0$-valued random variable X.*
Then:

(a) The function G_X is infinitely differentiable on $(-1,1)$, and for $k\in\mathbb{N}$ one has

$$\lim_{s\uparrow 1} G_X^{(k)}(s) = \sum_{n=k}^{\infty} \mathbb{P}(X=n)\cdot n\cdot(n-1)\cdot\ldots\cdot(n-k+1).$$

In particular,

$$\lim_{s\uparrow 1} G_X'(s) = \mathbb{E}[X].$$

(b) The distribution of X is completely determined by G_X and vice versa. In particular, there is a one-to-one correspondence between distributions on $\mathbb{N}_0$ and generating functions.

Proof. (a) The radius of convergence of the power series in (1.10.1) is at least 1, which can be seen e.g. by direct inspection or an application of the root test. From the elementary theory of power series we recall that at points in the interior of its domain of convergence, the series is differentiable infinitely often and the derivatives of the power series are obtained by differentiating each of its summand separately. This entails

$$G_X^{(k)}(s) = \sum_{n=k}^{\infty} \mathbb{P}(X=n)\cdot n\cdot(n-1)\cdot\ldots\cdot(n-k+1)\cdot s^{n-k},$$

and taking the limit as $s\uparrow 1$ gives the desired equality.

(b) Observe that $G_X^{(k)}(0) = \mathbb{P}(X=k)\cdot k\cdot(k-1)\cdot\ldots\cdot 1$. Hence the distribution of X is completely determined by the sequence $(G_X^{(k)}(0))_{k\in\mathbb{N}_0}$, and therefore in particular by G_X itself. On the other hand, random variables with different distributions give rise to different generating functions (e.g., since two generating functions, i.e. power series around 0, whose domain of convergence includes $(-1,1)$, coincide if and only if all their coefficients coincide), and hence the one-to-one correspondence follows.

□

Example 1.10.3. *We compute the moment-generating function for a geometrically distributed random variable X with success parameter $p\in(0,1)$, recall* (1.8.2).
We get for $s\in(-1,1)$ that

$$G_X(s) = \sum_{k\in\mathbb{N}_0} \mathbb{P}(X=k)s^k = \sum_{k\in\mathbb{N}_0} p(1-p)^{k-1}s^k = \frac{p}{1-p}\frac{(1-p)s}{1-(1-p)s} = \frac{ps}{1-(1-p)s}. \tag{1.10.2}$$

Thus,

$$G_X'(s) = \frac{p(1-(1-p)s)+ps(1-p)}{(1-(1-p)s)^2} = \frac{p}{(1-(1-p)s)^2},$$

and we obtain using Theorem 1.10.2 that

$$\mathbb{E}[X] = \lim_{s\uparrow 1} G_X'(s) = \frac{1}{p}.$$

This coincides with our previous computation in (1.9.15)*, but it did not require any trick pulled out of our sleeve.*

Generating functions of sums of independent $\mathbb{N}_0$-valued random variables have the nice property that they can be written as the product of the generating functions of the single summands. Also, note that the sum of $\mathbb{N}_0$-valued random variables is $\mathbb{N}_0$-valued again.

Lemma 1.10.4. *Let X and Y be independent $\mathbb{N}_0$-valued random variables. Then*

$$G_{X+Y}(s) = G_X(s) \cdot G_Y(s) \quad \forall s \in (-1, 1).$$

Proof. We have

$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X] \cdot \mathbb{E}[s^Y] = G_X(s)G_Y(s),$$

where we took advantage of the fact that if X and Y are independent random variables, then so are s^X and s^Y (see Remark 1.7.15), in combination with Proposition 1.9.7 (c). □

One of the reasons generating functions are so useful arises from the combination of Lemma 1.10.4 with Theorem 1.10.2 (b), as is illustrated in the following example.

Example 1.10.5. *Let $X \sim \mathrm{Poi}(\nu)$ and $Y \sim \mathrm{Poi}(\mu)$ be independent random variables. We compute*

$$G_X(s) = \sum_{n \in \mathbb{N}_0} \mathrm{Poi}_\nu(n) s^n = e^{-\nu} \sum_{n \in \mathbb{N}_0} \frac{\nu^n}{n!} s^n = e^{-\nu(1-s)}, \tag{1.10.3}$$

and furthermore, using Lemma 1.10.4,

$$G_{X+Y}(s) = G_X(s)G_Y(s) = e^{-\nu(1-s)} e^{-\mu(1-s)} = e^{-(\nu+\mu)(1-s)}.$$

Thus, combining this equality with Theorems 1.10.2 and (1.10.3) *we deduce that $X + Y \sim \mathrm{Poi}(\nu + \mu)$.*

We have seen that the concept of generating functions is quite powerful, and hence it is natural to ask for a generalization to arbitrary real random variables. It turns out that in this context the concept of so-called *characteristic functions* will play a similar role, and we will only completely introduce this concept once we have a fully-fledged theory of Lebesgue integration available in 'Probability I'.

1.11 Convergence of random variables

As in analysis, asymptotic investigations play a fundamental role in probability theory, in particular when it comes to the limit theorems that we will be considering in Sections 1.13 and 1.15. To build a theoretical base for this we will introduce the fundamental types of convergence that we will encounter in probability theory and investigate their dependencies.

1.11.1 Almost sure convergence

This is one of the strongest types of convergence that we will consider.

Definition 1.11.1. *Let (X_n) be a sequence of real random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and let X be a real random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X_n converges* almost surely *(or* a.s.*) (*'fast sicher' *(oder auch* 'f.s.'*) to X, and we write*

$$X_n \xrightarrow{a.s.} X \quad \text{as } n \to \infty,$$

or

$$\lim_{n\to\infty} X_n = X \quad \mathbb{P} - a.s.,$$

if

$$\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = \mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1. \tag{1.11.1}$$

Remark 1.11.2. *(a) In order for the probabilities in* (1.11.1) *to be well-defined, we need that*

$$\{\lim_{n\to\infty} X_n = X\} \in \mathcal{F}. \tag{1.11.2}$$

Now recall from Definition 1.7.3 that $\mathcal{B}(\mathbb{R})$ contains all intervals. Therefore, and since $|X_n - X|$ is a random variable again (see Lemma 1.7.5), we obtain that $\{|X_n - X| \leqslant \frac{1}{k}\} \in \mathcal{F}$ for all $k, n \in \mathbb{N}$. Recalling the definition of the limit

$$\Big\{ \lim_{n\to\infty} X_n = X \Big\} = \bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \Big\{ |X_n - X| \leqslant \frac{1}{k} \Big\}$$

from basic analysis lectures, we therefore get (1.11.2) *using the stability of $\mathcal{F}$ under countable unions and intersections. Hence we get the alternative characterization of*

$$\lim_{n\to\infty} X_n = X \quad a.s.$$

via

$$\mathbb{P}\Big(\bigcap_{k=1}^{\infty} \bigcup_{n_0=1}^{\infty} \bigcap_{n=n_0}^{\infty} \Big\{ |X_n - X| \leqslant \frac{1}{k} \Big\} \Big) = 1. \tag{1.11.3}$$

(b) (1.11.1) *is equivalent to the existence of $N \in \mathcal{F}$ with $\mathbb{P}(N) = 0$ (N is called a* null set *in this case) such that for all $\omega \in \Omega \backslash N$ one has*

$$X_n(\omega) \to X(\omega) \quad \text{as } n \to \infty, \tag{1.11.4}$$

where the latter is just a statement about the convergence of a sequence of real numbers.

More generally, if there exists some $N \in \mathcal{F}$ so that a statement (such as e.g. (1.11.4)*) holds for all $\omega \in \Omega \backslash N$, then we say that it* holds $\mathbb{P}$-almost surely / $\mathbb{P}$-a.s. / a.s. ('$\mathbb{P}$-fast sicher', oder '$\mathbb{P}$-f.s.', oder 'f.s.').

(c) In particular, note that if X_n converges to X pointwise, then we have almost sure convergence as well. The reason that pointwise convergence is not so important to us is that modifications that only effect null sets cannot be noticed from a point of view of the probability measure. Note for example that we have seen in Remark 1.8.24 that changing a probability density in finitely many points does not have any effect on the corresponding distribution function.

1.11.2 Convergence in $\mathcal{L}^p$

This is yet another fairly strong type of convergence which in a slightly more general form plays an important role in (functional) analysis, too.

Definition 1.11.3. *Let $p > 0$, let (X_n) be a sequence of random variables in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$, and let $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ as well. Then we say that X_n* converges to X in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$*, and write*

$$X_n \xrightarrow{\mathcal{L}^p} X$$

if

$$\mathbb{E}[|X_n - X|^p] \to 0 \quad \text{as } n \to \infty.$$

As long as we do not impose any further assumptions (which we don't do for the time being), none of the above two types of convergence is actually stronger than the other, but we will really only address $\mathcal{L}^p$ convergence in 'Probability I' (or you might do so in Analysis III before).

Example 1.11.4. *Consider the probability space obtained by endowing $[0, 1)$ with the respective Borel-σ-algebra and the probability measure $\mathbb{P}$ given by the uniform distribution on $[0, 1)$ (to the extent to which it has been introduced in Examples 1.8.21 and 1.8.25).*

(a) Consider for $n \geqslant 1$ and $k \in \{0, 1, \ldots, 2^n - 1\}$ the random variables

$$X_{n,k} := \mathbb{1}_{[k2^{-n}, (k+1)2^{-n})}$$

and define $Y_1 := X_{1,0}$, $Y_2 := X_{1,1}$, $Y_3 := X_{2,0}$, $Y_4 := X_{2,1}$, ... (this is the 'lexicographic ordering'). Then $\limsup_{n\to\infty} Y_n = 1$ and $\liminf_{n\to\infty} Y_n = 0$, and in particular Y_n does not converge almost surely. On the other hand, for $p > 0$, any $n \in \mathbb{N}$, and $k \in \{0, \ldots, 2^n - 1\}$ we have

$$\mathbb{E}[|X_{n,k} - 0|^p] = \mathbb{P}([0, 2^{-n})) = 2^{-n},$$

and the right-hand side converges to 0 as $n \to \infty$. Therefore, $Y_n \xrightarrow{\mathcal{L}^p} 0$ as $n \to \infty$.

This example shows that convergence in $\mathcal{L}^p$ does not imply almost sure convergence.

(b) Fix $p > 0$ and consider the random variables $X_n := n^{\frac{1}{p}} \mathbb{1}_{[0,1/n]}$. Then for any $\omega \in (0,1)$ fixed we have

$$X_n(\omega) = n^{\frac{1}{p}} \mathbb{1}_{[0,1/n]}(\omega),$$

and the right-hand side converges to 0 as $n \to \infty$. Therefore,

$$\big\{ \lim_{n\to\infty} X_n = 0 \big\} = (0,1),$$

and since $\mathbb{P}((0,1)) = 1$ this implies that $\lim_{n\to\infty} X_n = 0$ almost surely.

On the other hand, a moment's thought reveals that since $X_n \to 0$ holds $\mathbb{P}$-a.s. as $n \to \infty$, the only possible limit in $\mathcal{L}^p$ (modulo events of probability zero) would be the constant random variable $X = 0$. Now for all $n \in \mathbb{N}$ one has

$$\mathbb{E}[|X_n - 0|^p] = 1,$$

and therefore X_n does not converge to 0 in $\mathcal{L}^p$.

This example shows that almost sure convergence does not imply convergence in $\mathcal{L}^p$.

1.11.3 Convergence in probability

Definition 1.11.5. *Let (X_n) be a sequence of real random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and let X also be a real random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X_n* converges in probability ('konvergiert in Wahrscheinlichkeit'oder 'konvergiert stochastisch') *to X if for all $\varepsilon > 0$,*

$$\mathbb{P}(|X_n - X| \geqslant \varepsilon) \to 0, \quad \text{as } n \to \infty.$$

In this case we write

$$X_n \xrightarrow{\mathbb{P}} X \quad \text{as } n \to \infty.$$

1.11.4 Convergence in distribution

Definition 1.11.6. *Let (μ_n) be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and let μ be yet another probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In addition, let real random variables $X_n \sim \mu_n$ and $X \sim \mu$ which can be defined on different probability spaces $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ and $(\Omega, \mathcal{F}, \mathbb{P})$, respectively, be given. We say that (μ_n)* converges in weakly ('konvergiert schwach') *to μ if for all continuous bounded functions $f \in C_b(\mathbb{R})$ from $\mathbb{R}$ to $\mathbb{R}$ we have*

$$\mathbb{E}_n[f(X_n)] \to \mathbb{E}[f(X)] \quad \text{as } n \to \infty.$$

In this case we write

$$\mu_n \xrightarrow{w} \mu \quad \text{as } n \to \infty,$$

where w stands for 'weakly'.

Remark 1.11.7. *Note that according to Remark 1.8.1, random variables X_n and X as in the above definition exist.*

Definition 1.11.8. *Let (X_n) be a sequence of real random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and let X also be a real random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that X_n* converges in distribution ('konvergiert in Verteilung') *to X if $\mathbb{P} \circ X_n^{-1}$ converges weakly to $\mathbb{P} \circ X^{-1}$.*
In this case we write

$$X_n \xrightarrow{\mathcal{L}} X \quad \text{as } n \to \infty,$$

or also

$$X_n \xrightarrow{\mathcal{D}} X \quad \text{as } n \to \infty.$$

Here, $\mathcal{L}$ and $\mathcal{D}$ stand for 'law' and 'distribution', respectively. Yet another very common notation is

$$X_n \Longrightarrow X \quad \text{as } n \to \infty.$$

Since we are mostly dealing with real random variables in this course, the following equivalent criterion for convergence in distribution of real random variables will come in handy.

Lemma 1.11.9. *Let (X_n) be a sequence of real random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and let X also be a real random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Denote the corresponding distribution functions by F_n and F, respectively. Then the following are equivalent:*

(a)

$$X_n \xrightarrow{\mathcal{D}} X;$$

(b) For all points t of continuity of F, one has

$$F_n(t) \to F(t) \quad \text{as } n \to \infty;$$

Proof. '$(a) \Rightarrow (b)$':
Let t_0 be a point of continuity for F and for $\varepsilon > 0$ arbitrary choose $f_{t_0,\varepsilon} \in C_b(\mathbb{R})$ with

$$\mathbb{1}_{(-\infty,t_0]} \leqslant f_{t_0,\varepsilon} \leqslant \mathbb{1}_{(-\infty,t_0+\varepsilon]}.$$

We get

$$\limsup_{n\to\infty} F_n(t_0) \leqslant \lim_{n\to\infty} \mathbb{E}[f_{t_0,\varepsilon}(X_n)] = \mathbb{E}[f_{t_0,\varepsilon}(X)] \leqslant F(t_0+\varepsilon).$$

Since $\varepsilon > 0$ was chosen arbitrarily and F is right-continuous, this supplies us with

$$\limsup_{n\to\infty} F_n(t_0) \leqslant F(t_0).$$

Similarly, choosing $\widetilde{f}_{t_0,\varepsilon} \in C_b(\mathbb{R})$ with

$$\mathbb{1}_{(-\infty,t_0-\varepsilon]} \leqslant \widetilde{f}_{t_0,\varepsilon} \leqslant \mathbb{1}_{(-\infty,t_0]}.$$

and using that t_0 is in fact a point of continuity of F (which implies the left-continuity of F in t_0), we obtain in a similar manner as before the inequality

$$\liminf_{n\to\infty} F_n(t_0) \geqslant F(t_0).$$

Combining these two inequalities supplies us with $\lim_{n\to\infty} F_n(t_0) = F(t_0)$.
'$(b) \Rightarrow (a)$': Assume (b) to hold true. For $\varepsilon > 0$ given we choose points of continuity $t_0 < t_1 < \ldots < t_{m_\varepsilon}$ such that

(a) $F(t_0) \leqslant \varepsilon$ and $F(t_{m_\varepsilon}) \geqslant 1 - \varepsilon$;

(b) for all $i \in \{1, \ldots, m_\varepsilon\}$,

$$|f(t) - f(t_i)| \leqslant \varepsilon \quad \forall t \in [t_{i-1}, t_i]. \tag{1.11.5}$$

Then

$$\mathbb{E}[f(X_n)] = \mathbb{E}\big[f(X_n)\mathbb{1}_{X_n \leqslant t_0}\big] + \sum_{i=1}^{m_\varepsilon} \mathbb{E}\big[f(X_n)\mathbb{1}_{X_n \in (t_{i-1},t_i]}\big] + \mathbb{E}\big[f(X_n)\mathbb{1}_{X_n > t_{m_\varepsilon}}\big]. \tag{1.11.6}$$

Using the inequality

$$f(X_n)\mathbb{1}_{X_n\in(t_{i-1},t_i]} \leqslant \big(\inf_{t\in(t_{i-1},t_i]} f(t) + \varepsilon\big) \cdot \mathbb{1}_{X_n\in(t_{i-1},t_i]}$$

(which itself is a consequence of (1.11.5)), in combination with the monotonicity property of the expectation operator and assumption (b), the $\limsup_{n\to\infty}$ of the right-hand side of (1.11.6) can be upper bounded by

$$2\varepsilon \sup_{t\in\mathbb{R}} |f(t)| + \sum_{i=1}^{m_\varepsilon} \inf_{t\in(t_{i-1},t_i]} f(t)\big(F(t_i) - F(t_{i-1})\big) + \varepsilon.$$

This again can be upper bounded by

$$4\varepsilon \sup_{t\in\mathbb{R}} |f(t)| + 2\varepsilon + \mathbb{E}[f(X)].$$

In a similar way we can derive

$$\liminf_{n\to\infty} \mathbb{E}[f(X_n)] \geqslant -4\varepsilon \sup_{t\in\mathbb{R}} |f(t)| - 2\varepsilon + \mathbb{E}[f(X)].$$

Thus, since $\varepsilon > 0$ was arbitrary, all in all we get

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$$

Since $f \in C_b(\mathbb{R})$ was chosen arbitrarily, this implies (a). □

Remark 1.11.10. *In contrast to the characterization of convergence in distribution through the convergence of the distribution functions at the points of continuity of F given in Lemma 1.11.9, Definition 1.11.8 has the advantage that it can be easily generalized to random variables that take values in spaces which are more general than $\mathbb{R}$, since we only need to have the concept of bounded real-valued functions on that corresponding space.*

In the following section we provide some basic inequalities. These are interesting and significant in their own right, and they will serve us to study the interdependence of the types of convergence of random variables introduced in this section.

1.12 Some fundamental tools and inequalities

1.12.1 Markov's and Chebyshev's inequalities

We now introduce some fundamental inequalities. These play a central role in probability and are some of the standard tools one has to feel comfortable to apply.

Proposition 1.12.1 (Markov's inequality (Andrey Andreyevich Markov (1856–1922)))**.** *Let X be a real random variable and let $\varepsilon > 0$. Then, for any increasing function $\varphi : [0,\infty) \to [0,\infty)$ with $\varphi((0,\infty)) \subset (0,\infty)$ one has*

$$\mathbb{P}(|X| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[\varphi(|X|)]}{\varphi(\varepsilon)}. \tag{1.12.1}$$

Proof. Since φ is monotone increasing we have the inequality

$$\varphi(|X|) \geqslant \mathbb{1}_{|X|\geqslant\varepsilon}\varphi(\varepsilon),$$

and taking expectations on both sides supplies us with

$$\mathbb{E}[\varphi(|X|)] \geqslant \mathbb{P}(|X| \geqslant \varepsilon)\varphi(\varepsilon),$$

which implies (1.12.1). □

Corollary 1.12.2 (Chebyshev's inequality (Pafnuty Chebyshev (1821–1894)))**.** *Let X be in $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$. Then*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2}. \tag{1.12.2}$$

Proof. This follows from Proposition 1.12.1 by choosing the random variable in (1.12.1) as $X - \mathbb{E}[X]$ and $\varphi(x) := x^2$. □

Remark 1.12.3. *Inequalities of the type* (1.12.2) *which provide upper bounds for the probability that X deviates from a certain quantity, such as its expectation, are also referred to as 'concentration inequalities'.*

1.12.2 The Borel-Cantelli lemmas

In order to prove this theorem we need some further results, which are important and of interest on their own. For this purpose we introduce the following notation.

Definition 1.12.4. *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and assume given a sequence (A_n) of events $A_n \in \mathcal{F}$. Then the 'limes superior' of the sequence (A_n) is defined as*

$$\limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k.$$

The 'limes inferior' of the sequence (A_n) is defined as

$$\liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k$$

Exercise 1.12.5. *Show the following identities:*

- $$\limsup_{n\to\infty} A_n = \big\{\omega \in \Omega \,:\, \omega \in A_n \text{ for infinitely many } n\big\};$$
- $$\liminf_{n\to\infty} A_n = \big\{\omega \in \Omega \,:\, \text{ such that } \exists n_0 \in \mathbb{N} \text{ with } \omega \in A_n \ \forall n \geqslant n_0\big\};$$

Lemma 1.12.6 (Borel-Cantelli lemma (French mathematician and politician Émile Borel (1871–1956), Italian mathematician Francesco Paolo Cantelli (1875–1966))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and assume given a sequence (A_n) of events $A_n \in \mathcal{F}$.*

(a) If

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty, \tag{1.12.3}$$

we have

$$\mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = 0.$$

(b) If

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty,$$

and if in addition the (A_n) are independent, then

$$\mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = 1.$$

Proof. (a) For any $k \in \mathbb{N}$ we have

$$\limsup_{n\to\infty} A_n \subset \bigcup_{n=k}^{\infty} A_n,$$

and therefore

$$\mathbb{P}\big(\limsup_{n\to\infty} A_n\big) \leqslant \mathbb{P}\Big(\bigcup_{n=k}^{\infty} A_n\Big) \leqslant \sum_{n=k}^{\infty} \mathbb{P}(A_n)., \tag{1.12.4}$$

Now due to (1.12.3) we get that

$$\sum_{n=k}^{\infty} \mathbb{P}(A_n) \to 0 \quad \text{ as } k \to \infty.$$

Therefore, since (1.12.4) was valid for any k, we get

$$\mathbb{P}\big(\limsup_{n\to\infty} A_n\big) = 0.$$

(b) The usual proof considers the probability of the complement and takes advantage of De Morgan's rule that

$$(\limsup_{n\to\infty} A_n)^c = \Big(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\Big)^c = \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c.$$

Now observe that using the independence of the family (A_n) in combination with the continuity from above of probability measure, we infer for $N > k$ the identity

$$\mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) = \lim_{M\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{M} A_k^c\Big) = \lim_{M\to\infty} \prod_{k=n}^{N} \mathbb{P}\big(A_k^c\big) \cdot \mathbb{P}\Big(\bigcap_{k=N+1}^{M} A_k^c\Big) = \prod_{k=n}^{N} \mathbb{P}\big(A_k^c\big) \cdot \mathbb{P}\Big(\bigcap_{k=N+1}^{\infty} A_k^c\Big).$$

Combining the above, we therefore obtain

$$\begin{aligned}\mathbb{P}\big((\limsup_{n\to\infty} A_n)^c\big) &\leqslant \sum_{n=1}^{\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) = \sum_{n=1}^{\infty} \lim_{N\to\infty} \prod_{k=n}^{N} \mathbb{P}\big(A_k^c\big)\mathbb{P}\Big(\bigcap_{k=N+1}^{\infty} A_k^c\Big) \leqslant \sum_{n=1}^{\infty}\prod_{k=n}^{\infty} \mathbb{P}\big(A_k^c\big) \\ &= \sum_{n=1}^{\infty}\prod_{k=n}^{\infty}(1-\mathbb{P}(A_k)) \leqslant \sum_{n=1}^{\infty} \exp\Big\{-\underbrace{\sum_{k=n}^{\infty} \mathbb{P}(A_k)}_{\infty}\Big\} = 0,\end{aligned}$$

where we used that for $t \geqslant 0$ we have $e^{-t} \geqslant 1 - t$.

□

Remark 1.12.7. *It is important to note here that the independence assumption in part (b) of Lemma 1.12.6 cannot be dropped. To see this, consider for example a single fair coin toss modeled on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and denote for all $n \in \mathbb{N}$ by A_n the event that the coin shows tails. Then $\mathbb{P}(A_n) = \frac{1}{2}$ for all $n \in \mathbb{N}$, so $\sum_{n\in\mathbb{N}_0} \mathbb{P}(A_n) = \infty$, but $\mathbb{P}(\limsup_{n\to\infty} A_n) = \mathbb{P}(A_n) = \frac{1}{2} \neq 1$.*

Example 1.12.8. *A popular application is the so-called 'infinite monkey theorem'. It states that a monkey which is randomly hitting keys (in an i.i.d. fashion, and such that any key, lower and upper case, has a positive probability of being hit) of a computer keyboard will almost surely type any given text, such as e.g. Tolstoy's 'War and Peace'. It is left to the reader to make this statement more precise.*

1.12.3 Jensen's inequality

Theorem 1.12.9 (Jensen's inequality (Danish mathematician Johan Jensen (1859 – 1925)))**.** *Let X be a real random variable in $\mathcal{L}^1$ and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function (if X is a non-negative random variable, then it is sufficient for φ to be a convex function defined on $[0, \infty)$). Then*

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)] \in (-\infty, \infty]. \tag{1.12.5}$$

In particular, if in addition φ is strictly convex and $\mathbb{P}_X$ is not concentrated in a single point (i.e., there does not exist any $a \in \mathbb{R}$ such that $\mathbb{P}_X(\{a\}) = 1$), then the inequality in (1.12.5) is strict.

Proof. Since φ is convex, we get that for all $x_0 \in \mathbb{R}$ the one-sided derivatives

$$\varphi'_-(x_0) := \lim_{h\downarrow 0} \frac{\varphi(x_0) - \varphi(x_0 - h)}{h} \quad \text{and } \varphi'_+(x_0) := \lim_{h\downarrow 0} \frac{\varphi(x_0 + h) - \varphi(x_0)}{h}$$

exist (since the corresponding functions are monotone in h) and that $\varphi'_-(x_0) \leqslant \varphi'_+(x_0)$. Therefore, choosing any slope $s(x_0) \in [\varphi'_-(x_0), \varphi'_+(x_0)]$, we get for all $x \in \mathbb{R}$ that

$$\varphi(x) \geqslant \varphi(x_0) + s(x_0)(x - x_0).$$

Taking $x_0 := \mathbb{E}[X] \in (-\infty, \infty)$ and $x = X(\omega)$ we obtain

$$\varphi(X(\omega)) \geqslant \varphi(\mathbb{E}[X]) + s(\mathbb{E}[X])(X(\omega) - \mathbb{E}[X]). \tag{1.12.6}$$

Taking expectations on both sides yields (1.12.5), where the finiteness of the expectation on the right-hand side of (1.12.6) implies $\mathbb{E}[\varphi(X)] \in (-\infty, \infty]$.
In the case of φ being strictly convex, the inequality in (1.12.6) is strict once $X(\omega) \neq \mathbb{E}[X]$. If now in addition $\mathbb{P}_X$ is not concentrated in a single point, then $\{\omega \in \Omega : X(\omega) \neq \mathbb{E}[X]\}$ has positive probability and the desired strict inequality in (1.12.5) follows. □

Remark 1.12.10. *(a) If $\widetilde{\varphi}$ is a concave function on $\mathbb{R}$, then $-\widetilde{\varphi}$ is a convex function, hence Theorem 1.12.9 yields*

$$\widetilde{\varphi}(\mathbb{E}[X]) \geqslant \mathbb{E}[\widetilde{\varphi}(X)]$$

for $X \in \mathcal{L}^1$.

(b) This immediately supplies us with another proof for the inclusion $\mathcal{L}^q \subset \mathcal{L}^p$ for $q, p \in (0, \infty)$ with $q > p$ which we had derived in (1.9.8). Indeed, since the function $\widetilde{\varphi}(x) := x^{\frac{p}{q}}$ is concave on $[0, \infty)$ and since $|X|$ is non-negative, we get for $X \in \mathcal{L}^q$ that

$$\infty > \widetilde{\varphi}(\mathbb{E}[|X|^q]) \geqslant \mathbb{E}[\widetilde{\varphi}(|X|^q)] = \mathbb{E}[|X|^p].$$

Thus, $\mathbb{E}[|X|^p] < \infty$ which implies $X \in \mathcal{L}^p$.

Example 1.12.11. *Let $X \in \mathcal{L}^1$.*

(a) Consider the absolute value function $\varphi(x) := |x|$ and check that it is convex. Thus, Jensen's inequality yields $|\mathbb{E}[X]| \leqslant \mathbb{E}[|X|]$.

(b) Choosing the convex function $\varphi(x) := x^2$, Jensen's inequality supplies us with

$$\mathbb{E}[|X|]^2 \leqslant \mathbb{E}[X^2].$$

1.13 Interdependence of types of convergence of random variables

Having introduced all the above types of convergence, it is natural to try to order them in terms of strength. As we have seen in Example 1.11.4, there are no general implications between convergence in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{P}$-almost sure convergence. However, for the remaining ones we have the following hierarchy.

Theorem 1.13.1. *Let X_n, X be real random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $p > 0$.*

(a) If either $\lim_{n\to\infty} X_n = X$ almost surely, or if X, $X_n \in \mathcal{L}^p$ and $X_n \xrightarrow{\mathcal{L}^p} X$, then

$$X_n \xrightarrow{\mathbb{P}} X. \tag{1.13.1}$$

(b) If $X_n \xrightarrow{\mathbb{P}} X$, then

$$X_n \Longrightarrow X. \tag{1.13.2}$$

(c) If $0 < p < q < \infty$ and if (X_n) and X are in $\mathcal{L}^q$ such that $X_n \xrightarrow{\mathcal{L}^q} X$, then $X_n \xrightarrow{\mathcal{L}^p} X$ as well.

(d) If

$$\text{for all } \varepsilon > 0 \text{ one has} \quad \sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| \geqslant \varepsilon) < \infty, \tag{1.13.3}$$

then $\lim_{n\to\infty} X_n = X$ $\mathbb{P}$-a.s.[13]

In particular, if $X_n \xrightarrow{\mathbb{P}} X$, then there exists a subsequence (X_{n_k}) of (X_n) such that

$$X_{n_k} \longrightarrow X \quad \mathbb{P}-a.s.$$

Proof. (a) Assume that $\lim_{n\to\infty} X_n = X$ almost surely first, and for $\varepsilon > 0$ define the decreasing sequence of sets

$$B_{n,\varepsilon} := \big\{\exists m \geqslant n \,:\, |X_m - X| \geqslant \varepsilon\big\}, \quad n \in \mathbb{N}.$$

Denoting by N the null set introduced in the context of $\mathbb{P}$-a.s. convergence in Remark 1.11.2 (b), we get

$$\bigcap_{n\in\mathbb{N}} B_{n,\varepsilon} \subset N. \tag{1.13.4}$$

Thus,

$$\limsup_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \varepsilon) \leqslant \limsup_{n\to\infty} \mathbb{P}(B_{n,\varepsilon}) \leqslant \mathbb{P}(N) = 0,$$

where the last inequality follows from (1.13.4) and the continuity of $\mathbb{P}$ from above (recall Proposition 1.3.9 (g)). This implies (1.13.1).

Next, assume that instead of almost sure convergence we have X, $X_n \in \mathcal{L}^p$ and $X_n \xrightarrow{\mathcal{L}^p} X$. Then using Markov's inequality from Proposition 1.12.1 with $\varphi(x) := |x|^p$ supplies us with

$$\mathbb{P}(|X_n - X| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[|X_n - X|^p]}{|\varepsilon|^p} \to 0 \quad \text{as } n \to \infty.$$

Thus, (1.13.1) holds true.

(b) Fix an arbitrary $f \in C_b(\mathbb{R})$ and set $M := \sup_{x\in\mathbb{R}} |f(x)| < \infty$. Since the probability measure $\mathbb{P} \circ X^{-1}$ is continuous from above, for given $\varepsilon > 0$ we find $N_\varepsilon \in \mathbb{N}$ such that

$$\mathbb{P}(X \in [-N_\varepsilon, N_\varepsilon]^c) \leqslant \varepsilon.$$

Since f is uniformly continuous on $[-N_\varepsilon - 1, N_\varepsilon + 1]$, we find $\delta \in (0, 1)$ such that for all $x, y \in [-N_\varepsilon - 1, N_\varepsilon + 1]$ with $|x - y| \leqslant \delta$ we have

$$|f(x) - f(y)| \leqslant \varepsilon. \tag{1.13.5}$$

[13] If (1.13.3) holds true one says that X_n converges *fast* or *almost completely* to X.

As a consequence, using the triangle inequality we obtain

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| \leqslant \underbrace{\Big|\mathbb{E}\big[f(X_n) - f(X);\, X \in [-N_\varepsilon, N_\varepsilon], |X_n - X| < \delta\big]\Big|}_{\overset{(1.13.5)}{\leqslant}\ \varepsilon} + 2M \cdot \mathbb{P}(X \in [-N_\varepsilon, N_\varepsilon]^c) + 2M \cdot \mathbb{P}(|X_n - X| \geqslant \delta).$$

Taking the $\limsup_{n\to\infty}$ on both sides, we get that the resulting right-hand side is upper bounded by

$$\varepsilon + 4M\varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrarily small, this implies

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)],$$

and hence, since $f \in C_b(\mathbb{R})$ was chosen arbitrarily, (1.13.2) follows.

(c) Jensen's inequality 1.12.9 applied with the convex function $\varphi(x) := |x|^{q/p}$ supplies us with

$$\varphi(\mathbb{E}[|X_n - X|^p]) \leqslant \mathbb{E}[\varphi(|X_n - X|^p)]) \leqslant \mathbb{E}[|X_n - X|^q].$$

As $n \to \infty$, the right-hand side converges to 0 by assumption, and therefore so does the left-hand side, which implies $X_n \xrightarrow{\mathcal{L}^p} X$ as $n \to \infty$.

(d) From (1.13.3) we can infer the existence of a subsequence X_{n_k} such that

$$\sum_{m=n_k}^{\infty} \mathbb{P}(|X_m - X| \geqslant 2^{-k}) \leqslant 2^{-k}$$

for all $k \in \mathbb{N}$.

Setting

$$A_k := \bigcup_{m=n_k}^{\infty} \big\{|X_m - X| \geqslant 2^{-k}\big\}$$

we thus get

$$\mathbb{P}(A_k) \leqslant \sum_{m=n_k}^{\infty} \mathbb{P}(|X_m - X| \geqslant 2^{-k}) \leqslant 2^{-k}.$$

The first part of the Borel-Cantelli lemma then supplies us with

$$\mathbb{P}(\limsup_{k\to\infty} A_k) = 0,$$

which implies that there exists a null set $N \subset \mathcal{F}$ such that for each $\omega \in \Omega\backslash N$ there exists $N_\omega \in \mathbb{N}$ with the property that

$$\omega \in \bigcap_{n=N_\omega}^{\infty} A_k^c,$$

and this again means that $X_n(\omega) \to X(\omega)$ as $n \to \infty$. Therefore, $\lim_{n\to\infty} X_n = X$ $\mathbb{P}$-a.s.

If, on the other hand, X_n converges in probability to X, then we can choose a subsequence X_{n_k} of X_n such that $Y_k := X_{n_k}$ fulfills

$$\mathbb{P}\Big(|Y_k - X| \geqslant \frac{1}{k}\Big) \leqslant \frac{1}{k^2}$$

and hence condition (1.13.3), and due to the first part one then has $X_{n_k} \to X$ $\mathbb{P}$-a.s. as $k \to \infty$.

□

Exercise 1.13.2. *Show that the converses of the convergence implications given in Theorem 1.13.1 (a) to (c) do not hold true in general.*

1.14 Laws of large numbers

One central topic in probability theory is the asymptotic analysis of random systems and one of the simplest and more or less realistic situations to imagine is arguably a very long (or, possibly slightly less realistic, an infinite) sequence of independent coin tosses or dice rolls. For the sake of simplicity let's have a look at the situation of independent fair coin tosses, and define for $n \in \mathbb{N}$ a random variable X_n on $(\Omega, \mathcal{F}, \mathbb{P})$ that takes the value 1 if the coin of the n-th toss shows heads, whereas it takes the value -1 if the coin shows tails, i.e., the X_n are Rademacher-distributed. Now we know that $\mathbb{E}[X_n] = 0$, and also for the sum

$$S_n := \sum_{j=1}^{n} X_j \tag{1.14.1}$$

we have $\mathbb{E}[S_n] = 0$ by the linearity of expectation.

Definition 1.14.1. *The sequence S_n as defined in* (1.14.1) *is also called* simple random walk (SRW) ('einfache Irrfahrt')

Oftentimes, instead of investigating the expectation, one is interested e.g. in realizationwise statements, or statements concerning probabilities of certain events. In our current setting for example, one might want to ask what values $S_n(\omega)$ 'typically' takes. Now, although $\mathbb{E}[S_n] = 0$ for all $n \in \mathbb{N}$, it is obvious that $S_n(\omega) = 0$ can only hold true if n is even. And in fact, even when n is even, 0 is not the typical value for S_n to take, in the sense that it would be realized with a high probability or at least with a probability that is bounded away from 0 as $n \to \infty$. Indeed, for $n = 2k$ even we get with Stirling's formula that

$$\mathbb{P}(S_{2n} = 0) = \binom{2n}{n}\Big(\frac{1}{2}\Big)^{2n} \sim \frac{(2n/e)^{2n}\sqrt{2\pi \cdot 2n}}{\big((n/e)^n\sqrt{2\pi n}\big)^2} 2^{-2n} = \frac{1}{\sqrt{n\pi}}, \tag{1.14.2}$$

where for sequences (a_n) and (b_n) of positive real numbers we write $a_n \sim b_n$ if $\lim_{n\to\infty} a_n/b_n = 1$.

Exercise 1.14.2. *Using an explicit computation as in* (1.14.2)*, show that although $\mathbb{P}(S_n = 0) \to 0$ due to* (1.14.2)*, for any $n \in \mathbb{N}$, the function $\mathbb{Z} \ni m \mapsto \mathbb{P}(S_{2n} = m)$ is maximised for $m = 0$.*

Thus (1.14.2) tells us that $\mathbb{P}(S_n = 0)$ goes to zero at the order of $n^{-\frac{1}{2}}$. One might therefore be tempted to guess that if instead of just considering 0, we were replacing it by intervals of the type $[-c\sqrt{n}, c\sqrt{n}]$, then we would obtain a non-trivial limiting probability for S_n to take values in such intervals. This is indeed the case (and not only if the X_n describe coin tosses, but for far more general distributions of X) as will be established in the central limit theorem (see Theorem 1.15.1 below). For the time being, however, we start with having a look at a simpler result at cruder scales.

1.14.1 Weak law of large numbers

We will start with investigating the so-called empirical mean.

Definition 1.14.3. *Given a realization $X_1(\omega), \dots, X_n(\omega)$ of random variables, its* empirical mean *is defined as*

$$\frac{1}{n}S_n(\omega) = \frac{1}{n}\sum_{j=1}^{n} X_j(\omega). \tag{1.14.3}$$

Remark 1.14.4. *According to Lemma 1.7.5, the empirical mean $\frac{1}{n}S_n$ defined in* (1.14.3) *is a random variable again.*

In order to be able to prove something meaningful about the empirical mean, we will take advantage of Chebyshev's inequality introduced in Corollary 1.12.2 above.
As suggested by (1.14.2) and the heuristics developed subsequently, we might guess that the empirical mean defined in (1.14.3) will converge to 0 under suitable assumptions on the sequence (X_n).

Definition 1.14.5. *A sequence (X_n) of elements of $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ satisfies a* weak law of large numbers *if*

$$\frac{1}{n}\Big(\sum_{j=1}^{n} X_j - \mathbb{E}[X_j]\Big) \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty. \tag{1.14.4}$$

Historically, a weak law of large numbers had first been rigorously derived by Jakob Bernoulli in [Ber13]. Nevertheless, the intuition for such a statement must have been around at that time already since in a correspondence Jakob Bernoulli writes to Gottfried Wilhelm Leibniz in October 1703 [vdWB75, pp. 509–513]: 'Obwohl aber seltsamerweise durch einen sonderbaren Naturinstinkt auch jeder Dümmste ohne irgend eine vorherige Unterweisung weiss, dass je mehr Beobachtungen gemacht werden, umso weniger die Gefahr besteht, dass man das Ziel verfehlt, ist es doch ganz und gar nicht Sache einer Laienuntersuchung, dieses genau und geometrisch zu beweisen.'

Theorem 1.14.6 (Weak law of large numbers)**.** *Let (X_n) be a sequence of pairwise uncorrelated random variables in $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and let (α_n) be a sequence of real numbers such that*

$$\frac{\sum_{j=1}^n \operatorname{Var}(X_j)}{\alpha_n^2} \to 0. \tag{1.14.5}$$

Then for all $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\frac{\sum_{j=1}^n (X_j - \mathbb{E}[X_j])}{\alpha_n}\Big| \geqslant \varepsilon\Big) \leqslant \frac{\sum_{j=1}^n \operatorname{Var}(X_j)}{\alpha_n^2 \varepsilon^2} \to 0 \quad \text{as } n \to \infty. \tag{1.14.6}$$

In particular, if the sequence (X_n) is even i.i.d., then it satisfies a weak law of large numbers.

Proof. We set

$$Y_n := \frac{\sum_{j=1}^n (X_j - \mathbb{E}[X_j])}{\alpha_n}$$

and apply Bienaymé's formula from Corollary 1.9.25 in combination with (1.9.13) to get the equality in

$$\operatorname{Var}(Y_n) = \frac{1}{\alpha_n^2} \sum_{j=1}^n \operatorname{Var}(X_j) \to 0 \quad \text{as } n \to \infty,$$

where we used (1.14.5) to get the convergence. Plugging this bound into Chebyshev's inequality of Corollary 1.12.2, we obtain

$$\mathbb{P}(|Y_n| \geqslant \varepsilon) \leqslant \frac{\operatorname{Var}(Y_n)}{\varepsilon^2} = \frac{\sum_{j=1}^n \operatorname{Var}(X_j)}{\alpha_n^2 \varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

This establishes (1.14.6).
If the (X_n) are in addition i.i.d., then

$$\sum_{j=1}^n \operatorname{Var}(X_j) = n \operatorname{Var}(X_1)$$

with $\operatorname{Var}(X_1) < \infty$, so (1.14.5) holds true with any α_n such that $\alpha_n/\sqrt{n} \to \infty$. In particular, this is the case for $\alpha_n = n$, and thus the validity of (1.14.6) implies the law of large numbers for the sequence (X_n). and hence finishes the proof. □

Example 1.14.7. *Let a sequence the (X_n) as in Definition 1.14.1 of simple random walk be given. Then the sequence (X_n) satisfies a weak law of large numbers.*
Indeed, by assumption the (X_j) are centered, i.i.d., and in $\mathcal{L}^2$. Thus, the assumptions of the last part of Theorem 1.14.6 are satisfied and in particular

$$\frac{1}{n} \sum_{j=1}^n X_j \xrightarrow{\mathbb{P}} 0,$$

so the sequence (X_n) satisfies a weak law of large numbers.

Example 1.14.8 (Monte-Carlo integration)**.** *Assume we want to evaluate the integral*

$$\int_0^1 f(x)\,\mathrm{d}x,$$

where $f : [0,1] \to \mathbb{R}$ is a sufficiently nice (say continuous) function. If we now choose an i.i.d. sequence (X_n) or random variables with X_n being uniformly distributed on $[0,1]$, then we have

$$\int_0^1 f(x)\,\mathrm{d}x = \mathbb{E}[f(X_1)],$$

and hence the weak law of large numbers Theorem 1.14.6 supplies us with

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n} f(X_j) - \int_0^1 f(x)\,\mathrm{d}x\Big| \geqslant \varepsilon\Big) \leqslant \frac{n\,\mathrm{Var}(f(X_1))}{n^2\varepsilon^2} = \frac{\mathrm{Var}(f(X_1))}{n\varepsilon^2} \to 0, \quad \text{as } n \to \infty. \tag{1.14.7}$$

In this sense, $\frac{1}{n}\sum_{j=1}^{n} f(X_j)$ becomes a better and better approximation of $\int_0^1 f(x)\,\mathrm{d}x$ as $n \to \infty$.
Monte-Carlo integration is particularly useful in high dimensions. The reason for this is that naïve deterministic methods usually result in error bounds that grow exponentially in the dimension, whereas (1.14.7) *is independent of the dimension. There is a whole branch of research dedicated entirely to Monte-Carlo integration and improving Monte-Carlo algorithms leading to more sophisticated algorithms such as e.g. 'importance sampling' or 'Quasi Monte Carlo methods'.*

It occurs quite frequently in probability theory that triangular arrays $(X_{n,k})$, $1 \leqslant k \leqslant n$, of random variables play an important role. In this setting we get the following generalization of Theorem 1.14.6.

Theorem 1.14.9. *Let $(X_{n,k})$, $1 \leqslant k \leqslant n$, $n \in \mathbb{N}$ be a triangular array of random variables in $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ such that for each $n \in \mathbb{N}$, the random variables $X_{n,1}, \ldots, X_{n,n}$ are pairwise uncorrelated. Furthermore, let (α_n) be a sequence of real numbers such that setting*

$$S_n := \sum_{j=1}^{n} X_{n,j},$$

we have that

$$\frac{\mathrm{Var}(S_n)}{\alpha_n^2} \to 0. \tag{1.14.8}$$

Then

$$\frac{S_n - \mathbb{E}[S_n]}{\alpha_n} \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty.$$

This result will e.g. be useful in Homework sheet 9, where you find another problem concerning the Pinana stickers.

Proof of Theorem 1.14.9. Applying Chebyshev's inequality to the elements of the sequence of random variables

$$\frac{S_n - \mathbb{E}[S_n]}{\alpha_n}, \quad n \in \mathbb{N},$$

we obtain for $\varepsilon > 0$ arbitrary that

$$\mathbb{P}\Big(\Big|\frac{S_n - \mathbb{E}[S_n]}{\alpha_n}\Big| \geqslant \varepsilon\Big) \leqslant \frac{\mathrm{Var}(S_n)}{\alpha_n^2\varepsilon^2} \to 0 \quad \text{as } n \to \infty,$$

where the convergence follows from (1.14.8). □

1.14.2 Strong law of large numbers

Definition 1.14.10. *A sequence (X_n) of elements of $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ satisfies* the strong law of large numbers *if*

$$\mathbb{P}\Big(\limsup_{n\to\infty}\Big|\frac{1}{n}\sum_{j=1}^{n}\big(X_j - \mathbb{E}[X_j]\big)\Big| = 0\Big) = 1,$$

which is the same as saying that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\big(X_j - \mathbb{E}[X_j]\big) = 0 \quad \mathbb{P}-a.s.$$

Theorem 1.14.11 (Strong law of large numbers)**.** *Let (X_n) be a sequence of independent identically distributed random variables in $\mathcal{L}^4(\Omega,\mathcal{F},\mathbb{P})$. Then (X_n) satisfies a strong law of large numbers.*

Proof. Possibly replacing X_i by $X_i - \mathbb{E}[X_i]$ we can assume without loss of generality that $\mathbb{E}[X_i] = 0$. Setting $S_n := \sum_{i=1}^n X_i$, according to Theorem 1.13.1 (d) it is sufficient to show that for each $\varepsilon > 0$ we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|n^{-1}S_n| \geqslant \varepsilon) < \infty. \tag{1.14.9}$$

For this purpose, we apply Markov's inequality with the function $\varphi(x) = x^4$, which entails

$$\mathbb{P}(|n^{-1}S_n| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[n^{-4}S_n^4]}{\varepsilon^4}. \tag{1.14.10}$$

Now

$$\mathbb{E}[S_n^4] = \sum_{1\leqslant i,j,k,l\leqslant n} \mathbb{E}[X_iX_jX_kX_l].$$

Using that the (X_n) are independent we deduce that $\mathbb{E}[X_iX_jX_kX_l]$ can be non-zero only if each of the indices i,j,k,l appears at least twice among i,j,k,l. We can therefore continue the above equality to get

$$\mathbb{E}[S_n^4] = \sum_{i=1}^n \mathbb{E}[X_i^4] + 2\sum_{\substack{i,j=1\\ i\neq j}}^n \mathbb{E}[X_i^2X_j^2] = n\mathbb{E}[X_1^4] + 6n(n-1)\mathbb{E}[X_1^2]^2.$$

Plugging this into (1.14.10) we get

$$\mathbb{P}(|n^{-1}S_n| \geqslant \varepsilon) \leqslant \frac{n\mathbb{E}[X_1^4] + 6n(n-1)\mathbb{E}[X_1^2]^2}{n^4\varepsilon^4},$$

which is summable over $n \in \mathbb{N}$ since $\mathbb{E}[X_1^2], \mathbb{E}[X_1^4] < \infty$. Therefore, (1.14.9) follows which finishes the proof. □

Remark 1.14.12. *(a) The implications of Theorem 1.14.11 also hold if we replace the condition $X \in \mathcal{L}^4(\Omega,\mathcal{F},\mathbb{P})$ by $X \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$. This has been proven by Etemadi [Ete81]; the proof is elementary and you should feel encouraged to read it (the article is available online through the university network at* `http://link.springer.com/article/10.1007%2FBF01013465`*)*

(b) As the name suggests, if (X_n) satisfies a strong law of large numbers, it also satisfies a weak law of large numbers. This is a direct consequence of Theorem 1.13.1 (a) applied to the sequence $(n^{-1}\sum_{i=1}^n X_i)$ of random variables and where the limiting random variable in Theorem 1.13.1 (a) is given by the constant 0*.*

The converse is not necessarily true, as you will be asked to show in the homework problems.

Example 1.14.13. *(a) Let $p \in (1/2, 1)$ and let (X_n) be a sequence of independent identically distributed random variables on some probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that*

$$\mathbb{P}(X_1 = 1) = p, \quad \mathbb{P}(X_1 = -1) = 1 - p.$$

Then $S_n := \sum_{i=1}^n X_i$ is called a random walk with drift*.*

Since obviously $X_n \in \mathcal{L}^4(\Omega,\mathcal{F},\mathbb{P})$ for all $n \in \mathbb{N}$, the strong law of large of large numbers supplies us with

$$\lim_{n\to\infty} \frac{1}{n}S_n - \underbrace{\mathbb{E}[X_1]}_{=2p-1} = 0 \quad \mathbb{P}-a.s.,$$

so

$$\lim_{n\to\infty} \frac{1}{n}S_n = 2p - 1 > 0 \quad \mathbb{P}-a.s.$$

Therefore, we have '$S_n \to \infty$' $\mathbb{P}$-a.s., or, more formally, $\mathbb{P}(\liminf_{n\to\infty} S_n \geqslant M) = 1$ for all $M \in \mathbb{R}$.

(b) Consider a sequence (X_n) of discrete i.i.d. random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in the finite set E such that $\pi(x) := \mathbb{P}(X_1 = x) > 0$ for all $x \in E$. Then π is the law of X_1 on E, and the common interpretation is that (X_n) describes a sequence of random signals taking values in the alphabet E.

Due to the independence of the X_n, for any $\omega \in \Omega$, the probability of having observed $X_1(\omega), \dots, X_n(\omega)$ in the first n trials is given by

$$\prod_{j=1}^{n} \pi(X_j(\omega)).$$

Taking logarithms and applying the strong law of large numbers we get that $\mathbb{P}$-a.s.,

$$-\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \ln(\pi(X_j(\omega))) = -\mathbb{E}[\ln(\pi(X_1))] = -\sum_{x\in E} \pi(x) \ln \pi(x),$$

where we took advantage of the fact that the $\ln(\pi(X_n))$ form an independent family of random variables. The right-hand side of this expression is called the entropy of the probability measure π*, and it is a measure for the amount of information or complexity that is inherent to the distribution π.*

1.14.3 Are we investigating unicorns?

In the previous sections we have investigated the asymptotic behaviour of (infinite) sequences of independent (and sometimes identically distributed) random variables in depth. In particular, it seemed like we took for granted the fact that there *exist* such sequences in the first place. Is this a fair thing to do?
In Example 1.6.6 we have seen that a finite number of independent experiments can be modelled on a corresponding product space (in that setting, if we consider the n-fold product, one can choose X_k to be a random variable depending on the k-th coordinate of $\Omega_1 \times \ldots \times \Omega_n$ only, and in this case the random variables would be independent random variables on the product space given in (1.6.6)).
A self-evident choice for modeling an infinite number of independent experiments (such as independent rolls of a fair die) would be to try considering a corresponding infinite product space. However, it turns out that we run into difficulties defining a probability measure on an infinite space like $\{1,2,3,4,5,6\}^{\mathbb{N}}$ (endowed with a suitable σ-algebra) in such a way that for $\omega = (\omega_1, \omega_2, \ldots) \in \{1,2,3,4,5,6\}^{\mathbb{N}}$ the coordinates projections $X_j(\omega) = \omega_j$ are independent random variables and give the right probabilities (recall that this was not hard to accomplish in the case of finitely many die rolls). This issue will only be formally and positively resolved in the lecture 'Probability Theory I'. For the time being we tacitly assume that there are probability spaces on which we can find an infinite sequence of independent random variables, each of which is distributed according to a given distribution.

1.15 Central limit theorem

Besides the laws of large numbers from the previous section, the central limit theorem is the second main result of this course. On the one hand, it provides more precise information of the fluctuations of the sum of well-behaved random variables than the results we know from the laws of large numbers. On the other hand, it plays an important role in statistics since it justifies using the normal distribution in many models. This exemplifies how it has been given its name, which in fact is due to the 'central' role it plays in probability theory.
To motivate the central limit theorem, let us get back to (1.14.2) where we had shown that for simple random walk S_n,

$$\mathbb{P}(S_{2k} = 0) \sim \frac{1}{\sqrt{k\pi}}.$$

In fact, in this setting it is not hard to show that not only the probability of finding simple random walk in 0 at time $2k$ has a square root decay in k, but also the probabilities of finding simple random walk at a distance of order $\sqrt{k}$ at time $2k$ (we restrict ourselves to even times for simplicity). Indeed, we obtain

for any constant $c \in \mathbb{R}\backslash\{0\}$, setting $c_k := \lfloor c\sqrt{k} \rfloor$ for brevity, that

$$\begin{aligned}
\mathbb{P}(S_{2k} = 2c_k) &= \binom{2k}{k+c_k} 2^{-2k} \\
&\sim \frac{(k/e)^{2k}\sqrt{2\pi \cdot 2k}}{(k+c_k)^{k+c_k}(k-c_k)^{k-c_k}\sqrt{4\pi^2(k+c_k)(k-c_k)}} \sqrt{\frac{k}{\pi(k+c_k)(k-c_k)}} \\
&= \exp\Big\{2k\ln k - \underbrace{(k+c_k)\ln(k+c_k)}_{=(k+c_k)(\ln k + \ln(1+c_k/k))} - (k-c_k)\ln(k-c_k)\Big\} \sqrt{\frac{1}{\pi k(1-c_k^2/k^2)}} \\
&= \underbrace{\exp\Big\{-c_k(\ln k + \ln(1+c_k/k)) - k\ln(1+c_k/k) + c_k(\ln k + \ln(1-c_k/k)) + k\ln(1-c_k/k)\Big\}}_{\sim \widetilde{c} \in (0,\infty)} \\
&\quad \times \sqrt{\frac{1}{\pi k(1-c_k^2/k^2)}} \sim \widetilde{c}\sqrt{\frac{1}{\pi k}},
\end{aligned}$$

where we used Stirling's formula in the second line to obtain the first asymptotics.
Thus, we have shown that, at least for simple random walk, we have the same order of decay for all probabilities of the form $\mathbb{P}(2k = c_k)$ with c_k as above. As a consequence, if we are looking for a rescaling of S_n by some scale function $\varphi(n)$ such that $S_n/\varphi(n)$ might converge in distribution to a non-trivial limiting distribution, then the above suggests that $\sqrt{n}$ should be the correct order of $\varphi(n)$.
Yet another motivation for the central limit theorem can be derived from the laws of large numbers: From those we know that under suitable assumptions on a sequence of i.i.d. random variables we have

$$\lim_{n\to\infty} \Big(\frac{1}{n} S_n - \mathbb{E}[X_1]\Big) = 0.$$

To obtain information on a finer scale than in the central limit theorem we can now ask if there exists an exponent $\beta \in (0,\infty)$ such that the sequence $n^\beta(\frac{1}{n}S_n - \mathbb{E}[X_1])$ might hopefully converge to a non-trivial limiting random variable instead of 0. The first motivational thread via the investigation of simple random walk then suggests that $\beta = 1/2$. Indeed, this always has to be the case as long as the X_n are assumed to have finite variance since due to Bienaymé's formula we have

$$\operatorname{Var}\Big(n^\beta(n^{-1}S_n - \mathbb{E}[X_1])\Big) = n^{2\beta}\frac{1}{n^2}n = n^{2\beta-1},$$

which can only converge to a non-trivial limit if $\beta = \frac{1}{2}$.
While the central limit theorem will not give us any information on probabilities of finding e.g. simple random walk at single points, it does indeed imply that the right scale for rescaling is $\sqrt{n}$; and not only does it do so for simple random walk, but for a very general class of distributions.

Theorem 1.15.1 (Central limit theorem)**.** *Let a sequence (X_n) of independent identically distributed random variables with expectation $\mu = \mathbb{E}[X_1]$ and finite positive variance $\sigma^2 := \operatorname{Var}(X_1)$ be given. Then the sequence of random variables defined via*

$$Y_n := \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{\sigma^2 n}}, \quad n \in \mathbb{N}, \tag{1.15.1}$$

converges in distribution to a $\mathcal{N}(0,1)$ distributed random variable, i.e., for any $t \in \mathbb{R}$, (recall (1.8.7))

$$\mathbb{P}(Y_n \leqslant t) \to \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2}\,\mathrm{d}x \quad \text{as } n \to \infty.$$

There are at least two essentially different strategies to prove the central limit theorem. The first one is a more or less self-contained and direct proof taking advantage of the characterization of convergence in distribution given in Definition 1.11.8. A proof along these lines can e.g. be found as the proof of [Geo09, Theorem 5.28]. The second one uses the technique of characteristic functions. It has the disadvantage that for us it is less self-contained; it is, however, more robust under variations of the very setting given in Theorem 1.15.1 and can be generalized without too much effort to more general situations, such as different state spaces or dependencies between the random variables X_n. We refer to [Kle14, Section 15.5] for further detail.

Remark 1.15.2. *(a) The Y_n are shifted and rescaled in such a way that $\mathbb{E}[Y_n] = 0$ and $\mathrm{Var}(Y_n) = 1$, so expectation and variance already coincide with those of a $\mathcal{N}(0,1)$-distributed variable.*

(b) It is surprising that, as long as the X_n have finite second moments the limiting distribution is the normal distribution, independent of the specific distribution of the X_is. This type of phenomenon is also called universality (of the normal distribution)*.*

(c) There is a plethora of other, more general conditions which imply the validity (1.15.1)*. In particular, similarly to the case of the weak law of large numbers Theorem 1.14.9, there is a version of the central limit theorem for triangular arrays as well.*

(d) The finiteness of the second moment is in fact essential in Theorem 1.15.1. If it is not assumed, however, then one can still obtain other types of convergence results to non-trivial distributions (so-called α-stable distributions) for different rescalings than the division by $\sqrt{n}$ in (1.15.1)*.*

(e) It can be shown that the strongest type of convergence in Theorem 1.15.1 indeed is already given by convergence in distribution (and not e.g., in probability).

Almost sure convergence can be excluded setting $A_n := \{Y_n \geqslant 1\}$ and $B_n := \{Y_n \leqslant -1\}$. Then $\mathbb{P}(A_n), \mathbb{P}(B_n) \to \Phi(-1) > 0$, so the first part of Borel Cantelli yields

Exercise 1.15.3. *For a sequence of random variables (X_n) as in the assumption of Theorem 1.15.1, the central limit theorem implies the validity of a weak law of large numbers for (X_n).*
Indeed, since the distribution function Φ of the standard normal distribution (see (1.8.7)*) is continuous, Theorem 1.15.1 implies that for arbitrary $M > 0$ we have*

$$\mathbb{P}\Big(\underbrace{\frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{\sigma^2 n}} \notin (-M, M]}_{A_{n,M}} \Big) \to \Phi(-M) + (1 - \Phi(M)). \tag{1.15.2}$$

Now for any $M \in (0,\infty)$ and $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ one has

$$B_{n,\varepsilon} := \Big\{ \Big| \frac{\sum_{i=1}^n (X_i - \mu)}{n} \Big| > \varepsilon \Big\} \subset A_{n,M}.$$

As a consequence, we obtain for any such M and ε, in combination with (1.15.2)*, that*

$$\limsup_{n\to\infty} \mathbb{P}(B_{n,\varepsilon}) \leqslant \Phi(-M) + (1 - \Phi(M)).$$

Since M was arbitrary and $\lim_{M\to\infty} \Phi(-M) + (1 - \Phi(M)) = 0$, this implies

$$\lim_{n\to\infty} \mathbb{P}(B_{n,\varepsilon}) = 0.$$

As in addition $\varepsilon > 0$ was arbitrary, this implies the desired weak law of large numbers for (X_n).

Example 1.15.4. *Using the (e.g., weak) law of large numbers, we have seen in Example 1.14.13 that for a random walk with drift (i.e., $S_n = \sum_{j=1}^n X_j$ where the X_j are i.i.d. with $\mathbb{P}(X_1 = 1) = p$, $\mathbb{P}(X_1 = -1) = 1 - p$, and $p \in (1/2, 1)$) one has that for all $\varepsilon > 0$,*

$$\mathbb{P}(|S_n - n(2p-1)| \geqslant n\varepsilon) \to 0.$$

Therefore, the first order (i.e. linear in n) term of the position of S_n at time n will asymptotically be given by $2p - 1$. In order to obtain a better understanding, it is of course tempting to ask for the lower order corrections. For this purpose we apply the central limit theorem; using that the variance of X_n is given by

$$\mathrm{Var}(X_n) = \mathbb{E}[X_n^2] - \mathbb{E}[X_n]^2 = 1 - (2p-1)^2 = 1 - 4p^2 + 4p - 1 = 4p(1-p) := \sigma^2$$

we obtain

$$\frac{S_n - n(2p-1)}{\sqrt{\sigma^2 n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

In particular, this implies that the 'typical' fluctuations of S_n around its expected value $n(2p-1)$ are of the order $\sqrt{n}$.

1.16 Markov chains

So far we have mainly been focusing on the investigation of sequences of independent random variables (with the prime examples of independent dice rolls or coin flips). It turns out, however, that in order to model slightly more complicated situations one should admit at least a certain degree of dependence among random variables. As an example, consider a very simple stochastic model of weather forecasting. While experience probably tells you that modeling whether a day will be sunny or cloudy by independent (possibly biased) coin flips seems too simplistic, another approach might be to say that tomorrow has a higher probability to be sunny than cloudy if we already know that today is sunny.
Alternatively, assume that each minute you were scanning the traffic situation at a certain part of a highway. There could be three states, namely 'traffic jam', 'slow-moving traffic', and 'well-moving traffic'. Again, experience tells us that, empirically, if there is a traffic jam at present, then in the next minute the probability of still seeing a traffic jam is higher than if there was well-moving traffic at present, and similarly for the other possible states of the road. In particular, it does not seem to be a good idea to try to model the states of the highway using i.i.d. random variables.
This exemplifies of the easiest types of dependence in sequences of random variables, and this type of dependence is so important that this class of sequences has been given a name on its own, namely the class of *Markov chains*. Intuitively, a sequence of random variables is called a Markov chain (or is said to have the Markov property) if, at any time n, the transition probabilities for its state at time $n+1$ only depend on the value of X_n, and neither on the values of $X_1, \dots, X_{n-1}$ nor on the specific value of n (the latter property is also referred to as *homogeneity in time*).
In what follows for a finite or countable set S we will denote by $\mathbb{R}^{S\times S}$ the set of real matrices whose rows and columns are indexed by the elements of S.

Definition 1.16.1. *Let S be a finite or countable set. A sequence of random variables (X_n) on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a* (time homogeneous) Markov chain with state space S and transition matrix $P = (P(x,y))_{x,y\in S}$ *if for all $n \in \mathbb{N}_0$ and $x_0, x_1, \dots, x_{n+1} \in S$, one has*

$$\begin{aligned} P(x_n, x_{n+1}) &= \mathbb{P}(X_{n+1} = x_{n+1} \,|\, X_n = x_n) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} \,|\, X_0 = x_0,\, X_1 = x_1, \dots, X_n = x_n), \end{aligned} \tag{1.16.1}$$

whenever $\mathbb{P}(X_0 = x_0,\, X_1 = x_1, \dots, X_n = x_n) > 0$.

Thus, the probability that a (time homogeneous) Markov chain jumps from x to y from time n to time $n+1$ depends only on x and y and is given by the entry $P(x,y)$ of the transition matrix.

Definition 1.16.2. *A real $n \times n$ matrix $P \in \mathbb{R}^{S\times S}$ with the following property is called* stochastic*: For each $x \in S$, one has that $P(x,\cdot)$ is a probability distribution on $(S, 2^S)$, i.e., $P(x,y) \geqslant 0$ for all $y \in S$, and*

$$\sum_{y\in S} P(x,y) = 1.$$

In particular, transition matrices are stochastic matrices. Vice versa, for any stochastic matrix $P \in \mathbb{R}^{S\times S}$ one can define a Markov chain (X_n) that has transition matrix P.

Example 1.16.3. *(a) Simple random walk on $\mathbb{Z}$ (see Definition 1.14.1) and random walk with drift $\mathbb{Z}$ (see Example 1.14.13) are both Markov chains with state space $S = \mathbb{Z}$. Their transition matrices are given by*

$$P(x,y) := \begin{cases} \frac{1}{2}, & \text{for each } y \in \mathbb{Z} \text{ with } |x-y| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

and

$$P(x,y) := \begin{cases} p, & \text{if } y = x+1, \\ 1-p, & \text{if } y = x-1, \\ 0, & \text{otherwise.} \end{cases}$$

respectively.

(b) Random walks on graphs:

A finite (undirected) graph $G = (V, E)$ *consists of a finite set of vertices* V *and a set of edges* $E \subseteq \{\{x, y\} : x, y \in V\}$. *We say that* $x, y \in V$ *are* neighbors *if* $\{x, y\} \in E$, *and in this case one usually writes* $x \sim y$. *The* degree $d(x)$ *of a vertex* $x \in V$ *is defined as the number of its neighbors*

$$d(x) := |\{y \in V : x \sim y\}|.$$

Assume $d(x) < \infty$ *for all* $x \in V$. *Then* simple random walk on G *is the Markov chain with state space* $S = V$ *and transition probabilities given by*

$$P(x, y) := \begin{cases} \frac{1}{d(x)}, & \text{if } y \sim x, \\ 0, & \text{otherwise.} \end{cases}$$

I.e., when at vertex x *at time* n, *the probability of its position at time* $n+1$ *is distributed uniformly over the neighbors of* y.

A specific example is the simple random walk on $\mathbb{Z}$ *that we had introduced in Example 1.14.1, but much more is covered by this general setting, such as simple random walk on the complete graph of* $n \in \mathbb{N}$ *vertices, which is the graph consisting of a state space* S *with* $|S| = n$ *and edges between any two vertices, i.e., the edge set is* $\{\{x, y\} : x, y \in V\}$.

In the context of Markov chains it will frequently turn out useful to apply tools from linear algebra. Therefore, probability measures on $(S, 2^S)$ will oftentimes be interpreted as possibly infinite row (or column) vectors (called *probability vectors*), and similarly we will switch between interpreting elements of $\mathbb{R}^S$ as (row / column) vectors and as functions from S to $\mathbb{R}$ without further mentioning. In particular, for the distribution of a Markov chain (X_n) at time n we introduce the shorthand notation $\mu_n(x)_{x \in S} \in [0, 1]^S$ via

$$\mu_n(x) := \mathbb{P}(X_n = x), \quad \forall x \in S,\, n \in \mathbb{N}_0. \tag{1.16.2}$$

A priori, given the setting of Definition 1.16.1, the *initial distribution* of the Markov chain, i.e., its distribution at time 0, is given by the distribution of X_0 and thus equal to $\mu_0(x)$, $x \in S$. The probability that the chain is in state y at time 1 is then given by

$$\begin{aligned} \mu_1(y) = \mathbb{P}(X_1 = y) = \sum_{x \in S} \mathbb{P}(X_1 = y,\, X_0 = x) &= \sum_{\substack{x \in S \\ \mathbb{P}(X_0 = x) > 0}} \mathbb{P}(X_1 = y \,|\, X_0 = x)\mathbb{P}(X_0 = x) \\ &= \sum_{x \in S} \mu_0(x) P(x, y) = (\mu_0 \cdot P)(y). \end{aligned} \tag{1.16.3}$$

Hence, the entire distribution of the chain at time 1 is given by the vector

$$\mu_1 = \mu_0 \cdot P, \tag{1.16.4}$$

where the right-hand side is the product of a row vector with an element of $\mathbb{R}^{S \times S}$. I.e., *starting from today's distribution of the chain we obtain tomorrow's distribution by multiplying the former to the right by the matrix* P. Inductively we get

$$\mu_n = \mu_0 \cdot P^n, \tag{1.16.5}$$

where, P^n denotes the n-th power of the matrix P.

Remark 1.16.4. *We should note here that even in the case of* S *being infinite (in which case it will be countable due to our previous assumptions), the product of two stochastic matrices* $P, Q \in \mathbb{R}^{S \times S}$ *is well-defined in the same simple way it is defined in linear algebra: For any* $x, y \in S$ *we set*

$$(P \cdot Q)(x, y) := \sum_{z \in S} P(x, z) Q(z, y), \tag{1.16.6}$$

and in this case the right-hand side is a well-defined since the sum is over non-negative numbers only.

Exercise 1.16.5. *If* $P, Q \in \mathbb{R}^{S \times S}$ *are stochastic matrices, then* $P \cdot Q$ *again is a stochastic* $\mathbb{R}^{S \times S}$ *matrix. Indeed, it is obvious that as a sum over non-negative terms,* $(P \cdot Q)(x, y)$ *is non-negative again. In addition, since both* P *and* Q *are stochastic matrices, we have for the row sum of the row indexed by* x *that*

$$\sum_{y \in S} (P \cdot Q)(x, y) = \sum_{y \in S} \sum_{z \in S} P(x, z) Q(z, y) = \sum_{z \in S} P(x, z) \underbrace{\sum_{y \in S} Q(z, y)}_{=1} = 1,$$

where the sums could be interchanged since all summands are non-negative.

In what comes below, we will separate the initial distribution μ_0 of the Markov chain from the transition dynamics induced by the transition matrix P. Therefore, if we want to stress the initial distribution, we will use the notation $\mathbb{P}_{\mu_0}$ for $\mathbb{P}$. Similarly, for any time $n \in \mathbb{N}_0$ and any function $f : S \mapsto \mathbb{R}$ we write $\mathbb{E}_{\mu_0}$ for $\mathbb{E}$ in order to stress the initial distribution:

$$\mathbb{E}_{\mu_0}[f(X_n)] = \mathbb{E}[f(X_n)] = \sum_{y \in S} f(y)\mathbb{P}(X_n = y) = \sum_{y \in S} f(y)\mu_n(y)$$

as long as

$$\text{either} \quad \sum_{y \in S} (f(y) \vee 0)\mathbb{P}(X_n = y) < \infty \quad \text{or} \quad \sum_{y \in S} (f(y) \wedge 0)\mathbb{P}(X_n = y) > -\infty \tag{1.16.7}$$

(which guarantees that the right-hand side is well-defined in $[-\infty, \infty]$). In this case, bearing in mind that $X_n(\Omega) \subset S$ and taking advantage of (1.16.5), we can rewrite the expectation as

$$\mathbb{E}_{\mu_0}[f(X_n)] = \sum_{y \in S} (\mu_0 P^n)(y) \cdot f(y) = \mu_0 \cdot P^n \cdot f, \tag{1.16.8}$$

where all multiplications appearing on the right-hand side are to be interpreted as matrix multiplications. The nice thing about the formula on the right-hand side of (1.16.8) is that we have clearly separated the influence of the three parameters initial distribution μ_0, transition dynamics P, and the functional f. Hence, in the same vein as multiplication of a probability distribution μ on $(S, 2^S)$ from the right by P gives the distribution of the Markov chain after one time step with initial distribution μ (cf. (1.16.4)), formula (1.16.8) can be interpreted in the way that multiplication of a function $f : S \to \mathbb{R}$ (considered as a vector in $\mathbb{R}^S$) from the left by P corresponds to taking the expectation of $f(X_1)$.
In particular, although we have noted that in the context of Definition 1.16.1, the initial distribution of a Markov chain is already characterized by the distribution of X_0, using (1.16.8) we can think of a Markov chain started under any arbitrary initial probability distribution $\widetilde{\mu}_0$ on S by demanding that its distribution at time n be given by $\widetilde{\mu}_0 \cdot P^n$.
Indeed, we would then want to have

$$\mathbb{P}(X_i = x_i \,\forall\, 0 \leqslant i \leqslant n) = \widetilde{\mu}_0(x_0) \prod_{i=1}^{n} P(x_{i-1}x_i). \tag{1.16.9}$$

We do not have the tools yet to rigorously deduce the existence of a sequence of random variables (X_n) which has the desired properties (in the sense that it constitutes a Markov chain with the desired initial distribution and transition dynamics); for doing so, we refer to the theorem of Ionescu-Tulcea or, more generally, Kolmogorov's existence and uniqueness theorem, which is to be covered in probability theory I.

Remark 1.16.6. *The finite dimensional distributions of a Markov chain are the distributions of all the vectors of discrete random variables $(X_{i_1}, X_{i_2}, \ldots, X_{i_n})$ taking values in $(S^n, 2^{S^n})$, for $n \in \mathbb{N}$ and $1 \leqslant i_1 < i_2 < \ldots < i_n$ with $i_j \in \mathbb{N}$ for all $1 \leqslant j \leqslant n$.*
Then, due to (1.16.9)*, the finite dimensional distributions are indeed completely determined by the initial distribution $\widetilde{\mu}_0$ and the transition matrix P.*

Nevertheless, we note the following without proof.

Claim 1.16.7. *Let $S \neq \varnothing$ be finite or countable. Then there is a bijection between the set*

$$\Big\{(\mu, P) \,:\, \mu \text{ a distribution on } S, \text{ and } P \text{ a stochastic } S \times S \text{ matrix}\Big\},$$

and the set of distributions of time homogeneous Markov chains with state space S.

As a shorthand in the context above, if $\mu_0 = \delta_x$ (recall the Dirac measure introduced in Definition 1.3.8), then we write $\mathbb{P}_x$ and $\mathbb{E}_x$ as shorthands.

1.16.1 Stationary distribution

Definition 1.16.8. *A probability distribution π on $(S, 2^S)$ is called a* stationary distribution / invariant distribution / equilibrium distribution / steady state *for a transition matrix P if one has*

$$\pi = \pi \cdot P. \tag{1.16.10}$$

(i.e., if π is a left eigenvector of P with eigenvalue 1)

This means that starting a Markov chain with transition matrix P from its stationary distributions $\mu_0 := \pi$, one has $\mu_n = \pi$ for all $n \in \mathbb{N}_0$.

Remark 1.16.9. *Note that a real m by n matrix M has left eigenvector v and corresponding eigenvalue λ if and only if M^T has right eigenvector v^T and eigenvalue λ.*

Stationary distributions are a very important topic in Markov chains: Under reasonable assumptions, the Markov chain converges towards its (unique?) stationary distribution, and it is interesting and important to understand how fast this convergence happens (think of how often you will have to shuffle a deck of cards in order for the cards to be distributed 'sufficiently random').

Example 1.16.10. *(a) For random walk on a finite graph (V, E) as introduced in Example 1.16.3 (b), a stationary distribution is given by*

$$\pi(x) := \frac{d(x)}{2|E|}. \tag{1.16.11}$$

Indeed, for any $y \in S$ we have

$$(\pi \cdot P)(y) = \sum_{x \in V} \pi(x) \cdot P(x,y) = \sum_{x \in V,\, \{x,y\} \in E} \frac{d(x)}{2|E|} \cdot \frac{1}{d(x)} = \frac{1}{2|E|} \underbrace{\sum_{x \in V,\, \{x,y\} \in E} 1}_{=d(y)} = \frac{d(y)}{2|E|} = \pi(y).$$

In addition we have that π has mass 1, i.e.,

$$\sum_{x \in V} \pi(x) = \sum_{x \in V} \frac{d(x)}{2|E|} = \frac{1}{2|E|} \sum_{x \in V} d(x) = 1,$$

where the last equality follows since any edge is counted twice in the sum over $\sum_{x \in V}$ due to the fact that each of the two vertices it contains is counted.

(b) For $d \in \mathbb{N}$, d-dimensional simple random walk on $\mathbb{Z}^d$ is defined as the sequence $S_n = \sum_{j=1}^n X_j$, where the X_j are i.i.d. with

$$\mathbb{P}(X_1 = e) = \begin{cases} \frac{1}{2d}, & \text{if } \|e\|_1 = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Show that d-dimensional simple random walk on $\mathbb{Z}^d$ and random walk with drift on $\mathbb{Z}$ do not *have a stationary distribution.*

Show that, nevertheless, defining the counting measure π on $(\mathbb{Z}^d, 2^{\mathbb{Z}^d})$ via $\pi(A) := |A|$ for $A \in 2^{\mathbb{Z}^d}$ we have

$$\pi P = \pi,$$

if P denotes the transition matrix of random walk with drift on $\mathbb{Z}^d$. This shows that, although there might not exist a stationary distribution, one might still be able to find a so-called stationary measure *which then has infinite mass.*

We start with showing that a stationary distribution does not exist (in the case of simple random walk). For purpose, assume to the contrary that π is a stationary distribution for simple random walk on $\mathbb{Z}^d$. In particular, since π is non-negative and

$$\sum_{x \in \mathbb{Z}^d} \pi(x) = 1,$$

we deduce that $M := \max_{x \in \mathbb{Z}^d} \pi(x) \in (0, 1]$ exists. We can then find $x_M \in \mathbb{Z}^d$ such that

(a) $\pi(x_M) = M$, and

(b) there exists $y \in \mathbb{Z}^d$ with $\|x_M - y\|_1 = 1$ with $\pi(y) < \pi(x_M)$.

As a consequence,

$$(\pi P)(x_M) = \sum_{z \in \mathbb{Z}^d \,:\, \|z - x_M\|_1 = 1} \frac{1}{2d} \pi(z) < \pi(x_M).$$

In particular, π cannot be a stationary distribution for simple random walk on $\mathbb{Z}^d$.

On the other hand, denoting by π the counting measure on $\mathbb{Z}^d$ we deduce for any $y \in \mathbb{Z}^d$ that

$$(\pi \cdot P)(y) = \sum_{x \in \mathbb{Z}^d \,:\, \|x-y\|_1 = 1} \frac{1}{2d} \pi(x) = 1,$$

which equals $\pi(y) = 1$. Thus, at least we have $\pi = \pi \cdot P$, and π is a stationary measure for simple random walk on $\mathbb{Z}^d$.

For the case of random walk with a drift the corresponding results are proven in a similar fashion.

(c) For $p \in (0,1)$ and $N \in \mathbb{N}$, random walk with absorption on $\{0, 1, \ldots, N\}$ (and drift) *is defined as the Markov chain with transition matrix*

$$P = \begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 1-p & 0 & p \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{pmatrix}$$

One interpretation of this Markov chain is as follows. At time 0, two players A and B have a capital of $C_A, C_B \in \mathbb{N}$ such that $C_A + C_B = N$. They then start playing a sequence of (independent) matches that player A wins with probability p and loses with probability $1-p$ (hence, the matches are called fair *if $p = 1/2$, and* not fair *otherwise), and the player winning a match pays one unit to her opponent. The sequence stops once one of the two players is bankrupt. If X_n denotes the Markov chain at time n (i.e., the capital of player A after n matches), then, if it hits N before hitting 0, player A has won the sequence of matches, whereas otherwise player B has won the sequence.*

Show that the stationary distributions for random walk with absorption on $\{0, 1, \ldots, N\}$ are exactly the probability measures $\lambda\delta_0 + (1-\lambda)\delta_N$, for any $\lambda \in [0,1]$.

(d) For $p \in (0,1)$ and $N \in \mathbb{N}$, random walk with reflection on $\{0, 1, \ldots, N\}$ (and drift) *is defined as the Markov chain with transition probabilities*

$$P = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 1-p & 0 & p \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}$$

(e) For $p \in (0,1)$ random walk with reflection on $\{0, 1, 2, \ldots\}$ (and drift) *is defined as the Markov chain with transition probabilities given by the infinite matrix*

$$P = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 & \cdots \\ 1-p & 0 & p & 0 & \cdots & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & 0 & 0 & \cdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & \cdots \\ 0 & \cdots & \cdots & 1-p & 0 & p & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

According to (1.16.10)*, a probability vector $\pi \in [0,1]^{\mathbb{N}_0}$ is invariant for P if*

$$\pi \cdot P = \pi,$$

which is equivalent to the system

$$(1-p)\pi(1) = \pi(0) \tag{1.16.12}$$
$$\pi(0) + (1-p)\pi(2) = \pi(1) \tag{1.16.13}$$
$$p\pi(i-1) + (1-p)\pi(i+1) = \pi(i) \qquad \text{for all } i \geqslant 2, \tag{1.16.14}$$

of difference equations, where the last equation can be rewritten as

$$p\pi(i-1) - \pi(i) + (1-p)\pi(i+1) = 0. \tag{1.16.15}$$

We observe that (1.16.12) *implies that*

$$\pi(1) = \frac{\pi(0)}{1-p}, \tag{1.16.16}$$

which inserted in (1.16.13) *yields*

$$\pi(2) = \frac{\pi(1) - \pi(0)}{1-p} = \frac{p\pi(0)}{(1-p)^2}. \tag{1.16.17}$$

Plugging the previous two equations into (1.16.14) *we deduce*

$$\pi(3) = \frac{\pi(2) - p\pi(1)}{1-p} = \frac{p\pi(0) - p(1-p)\pi(0)}{(1-p)^3} = \frac{p^2\pi(0)}{(1-p)^3}. \tag{1.16.18}$$

We can now try to guess the general solution from these considerations and prove it using complete induction.

An alternative approach to solving the above system of equations is to use the Ansatz $\pi(i) = \lambda^i$*, which plugged into* (1.16.15) *yields*

$$p\lambda^{i-1} - \lambda^i + (1-p)\lambda^{i+1} = 0,$$

and, as is done in a similar fashion for linear ordinary differential equations ('gewöhnliche Differentialgleichungen') with constant coefficients, solve the corresponding characteristic polynomial which is obtained from the previous display via division by λ^{i-1} *to get*

$$\lambda \mapsto p - \lambda + (1-p)\lambda^2 = 0$$

for which we deduce the solutions

$$\lambda_1 = 1 \quad \text{and} \quad \lambda_2 = \frac{p}{1-p}.$$

Therefore, we obtain the solutions

$$\pi(i) = C_1 + C_2 i, \qquad \text{if } p = 1-p = \frac{1}{2}, \tag{1.16.19}$$

and

$$\pi(i) = C_1 + C_2\Big(\frac{p}{1-p}\Big)^i, \qquad \text{if } p \neq \frac{1}{2}. \tag{1.16.20}$$

As a consequence we deduce that for $p \geqslant 1/2$ *there exists no stationary (probability) distribution. For* $p \in (0, 1/2)$*, in order to have* $\pi(i) \geqslant 0$ *for all* $i \in \mathbb{N}_0$ *and*

$$\sum_{i \in \mathbb{N}_0} \pi(i) = 1, \tag{1.16.21}$$

we need $C_1 = 0$ *and* $C_2 \geqslant 0$*. In particular, we deduce*

$$\pi(i) = C_2\Big(\frac{p}{1-p}\Big)^i, \quad i \geqslant 1, \tag{1.16.22}$$

which plugged into (1.16.12) *yields*

$$\pi(0) = (1-p)C_2\frac{p}{1-p} = C_2 p. \tag{1.16.23}$$

Thus, (1.16.21) *becomes*

$$1 \overset{!}{=} C_2\Big(p + \sum_{i\geqslant 1}\Big(\frac{p}{1-p}\Big)^i\Big) = C_2\Big(p + \frac{p/(1-p)}{1-p/(1-p)}\Big) = C_2\frac{2p(1-p)}{1-2p},$$

hence

$$C_2 = \frac{1-2p}{2p(1-p)},$$

which in combination with (1.16.22) *and* (1.16.23) *completely determines the stationary distribution.*

The definition of a stationary distribution in (1.16.10) states that a stationary distribution is a fixed point of the right multiplication by P. One might therefore be tempted to hope that the distributions $\mu_n = \mu_0 \cdot P^n$ of the Markov chain at time n could possibly converge towards a stationary distribution of the Markov chain. As a consequence, and since there are numerous applications to this, one of the principal things we will investigate is the question of convergence of the distribution μ_n to the stationary distribution of the Markov chain. It is therefore of course helpful to know criteria which ensure that a Markov chain has a stationary distribution, and also under which condition such a stationary distribution might be unique. One of the easiest results in this context and with least assumptions is the following.

Proposition 1.16.11. *Every Markov chain with finite state space has at least one stationary distribution.*

Proof. Let S denote the state space and

$$K := \Big\{(\pi(x))_{x\in S} \in \mathbb{R}^S \,:\, \pi(x) \geqslant 0 \ \forall x \in S,\ \sum_{x\in S}\pi(x) = 1\Big\},$$

which can be identified with the simplex of probability measures on S (recall Exercise 4 on the second homework sheet). Choose an arbitrary initial distribution μ_0 (which is an element of K by definition), and denote by P the transition matrix of the Markov chain. Then the sequence of its distributions (μ_n) is a sequence whose elements take values in K, and the same applies to the so-called Cesàro means

$$\nu_m := \frac{1}{m}\sum_{n=0}^{m-1}\mu_n.$$

Since K is a compact subset of $\mathbb{R}^S$ (check!), the sequence (ν_m) has an accumulation point $\nu \in K$; i.e., there exists a subsequence ν_{m_k}, $k \in \mathbb{N}_0$, such that $\lim_{k\to\infty}\nu_{m_k} = \nu \in K$. Hence, we get

$$\begin{aligned}\nu \cdot P &= (\lim_{k\to\infty}\nu_{m_k})\cdot P = \lim_{k\to\infty}(\nu_{m_k}\cdot P)\\ &= \lim_{k\to\infty}\Big((\frac{1}{m_k}\sum_{n=0}^{m_k-1}\mu_n)\cdot P\Big) = \lim_{k\to\infty}(\frac{1}{m_k}\sum_{n=0}^{m_k-1}\mu_{n+1}) \overset{\text{check!}}{=} \nu,\end{aligned} \tag{1.16.24}$$

where in the second equality we used that matrix multiplication is a continuous mapping. Thus, ν is an invariant distribution. □

Remark 1.16.12. *The equalities in* (1.16.24) *are of course equalities of vectors, and the limits taken in that display are to be understood as pointwise limits. Observe that this is the same as interpreting the vectors as probability measures (in the same spirit as outlined before* (1.16.2)*) and interpreting the limits of probability measures in* (1.16.24) *as weak limits (recall* (1.11.6)*) if we endow the space S with the metric $d(x,x) = 0$ for all $x \in S$ and $d(x,y) = 1$ if $x \neq y$ (note that this ensures that any function $f : S \to \mathbb{R}$ is continuous).*
Indeed, if (μ_n) is a sequence of vectors in $\mathbb{R}^S$ such that $\mu_n(x) \geqslant 0$ for all $x \in S$ and $\sum_{x\in S}\mu_n(x) = 1$ (and the same for μ) then this implies the following: If $f : S \to \mathbb{R}$ is any bounded (continuous) function, and X_n and X are S-valued random variables distributed according to μ_n and μ, respectively, we have

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[f(X_n)] &= \lim_{n\to\infty}\sum_{x\in S} f(x)\cdot\mathbb{P}(X_n = x) = \sum_{x\in S} f(x)\cdot\lim_{n\to\infty}\mathbb{P}(X_n = x)\\ &= \sum_{x\in S} f(x)\cdot\mathbb{P}(X = x) = \mathbb{E}[f(X)],\end{aligned}$$

where interchanging the limits to obtain the second equality is justified since f is bounded and since for any n we have that the $\mathbb{P}(X_n = x)$, $x \in S$, are non-negative and sum to 1 over $x \in S$ (later on the fact that this interchange is allowed will be generalized to integrals and it will be called 'dominated convergence'). Thus, the above equality implies that X_n converges in distribution to X, and hence μ_n converges weakly to μ.
Assume now, on the other hand, that the probability measures μ_n on $(S, 2^S)$ converge weakly to a probability measure μ. Denote by X_n and X random variables taking values in S that are distributed accordingly to μ_n and μ. Then, since for any $x \in S$, the function $\mathbb{1}_{\{x\}}$ is bounded and continuous, we get

$$\mu_n(x) = \mathbb{E}[\mathbb{1}_{\{x\}}(X_n)] \to \mathbb{E}[\mathbb{1}_{\{x\}}(X)] = \mu(x),$$

as $n \to \infty$. Thus, in particular we have for all $x \in S$ that $\mu_n(x) \to \mu(x)$ as $n \to \infty$.
This establishes the above claim.

The following questions immediately arise from our previous observations: Is the stationary distribution unique? Since one stationary distribution is given by the limit of the Cesàro means of the $\mu_0 P^n$ and since we have the fixed point interpretation of the stationary distribution from (1.16.10), might $\mu_0 P^n$ even converge to it?
Without further assumptions, these hopes turn out to be false.

Exercise 1.16.13. *Find a Markov chain on a finite state space such that*

(a) there are several stationary distributions for this Markov chain;

(b) the sequence $(\mu_0 P^n)$ of distributions does not converge weakly (or, if we interpret $\mu_0 P^n$ as a vector, pointwise, which is the same according to Remark 1.16.12).

The considerations of Example 1.16.13 lead us to the following definition:

Definition 1.16.14. *A transition matrix P is called* irreducible *if for any $x, y \in S$ there exists $m \in \mathbb{N}_0$ such that $P^m(x, y) > 0$.*

In words, this means that for any $x, y \in S$ there exists a finite time m such that if the chain is in x at time 0, then there is a positive probability that the chain moves from x to y in m time steps. It turns out that this is already sufficient to get the uniqueness of the stationary distribution as is shown in the following result.

Proposition 1.16.15. *Let P be an irreducible transition matrix of a Markov chain on a finite state space. Then there exists a unique stationary distribution π for P.*

In order to prove Proposition 1.16.15 we need to introduce some notation and the auxiliary result Lemma 1.16.17.

Definition 1.16.16. *For an at most countable set S and P a transition matrix, a function $h : S \to \mathbb{R}$ is called* harmonic (with respect to $P - I$) *if $(P - I)h = 0$, i.e., $Ph = h$, so h is a right eigenvector to P with eigenvalue 1.*
If the matrix P is clear from the context, we also just call h harmonic*, and we call it* harmonic at x *if $(Ph)(x) = h(x)$.*

Heuristically, a harmonic function can be interpreted in the following way: If we start the Markov chain at an arbitrary site $x \in S$, then the expected value of $h \circ X_n$ is constant in $n \in \mathbb{N}_0$; i.e., we have $h(x) = \mathbb{E}_x[h(X_n)]$ (alternatively, $h(x) = \sum_{y \in S} P(x, y)h(y)$, and inductively $h(x) = \sum_{y \in S} P(x, y)h(y)$) for all $n \in \mathbb{N}$. It is immediate that constant functions are harmonic, and the following lemma even tells us that these are the only functions which are harmonic, as long as P is irreducible and the state space is finite.

Lemma 1.16.17. *If $h : S \to \mathbb{R}$ is harmonic with respect to the irreducible transition matrix $P - I$ on the finite state space S, then h is constant.*

Proof. Let $M := \max_{x \in S} h(x)$ (in particular, note that the maximum is well-defined since we assume S to be finite) and let $x_0 \in S$ be such that $M = h(x_0)$. Then, since h is harmonic, we get

$$h(x_0) = \sum_{x \in S} P(x_0, x)h(x).$$

Since the $P(x_0, x)$ are non-negative and sum to 1 over $x \in S$, and since $h(x_0) = M$, we deduce that $h(x) = h(x_0)$ for all x such that $P(x_0, x) > 0$. Using induction, we obtain that $h(x) = h(x_0)$ for all $x \in S$ for which there exists $n \in \mathbb{N}$ such that $P^n(x_0, x) > 0$. Since we have assumed that P is irreducible, the latter is the case for all $x \in S$, which implies $h(x_0) = h(x)$ for all $x \in S$, and in particular h is constant. □

Exercise 1.16.18. *Give an example which shows that the conclusion of Lemma 1.16.17 fails to hold in general if one does not assume that S is finite or that P is irreducible.*

Proof of Proposition 1.16.15. Lemma 1.16.17 implies that the kernel of the linear map $P - I$ has dimension 1. Since the column rank and row rank of a matrix coincide, the kernel of $(P - I)^T = P^T - I$ also has dimension 1, and this implies that there is at most one vector π whose entries are non-negative, sum to 1, and that satisfies $\pi P = \pi$. Thus, in combination with Proposition 1.16.11 we conclude that there exists exactly one stationary distribution. □

Wrapping things up we observe that so far we have conditions for the existence and uniqueness of a stationary distribution. A central issue in applications to real world problems is to get information about convergence to it. In order to understand this topic a little bit better, again we introduce some more relevant notation.

1.16.2 Classification of states

The context of this subsection is interesting in its own right and will also be useful in applying the Perron-Frobenius theorem (Theorem 1.16.30 below) to transition matrices in order to prove our main convergence result for Markov chains (Theorem 1.16.39 below)

Definition 1.16.19. *For a Markov chain with state space S, transition matrix P, and for states $x, y \in S$, we say that y* can be reached from *x if there exists $n \in \mathbb{N}_0$ such that $P^n(x, y) > 0$, where P^0 is defined as the identity matrix in $\mathbb{R}^{S \times S}$. In this case we write $x \rightsquigarrow y$. We write $x \leftrightsquigarrow y$ if $x \rightsquigarrow y$ and $y \rightsquigarrow x$.*

Lemma 1.16.20. *The relation $\leftrightsquigarrow$ defines an equivalence relation on the state space S.*

Proof. Exercise. □

Exercise 1.16.21. *Convince yourself that it is necessary to take $n \in \mathbb{N}_0$ instead of $n \in \mathbb{N}$ here in order to get an equivalence relation on S.*

Corollary 1.16.22. *A transition matrix $P \in \mathbb{R}^{S \times S}$ is irreducible if and only if S has only one equivalence class with respect to $\leftrightsquigarrow$.*

Proof. If P is irreducible, then by definition, for any $x, y \in S$ there exist $m, n \in \mathbb{N}_0$ such that $P^m(x, y), P^n(y, x) > 0$. In particular, $x \leftrightsquigarrow y$. Since x and y were chosen arbitrarily, this means that there is exactly one equivalence class with respect to the relation $\leftrightsquigarrow$.
To prove the reverse direction, assume we have only one equivalence class with respect to $\leftrightsquigarrow$. As a consequence, for any $x, y \in S$ there is $n \in \mathbb{N}_0$ with $P^n(x, y) > 0$, and this implies that P is irreducible. □

Definition 1.16.23. *A state x is called* essential ('wesentlich') *if for all y, $x \rightsquigarrow y$ implies $y \rightsquigarrow x$. Otherwise x is called* inessential ('unwesentlich').

Claim 1.16.24. *Let $x \in S$ be arbitrary. Then:*

(a) If x is essential then y is essential for all $y \in S$ with $x \leftrightsquigarrow y$.

(b) If x is inessential then y is inessential for all $y \in S$ with $x \leftrightsquigarrow y$.

Proof. Exercise □

Definition 1.16.25. *For a transition matrix P and a state $x \in S$, the* period of *x is defined as*

$$p_x := \gcd\{n \in \mathbb{N} : P^n(x, x) > 0\}.$$

A state x is called aperiodic *if $p_x = 1$.*
The transition matrix P (and the induced Markov chain) is called aperiodic *if for all $x \in S$ one has $p_x = 1$.*

Lemma 1.16.26. *The period is a class property of the equivalence relation induced by $\leftrightsquigarrow$. I.e., $x \leftrightsquigarrow y$ implies $p_x = p_y$.*

Proof. Exercise. □

Remark 1.16.27. *Often aperiodic chains are easier to deal with than periodic ones. One way to make a Markov chain P aperiodic is by considering its lazy version which has transition matrix $\frac{P+I}{2}$ (check that this defines a transition matrix also). Intuitively, the difference of the corresponding Markov chain to the original one is that each time this chain tosses an extra coin to decide whether it moves in the first place or not. Many characteristics of the original chain are preserved by the lazy chain (like e.g. the equivalence classes induced by the relation $\leftrightsquigarrow$, or the law of its infinite trace and hence properties such as recurrence or transience), but it is often easier to investigate due to its aperiodicity.*

1.16.3 The Perron-Frobenius theorem

Definition 1.16.28. *The* spectral radius $\varrho(A)$ *of a square matrix $A \in \mathbb{R}^{n\times n}$ is defined as the maximum of the moduli of its eigenvalues:*

$$\varrho(A) := \max\big\{|\lambda| \,:\, \lambda \text{ is an eigenvalue of } A\big\}.$$

Remark 1.16.29. *It might be useful to note here that if A is a real square matrix, then $\varrho(A) = \varrho(A^T)$ due to Remark 1.16.9.*

Theorem 1.16.30 (Perron-Frobenius theorem; Oskar Perron ((1880–1975) and Ferdinand Frobenius (1849 –1917)))**.** *Let M be a non-negative $n \times n$ matrix such that $M^r > 0$ some $r \in \mathbb{N}$ (i.e., $M^r(i,j) > 0$ for all $i, j \in \{1, \dots, n\}$).*
Then

(i) the spectral radius $\varrho(M) \in (0, \infty)$ is a simple eigenvalue of M (recall that 'simple' means its algebraic multiplicity is 1);

(ii) there is a strictly positive left eigenvector and a strictly positive right eigenvector of M with eigenvalue $\varrho(M)$;

(iii) for any eigenvalue $\lambda \neq \varrho(M)$ one has $|\lambda| < \varrho(M)$.

Proof. Since the proof is long and not probabilistic at its core, we omit it here and refer the interested reader to [Sen06, Theorem 1.1] for example. □

Remark 1.16.31. *The proof of Theorem 1.16.30 is relatively long and technical. Therefore, it is worth to note here that Proposition 1.16.15, which has been obtained using relatively neat and simple arguments, is not too far off from the statement of Theorem 1.16.30 applied to a stochastic matrix P. In fact, Proposition 1.16.15 provides us with the fact that P has a left eigenvector with eigenvalue 1, which has non-negative entries only; in fact, due to the irreducibility it has positive entries only.*[14] *Furthermore, the proof of Proposition 1.16.15 immediately yields that the geometric multiplicity of the eigenvalue 1 is 1.*
In addition, it is easy to obtain that $\varrho(P) = 1$.[15]
Comparing the above to Theorem 1.16.30 applied to a stochastic matrix P, we see that the only relevant properties which we have not derived so far via our previous results are the following:

- *1 is a simple eigenvalue;*
- *$|\lambda| < \varrho(P) = 1$ for all eigenvalues λ different from the eigenvalue 1.*

[14] Indeed, for irreducible P we get that π has positive entries only: Choose $x \in S$ with $\pi(x) > 0$. For any $y \in S$, since P is irreducible, we obtain that there exists $n_{x,y} \in \mathbb{N}$ such that $P^{n_{x,y}}(x,y) > 0$. Thus, since $\pi = \pi P = \pi P^{n_{x,y}}$ we deduce that $\pi(y) = (\pi P^{n_{x,y}})(y) = \sum_{z\in S} \pi(z) P^{n_{x,y}}(z,y) > \pi(x) P^{n_{x,y}}(x,y) > 0$.

[15] We know that 1 is a left eigenvalue, so it remains to show that $\varrho(P) \leqslant 1$. Assume to the contrary that $\varrho(P) > 1$. Then there exists a left eigenvalue λ with

$$|\lambda| > 1 \tag{1.16.25}$$

to a left eigenvector $w \neq 0$. As a consequence, $wP^n = \lambda^n w$. Since P^n is a stochastic matrix again (see Exercise 1.16.5), we get that

$$\|wP^n\|_\infty \leqslant \|w\|_\infty \cdot |S|. \tag{1.16.26}$$

On the other hand, we have

$$\|\lambda^n w\|_\infty = |\lambda|^n \|w\|_\infty. \tag{1.16.27}$$

Now for $n \to \infty$, the left-hand side of (1.16.26) stays bounded, whereas the expression in (1.16.27) tends to infinity due to (1.16.25). In particular, they cannot be equal for all n, and therefore we must have $|\lambda| \leqslant 1$, which implies $\varrho(P) \leqslant 1$ since the eigenvalue λ was chosen arbitrarily.

1.16.4 Quantitative convergence of Markov chains to equilibrium

We start with a basic equality for Markov chains which is a generalization of (1.16.3).

Proposition 1.16.32 (Chapman-Kolmogorov equation (Sydney Chapman (1888–1970), Andrey Kolmogorov (1903–1987)))**.** *Let a Markov chain (X_n) with state space S and transition matrix P be given. Then for any $m, n \in \mathbb{N}_0$ and all $x, y \in S$ we have*

$$P^{m+n}(x,y) = \sum_{z \in S} P^m(x,z) P^n(z,y).$$

Proof. This is a direct consequence of the fact that $P^{m+n} = P^m P^n$, which even if S is infinite countable follows in the same way from the associativity of matrix multiplication of stochastic matrices as it does in the case of S finite (see also (1.16.6)). □

The following lemma is an auxiliary result for showing that if P is the transition matrix of an irreducible aperiodic Markov chain on a finite state space, then there exists $n \in \mathbb{N}$ such that $P^n > 0$. Essentially, it will give us a condition which ensures that the assumptions of the Perron-Frobenius theorem are fulfilled.

Lemma 1.16.33. *Let $\Lambda \subset \mathbb{N}$ be nonempty and set $d := \gcd(\Lambda)$, the greatest common divisor of all elements of Λ.*
If d is finite, then there exist $m \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_m \in \Lambda$, and $N_0 \in \mathbb{N}$, such that for all $n \geqslant N_0$ there exist coefficients $\alpha_{1,n}, \ldots, \alpha_{m,n} \in \mathbb{N}$ with

$$nd = \sum_{j=1}^{m} \alpha_{j,n} \lambda_j.$$

Proof. Denote by

$$G := \left\{ \sum_{j=1}^{m} r_j \lambda_j \,:\, m \in \mathbb{N},\, r_1, \ldots, r_m \in \mathbb{Z},\, \lambda_1, \ldots, \lambda_m \in \Lambda \right\}$$

the smallest additive subgroup of $\mathbb{Z}$ that contains Λ, and denote by d' the smallest positive element of G. We start with showing that $d' = d$.
Since d is a divisor of λ (write $d|\lambda$) for all $\lambda \in \Lambda$, it follows that $d|g$ for all $g \in G$, and hence $d|d'$, i.e., in particular

$$d \leqslant d'. \tag{1.16.28}$$

On the other hand, each $g \in G$ can be written as $g = rd' + s$ for some s with

$$0 \leqslant s < d'. \tag{1.16.29}$$

Now $s = g - rd' \in G$, and since d' was the minimal positive element in G we obtain from (1.16.29) that $s = 0$. Thus, d' divides each $g \in G$ and hence in particular each $\lambda \in \Lambda$, which implies that $d'|d$ and therefore $d' \leqslant d$. Thus, in combination with (1.16.28) we deduce $d' = d$.
As a consequence, we find $\lambda_1, \ldots, \lambda_m \in \Lambda$ und $r_1, \ldots, r_m \in \mathbb{Z}\backslash\{0\}$ such that

$$d = \sum_{j=1}^{m} r_j \lambda_j. \tag{1.16.30}$$

We still have to deal with the issue that some of the $r_1, \ldots, r_m$ can be negative. For this purpose observe now that we can find $b_1, \ldots, b_m \in \mathbb{N}$ such that

$$\lambda_j = b_j d \quad \text{for all } j = 1, \ldots, m. \tag{1.16.31}$$

Write

$$\begin{aligned} b &:= \min\{b_j \,:\, 1 \leqslant j \leqslant m\} \quad \text{and} \\ N_0 &:= \sum_{j=1}^{m} b \cdot |r_j| \cdot b_j. \end{aligned}$$

Then each $n \geqslant N_0$ can be written as

$$n = s_1 b_1 + \cdots + s_m b_m + s \tag{1.16.32}$$

with

$$0 \leqslant s < b \quad \text{and} \quad s_j \geqslant b \cdot |r_j| \quad \forall j \in \{1, \ldots, m\}. \tag{1.16.33}$$

Hence,

$$\begin{aligned} nd &\overset{(1.16.32)}{=} \sum_{j=1}^{m} s_j b_j d + sd \\ &\overset{(1.16.30),(1.16.31)}{=} \sum_{j=1}^{m} s_j \lambda_j + s \sum_{j=1}^{m} r_j \lambda_j \\ &= \sum_{j=1}^{m} (s_j + s r_j) \lambda_j \end{aligned}$$

This finishes the proof, since it follows from (1.16.33) that we have $s_j + s r_j \geqslant 1$. □

If the transition matrix P is irreducible we have that for any $x, y \in S$ there exists $n \in \mathbb{N}$ with $P^n(x, y) > 0$. The following result implies that under the additional assumption of aperiodicity, the number n can be chosen independently of the choice of $x, y \in S$. In particular, this will enable us to apply the Perron-Frobenius theorem to Markov chains with such transition matrices.

Corollary 1.16.34. *Let P be the transition matrix of an irreducible and aperiodic Markov chain with finite state space. Then for all $n \in \mathbb{N}$ large enough, $P^n > 0$.*

Proof. Let S denote the state space. Since S is finite it is sufficient to show that for each $x, y \in S$ there exists $N_{x,y} \in \mathbb{N}$ such that

$$P^n(x, y) > 0 \quad \forall\, n \geqslant N_{x,y}. \tag{1.16.34}$$

Thus, let $x, y \in S$ be arbitrary and consider $\Lambda := \{n \in \mathbb{N} : P^n(y, y) > 0\}$. As P is aperiodic, Lemma 1.16.26 implies that $\gcd(\Lambda) = 1$, and the previous lemma implies that there exist $\lambda_1, \ldots, \lambda_m \in \Lambda$ and $\widetilde{N} \in \mathbb{N}$ such that for any $n \geqslant \widetilde{N}$ we find $\alpha_1, \ldots, \alpha_m \in \mathbb{N}$ with

$$n = \sum_{j=1}^{m} \alpha_j \lambda_j. \tag{1.16.35}$$

Since P is irreducible, we find $n_{x,y}$ such that

$$P^{n_{x,y}}(x, y) > 0$$

We are going to show that (1.16.34) holds true for $N_{x,y} := \widetilde{N} + n_{x,y}$. For this purpose, observe that due to (1.16.35), any $n \geqslant N_{x,y}$ can be written as

$$n = \sum_{j=1}^{m} \widetilde{\alpha}_j \lambda_j + n_{x,y},$$

some $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_m \in \mathbb{N}$. As a consequence of the Chapman-Kolmogorov equation Proposition 1.16.32, we obtain

$$P^n(x, y) \geqslant P^{n_{x,y}}(x, y) \prod_{j=1}^{m} \left(P^{\lambda_j}(y, y)\right)^{\widetilde{\alpha}_j},$$

and the right-hand side is positive since by assumption we have $P^{n_{x,y}}(x, y) > 0$ and $P^{\lambda_j}(y, y) > 0$. □

Example 1.16.35. *Find an example of an irreducible aperiodic Markov chain on an infinite state space which shows that the assumption of the state space being finite is essential for the conclusion of the above Corollary 1.16.34 to hold.*

The following corollary specialises the results that we have obtained so far to the setting of Markov chains, in particular with a view towards proving Theorem 1.16.39.

Corollary 1.16.36. *Let P be the transition matrix of an irreducible and aperiodic Markov chain with finite state space. Then 1 is a simple eigenvalue of P and for any other eigenvalue λ of P one has $|\lambda| < 1$. The unique (according to Proposition 1.16.15 or otherwise according to Theorem 1.16.30) strictly positive left eigenvector of P associated to 1, whose entries sum to one, will be denoted by π, and it is the stationary distribution of the Markov chain associated to P.*

Proof. As in Remark 1.16.31 we obtain that $\varrho(P) = 1$. Therefore, and since due to Corollary 1.16.34 we may apply Theorem 1.16.30, all remaining implications of the corollary follow directly from Theorem 1.16.30.

□

As the title of the subsection suggests, we want to study the quantitative convergence of the distribution of a Markov chain to its equilibrium. For this purpose we need to be able to quantify the distance between two (probability) measures.

Definition 1.16.37. *Let μ, ν be probability measures on an at most countable measurable space $(S, 2^S)$. The total variation distance of μ and ν is given by*

$$\|\mu - \nu\|_{\mathrm{TV}} := \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_{x\in S}|\mu(x) - \nu(x)| = \sup_{A\in 2^S}|\mu(A) - \nu(A)|, \tag{1.16.36}$$

where $\|\cdot\|_1$ is the 1-norm on the space $\mathbb{R}^S$.

Exercise 1.16.38. *Prove the rightmost equality of* (1.16.36).

For the transition matrix P of an aperiodic and irreducible Markov chain with finite state space S, denote by $\lambda_1, \lambda_2, \ldots, \lambda_N$, $N \leqslant |S|$, the *distinct* eigenvalues of P in such a way that

$$1 = \lambda_1 > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_N|. \tag{1.16.37}$$

We denote by $\sigma := \lambda_1 - |\lambda_2|$ the so-called *spectral gap*. The following theorem tells us that the spectral gap plays a prominent role in determining the speed of convergence to equilibrium.

Theorem 1.16.39 (Convergence theorem for Markov chains)**.** *Consider an irreducible and aperiodic Markov chain on a finite state space S with transition matrix P and (unique, see Corollary 1.16.36) stationary distribution π. Then there exists a constant $C \in (0, \infty)$ such that*

$$\sup_{x\in S}\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leqslant C(1 - \sigma + \varepsilon(n))^n \tag{1.16.38}$$

for all $n \in \mathbb{N}$, and where the error term ε is in $\mathcal{O}(\ln n/n)$; i.e., there exists a constant $\widetilde{C} \in (0, \infty)$ such that $|\varepsilon(n)| \leqslant C \ln n/n$ for all $n \in \mathbb{N}$.

Remark 1.16.40. *If one just aims for a convergence of $\sup_{x\in S}\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}}$ to 0 which is exponential in n (without a good bound on the exponential rate $\ln(1-\sigma+\varepsilon(n))$ as we have it on the RHS of* (1.16.38) *with $1 - \sigma$), then there are slightly easier proofs available, see e.g. [LPW09, Theorem 4.9].*
A close look at our proof will reveal that the bound we have on the RHS of (1.16.38) *is essentially optimal (not in the constant C though)*

Proof of Theorem 1.16.39. The driving idea of the proof is easier to understand once the transition matrix is transformed to a similar Jordan matrix, which has the corresponding eigenvalues on the diagonal (and some 1's next to it). Theorem 1.16.30 tells us that there is a simple eigenvalue 1, and all other eigenvalues have modulus smaller than one. Therefore, taking higher and higher powers of the transition matrix, we can deduce that the contribution of those eigenvectors that correspond to eigenvalues with modulus smaller than one vanishes exponentially. Furthermore, since at each time the distribution of the Markov chain is a probability vector, that means that 'more and more mass' will be put on the stationary distribution π.
Formally: We recall that using a basis transformation, P can be written as

$$P = TJT^{-1}, \tag{1.16.39}$$

where the matrix $T = (u_1, u_2, \dots, u_{|S|})$ consisting of the column vectors $u_1, \dots, u_{|S|}$, is invertible in $\mathbb{R}^{S\times S}$, its inverse is T^{-1}, and J is a matrix in Jordan normal form, i.e., we have

$$J = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 1 & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \ddots & \ddots & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \ddots & 1 & \lambda_2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \ddots & 0 & \lambda_3 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \ddots & 1 & \lambda_3 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \lambda_3 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & 1 & \lambda_3 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \lambda_N & 0 & \cdots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & \lambda_N & \ddots & \cdots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \lambda_N & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 1 & \lambda_N
\end{pmatrix},$$

where without loss of generality we still assume the eigenvalues to be ordered in such a way that (1.16.37) holds.

The nice feature that we will be taking advantage of is that powers of Jordan blocks are easily computed and, for blocks corresponding to eigenvalues other than λ_1, they are nicely upper bounded in their 'total contributions' to the matrix power. In fact, as you may either know from linear algebra or otherwise prove using induction, in the n-th power of a Jordan block (say the one corresponding to eigenvalue λ_2) we will see λ_2^n's on the diagonal, $\binom{n}{1}\lambda_2^{n-1}$'s on the lower diagonal, $\binom{n}{2}\lambda_2^{n-2}$'s on the diagonal below the lower diagonal, and so on. Now note that for any $2 \leqslant j \leqslant N$, $k \in \mathbb{N}$ fixed, we have that there exists a constant $C_k \in (0, \infty)$ such that

$$\left|\binom{n}{k}\lambda_j^{n-k}\right| \leqslant C_k n^k |\lambda_j|^n \leqslant C_k(1 - \sigma + \varepsilon(n))^n \tag{1.16.40}$$

for all $n \in \mathbb{N}$. Now for an arbitrary probability row vector $w \in \mathbb{R}^S$ (a.k.a. distribution on S), we have

$$w \cdot T = \sum_{j=1}^{|S|} (w^T \cdot u_j)\, \mathrm{e}_j, \tag{1.16.41}$$

with (e_j) the canonical basis of $\mathbb{R}^S$. Taking $n \to \infty$ the components in directions e_j, $2 \leqslant j \leqslant |S|$, of $w \cdot T \cdot J^n$ converge to 0 uniformly in j due to $\sigma < 1$ and (1.16.40); but still we have that

$$w \cdot P^n = w \cdot T \cdot J^n \cdot T^{-1} \tag{1.16.42}$$

as well as its limit is a probability vector. This can be used to infer

$$(w^T \cdot u_1)\mathrm{e}_1 \cdot T^{-1} \text{ is a probability vector, and it must equal } \pi; \tag{1.16.43}$$

Indeed, the latter is true since due to (1.16.42) we have that wP^n converges to

$$w \cdot T \cdot \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \cdots & \ddots \end{pmatrix} T^{-1} = (w^T \cdot u_1)\mathrm{e}_1 \cdot T^{-1},$$

and since

$$\lim_{n\to\infty} wP^n = (\lim_{n\to\infty} wP^{n-1}) \cdot P,$$

we get from the uniqueness of the stationary distribution that $\pi = \lim_{n\to\infty} w \cdot P^n$. Hence,

$$w \cdot P^n = \pi + \sum_{j=2}^{|S|} (w^T \cdot u_j) \mathrm{e}_j \cdot J^n \cdot T^{-1}.$$

Thus, using (1.16.39) to (1.16.43) we infer that there exists a constant $C \in (0, \infty)$ depending on σ and N but not on the probability vector w, such that

$$\|\pi - wP^n\|_{\mathrm{TV}} \leqslant C(1 - \sigma + \varepsilon(n))^n, \tag{1.16.44}$$

Since w had been chosen to be an arbitrary probability vector on S, this finishes the proof. □

Remark 1.16.41. *One should note here that even in the case of the proof of the speed of convergence in the above result, this is only asymptotic as time goes to infinity. Especially for practical purposes it would be helpful to have non-asymptotic results available also (see so-called 'cut-off phenomena', which, however, are generally much harder to prove).*

An example of groundbreaking importance which shows that already on a finite state space Markov chains can yield very important results is the PageRank algorithms for ordering webpages according to their relevance.

Example 1.16.42 (PageRank algorithm)**.** *The original PageRange algorithm has been introduced by Google co-founders Larry Page and Sergei Brin in [BP98] (which makes an interesting read anyway, even though there is no rigorous mathematics in there) as an algorithm to rank the importance of websites in the internet.*
A simple description of it is as follows: Consider the directed graph whose vertex set is given by the set of webpages S and whose edge set E is defined as follows. There is a directed edge $(x, y) \in E$ from $x \in S$ to $y \in S$ if and only if the webpage x has a link to the webpage y. For $\alpha \in (0, 1)$ we now consider the Markov chain on this graph whose dynamics is given as follows. If the chain is at state x at time n, then with probability α the chain chooses one of the sites $y \in S$ for which we have $(x, y) \in E$ uniformly and jumps there at time $n + 1$. With probability $1 - \alpha$, the chain choose a vertex $y \in S$ uniformly and jumps there at time $n + 1$.
This dynamics models the behavior of a person surfing the web who, if she is browsing a certain webpage, with probability α chooses uniformly one of the links of this webpage and jumps to the webpage that this link points to. With probability $1-\alpha$ she gets bored and restarts the search at a webpage chosen uniformly at random in the entire universe of webpages.
Formally, writing

$$d_x := \big|\{y \in S : (x, y) \in E\}\big|$$

for the so-called out-degree *of $x \in S$, the transition matrix is characterized via*

$$P(x, y) = \begin{cases} \frac{\alpha}{d_x} + \frac{1-\alpha}{|S|} & \text{if } (x, y) \in E, \\ \frac{1-\alpha}{|S|}, & \text{if } (x, y) \notin E. \end{cases}$$

Since $\alpha \in (0, 1)$, we see that the Markov chain thus described is irreducible and aperiodic, which makes our results applicable. In particular, Proposition 1.16.15 now tells us that this Markov chain has a unique stationary distribution π. The PageRank of a certain webpage $x \in S$ is now defined as $\pi(x)$, i.e., as the weight of the webpage under the stationary distribution of the above Markov chain.
The heuristics behind is that the more links a webpage receives from other webpages, and the more important these webpages are, the more important the original webpage should be.
As you might have noticed, the above does not include the actual search term in the computation of the PageRank of a webpage. As a consequence, the PageRank of a vector does not yet give a useful result of a web search, and indeed it is only one out of numerous criteria that contribute to the final result of a web search.
Computing the PageRank of all webpages is central to ranking the webpages according to the above algorithm, and it amounts to computing the stationary distribution of the above Markov chain. In theory this issue is easy to deal with, we just have to find the solution $\pi = \pi \cdot P$, which according to Proposition 1.16.15 is unique if we impose that π be a probability vector. In practice, however, this means we have to solve a huge linear system of equations (since there are a lot of webpages).

Chapter 2

Statistics

In the first Chapter, which dealt with probability theory, we have mainly been concerned with situations in which we assumed that certain experiments (the outcomes of which we understood to be random) were realized according to given mechanisms that we usually were aware of (e.g., we assumed the coin of coin tossing experiments to be fair, or otherwise that one of the two outcomes was favoured by choosing the probabiliy p of its occurrence in $(\frac{1}{2}, 1)$; however, we did not investigate as to whether choosing $p = \frac{1}{2}$ or $p \in (\frac{1}{2}, 1)$ was actually a reasonable thing to do). In statistics, which is the subject of this second Chapter, we take a different point of view which is sort of reverse to what we did before. In fact, in statistics we are usually given observations of experiments and then want to understand in more or less detail the mechanisms according to which these observations have been realized (by using the techniques developed in probability theory). Thus, probability theory and statistics can be seen as complementing each other.
We will focus on the following three topics, the precise nature of which will become clear as we investigate them:

(a) parameter estimation;

(b) confidence intervals / regions;

(c) tests;

A standard problem in statistics is the following.

Example 2.0.1. *What is the number of persons in an entire population of size N that has a certain illness? Or, equivalently, what is the probability*[1] *p that a person in that population is suffering from a certain illness? One can take a sample of n persons out of the entire population and find out the number m of sick persons among them; if n is chosen reasonably large (whatever that means for the time being), one might be tempted to hope that for example due to the law of large numbers, m/n would be a good guess for p (and we will investigate later on how good this guess is): If the n persons we chose for our sample could be interpreted as independent and identically distributed samples out of the entire population, then the law of large numbers would tell us that the ratio m/n would converge to p as $n \to \infty$ (of course, m depends on n here as well, even though this is not emphasised in the notation).*

In order to be able to treat examples such as the above systematically, we introduce some suitable definitions.

Definition 2.0.2. *A* statistical model ('statistisches Modell') *is a triplet $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, where $\mathfrak{X}$ is a set called the* sample space ('Stichprobenraum', 'Beobachtungsraum')*, $\mathcal{A}$ is a σ-algebra on $\mathfrak{X}$, and $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ is a family of probability measures on the measurable space $(\mathfrak{X}, \mathcal{A})$.*
Since we now have several probability measures $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ at our disposal, when taking expectations we have to specify with respect to which we want to do so – for this reason, we write $\mathbb{E}_\vartheta[\cdot]$ for the expectation with respect to $\mathbb{P}_\vartheta$.

This is to be interpreted in the sense that the outcomes of the experiments we observe will be considered elements of $\mathfrak{X}$, realized according to $\mathbb{P}_\vartheta$. Usually it is not too hard to come up with a suitable measurable

[1] It might help intuition to recall the frequentist approach to probability that had been introduced in (1.2.1); the probability p is then given by the number of sick people divided by the entire population size N.

space $(\mathfrak{X}, \mathcal{A})$ (in fact, most of the times we will be in the situation that $\mathfrak{X} = \mathbb{R}^d$ and $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$, or otherwise $\mathfrak{X} \in \mathcal{B}(\mathbb{R}^d)$ and $\mathcal{A} = \mathfrak{X} \cap \mathcal{B}(\mathbb{R}^d)$). However, it is generally more demanding to find the 'real' ϑ (or a function $\tau(\vartheta)$ of ϑ which is of interest, for that matter) and hence the $\mathbb{P}_\vartheta$ according to which the observed data has been realized.

Also recall that so far we have generally used the notation $(\Omega, \mathcal{F}, \mathbb{P})$ to denote probability spaces. The reason for denoting our measurable space to be endowed with the probability measures $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ by $(\mathfrak{X}, \mathcal{A})$ instead of $(\Omega, \mathcal{F})$ is the following: We assume that the data we observe is given by realizations of random variables $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathfrak{X}, \mathcal{A})$. As we have seen in Chapter 1 on probability theory, the very structure of $(\Omega, \mathcal{F}, \mathbb{P})$ is oftentimes irrelevant to us – what is of interest, however, is the law $\mathbb{P} \circ X^{-1}$ of the random variable X, which describes a probability measure on $(\mathfrak{X}, \mathcal{A})$ (cf. Theorem 1.7.6). Thus, it will be our task to find that $\mathbb{P}_\vartheta$ in the family $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$, for which we have $\mathbb{P} \circ X^{-1} = \mathbb{P}_\vartheta$.[2]

Example 2.0.3 (Product model)**.** *We will often be in the situation that we have observed n realizations of an experiment for which we assume that the random variables X_i, $1 \leqslant i \leqslant n$, with $X_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ describing the outcome of the i-th experiment, are independent and identically distributed. In that case it lends itself to consider the corresponding* product model *(see e.g. the continuation of Example 2.0.1 below), i.e., the statistical model*

$$(\mathfrak{X}, \mathfrak{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}) = (E^n, \mathcal{E}^{\otimes n}, (Q_\vartheta^{\otimes n})_{\vartheta\in\Theta});$$

here, if E is an at most countable state space we have that $\mathcal{E} = 2^E$ and $\mathcal{E}^{\otimes n} = 2^{(E^n)}$, and for a measure Q_ϑ on $(E, \mathcal{E})$ (through which we hope to describe the distribution of a single X_i) we denote by $Q_\vartheta^{\otimes n}$ the corresponding product measure on $(E, \mathcal{E})$ (recall (1.6.7)*). On the other hand, if $E \in \mathcal{B}(\mathbb{R}^d)$ we have $\mathcal{E}^{\otimes n} = (E \cap \mathcal{B}(\mathbb{R}^d))^{\otimes n}$, where the latter is a to be interpreted as the smallest σ-algebra on $(\mathbb{R}^d)^n$ such that for any $i \in \{1, \dots, n\}$, the projection*

$$\pi_i : (\mathbb{R}^d)^n \ni (\bar{x}_1, \dots \bar{x}_n) \mapsto \bar{x}_i \in \mathbb{R}^d$$

is an $\mathcal{E}^{\otimes n}$–$\mathcal{B}(\mathbb{R}^d)$-measurable function (recall Definition 1.7.1). In addition, $Q_\vartheta^{\otimes n}$ is a probability measure on $(E^n, \mathcal{E}^{\otimes n})$ which corresponds to the situation that the X_i are i.i.d. and each X_i has law Q_ϑ; to be more precise, $Q_\vartheta^{\otimes n}$ is the unique probability measure on $(E^n, \mathcal{E}^{\otimes n})$ such that

$$Q_\vartheta^{\otimes n}(A_1 \times \dots A_n) = \prod_{i=1}^n Q_\vartheta(A_i) \quad \text{for all } A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}^d). \tag{2.0.1}$$

The very existence of a measure with the properties as postulated in (2.0.1) *will only be established in the sequel class 'Probability Theory I'; for the time being we will take its existence for granted.*

Definition 2.0.4. *A statistical model $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ will be called*

(a) a parametrical model ('parametrisches Modell')*, if $\Theta \subset \mathbb{R}^d$;*

(b) a discrete model ('diskretes Modell')*, if $\mathfrak{X}$ is finite or countable and $\mathcal{A} = 2^{\mathfrak{X}}$; in this case we define*

$$\varrho_\vartheta(x) := \mathbb{P}_\vartheta(\{x\}). \tag{2.0.2}$$

It is called a continuous model *if $\mathfrak{X} \in \mathcal{B}(\mathbb{R}^d)$, $\mathcal{A} = \mathfrak{X} \cap \mathcal{B}(\mathbb{R}^d)$,[3] and if for each $\mathbb{P}_\vartheta$ there exists a density $\varrho_\vartheta : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\mathbb{P}_\vartheta(A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)\varrho_\vartheta(x)\,\mathrm{d}x, \tag{2.0.3}$$

for all $A \in \mathcal{A}$ for which we can evaluate the right-hand side.[4]

[2] Here, a natural problem arises: If, for example, we consider a finite sequence of tosses of the same coin, then it is impossible to say whether the coin is fair or not; indeed, for any $p \in (0, 1)$, a coin that shows heads with probability p and tails with probability $1 - p$ could have produced the observed realizations. Therefore, we will only be able to gauge how probable it is to observe a certain realization, knowing that p takes a certain value. In fact, if in our finite sequence of observed coin tosses there is a strong majority of heads, then this should *somehow* be an indicator that p is large for the coin in question. The precise meaning of this is important and oftentimes leads to confusion – we will make this more precise when we will be dealing with so-called *confidence intervals*.

[3] Recall the trace σ-algebra defined in Exercise 1.3.4.

[4] Since we have not introduced the Lebesgue integral yet, one might run into slight technical troubles when trying to evaluate the integral for 'not so nice' A. However, most of the times $\mathcal{A}$ is sufficiently nice (e.g., if we have $\mathfrak{X} = \prod_{i=1}^d [a_i, b_i]$ with $a_i < b_i$ for all $i \in \{1, \dots, d\}$).

We refer to the model as a standard statistical model *if one of the two cases of Definition 2.0.4 (b) occurs.*

Example (Example 2.0.1 cont'd)**.** *In this context it seems to suggest itself to choose* $\mathfrak{X} = \{0,1\}^n$*, where a* 1 *in the* k*-th place means that the* k*-th person is sick, whereas* 0 *means that this is not the case. This means we are dealing with a discrete model and consequently choose* $\mathcal{A} := 2^{\mathfrak{X}}$*. Furthermore, if we draw with replacement, then seeing a* 1 *in* k*-th place occurs with probability* $\vartheta \in [0,1]$*, and this event is independent of all events that do not depend on the* k*-th coordinate. As a consequence, we would choose*

$$\Theta = [0,1], \quad \text{and} \quad \mathbb{P}_\vartheta(x_1,\dots,x_n) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}, \quad \text{for any } (x_1,\dots,x_n) \in \mathfrak{X}.$$

We now want to formalize the procedure of finding the right parameter $\vartheta \in \Theta$.

2.1 Estimators

Definition 2.1.1. *Let* $(\Sigma, \mathcal{S})$ *be a measurable space.*

(a) A random variable $S : (\mathfrak{X}, \mathcal{A}) \to (\Sigma, \mathcal{S})$ *is called a* statistic ('Statistik').

(b) Let an arbitrary function $\tau : \Theta \to \Sigma$ *be given, which assigns to every parameter* $\vartheta \in \Theta$ *an element of the set* Σ*. Then a statistics* $T : (\mathfrak{X}, \mathcal{A}) \to (\Sigma, \mathcal{S})$ *is called an* estimator ('Schätzer', 'Schätzfunktion') for τ.

Remark 2.1.2. *(a) It seems unnecessary to introduce yet another terminology 'statistics' for a random variable in Definition 2.1.1. The reason for this is to highlight a difference in interpretation: Whereas we have used and will use the term 'random variable' in order to describe a random experiment, the term 'statistics' is supposed to highlight the fact that such a random variable does not naturally come as the description of a random experiment, but rather has to be constructed in such a way that it is useful to the statistician, see also (b) below.*

(b) It might seem surprising here that the definition of an estimator for τ *does not depend on the specifics of* τ *at all. It will indeed turn out that we will mostly be interested in estimators which are 'reasonable' for* τ*; in some sense, heuristically, we would consider an estimator reasonable if*

$$\text{for all } \vartheta \in \Theta, \text{ with high } \mathbb{P}_\vartheta\text{-probability, } T \text{ takes values close to } \tau(\vartheta). \tag{2.1.1}$$

However, in order not to curtail flexibility, this limitation is not imposed at the stage of this definition yet.

Example (Example 2.0.1 cont'd)**.** *We now want to find the (or, more realistically, a reasonable guess for the)* ϑ *that led to the realization we observed. For this purpose, we choose* $\tau(\vartheta) = \vartheta$*, and we want to construct an estimator for* τ *that maps into* $(\Sigma, \mathcal{S}) = (\Theta, \mathcal{B}([0,1]))$*.*
One possibility to define our estimator T *would be to set*

$$\begin{aligned} T : (\{0,1\}^n, 2^{\{0,1\}^n}) &\to [0,1] \\ (x_1,\dots,x_n) &\mapsto \frac{1}{2}. \end{aligned} \tag{2.1.2}$$

In the case $\vartheta = \frac{1}{2}$*, according to the heuristics of* (2.1.1)*, this would be a very good estimator since it always recovers the real parameter* $\vartheta = \frac{1}{2}$*. However, in* (2.1.1) *we demand that for* all $\vartheta \in \Theta$*, the estimator* T *be close to* $\tau(\vartheta)$*.*
Therefore, a presumably better option for choosing this estimator might be to set

$$\begin{aligned} T : (\{0,1\}^n, 2^{\{0,1\}^n}) &\to [0,1] \\ (x_1,\dots,x_n) &\mapsto \frac{\sum_{i=1}^n x_i}{n}, \end{aligned} \tag{2.1.3}$$

which takes into account the very specifics of τ*, and which again is motivated by the law of large numbers. On a very heuristic level the relation of the estimators defined in* (2.1.2) *and* (2.1.3) *can be compared to that of two clocks, one of which stopped running whereas of the second clock you know that the time it*

shows deviates at most say ten minutes from the real time. Generally, you would also prefer the second option over the first.
This example became easier to approach by our assumption that the event that the k-th person is sick was independent (and identically distributed in the ks) from the events not depending on the k-th coordinate since it gave rise to the fact that we could assume

$$\mathbb{P}_\vartheta = (\vartheta\delta_1 + (1-\vartheta)\delta_0)^n,$$

see (1.6.7) *also. If we did not make this assumption, one possibility would be to choose Θ to be the space of all probability measures on $(\{0,1\}^n, 2^{\{0,1\}^n})$ (in this case the model would not be a parametric one anymore) This, however, would make it way harder to obtain a good guess for ϑ or $\mathbb{P}_\vartheta$, respectively.*
We already observe here a general pattern: Choosing Θ too large in relation to the observed data available, it is hard to obtain a good guess for ϑ. On the other hand, choosing Θ too small one might run the risk that the actual probability distribution that governs the generation of the observed data is not contained in the family $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$.

2.1.1 Properties of estimators

We have seen that Definition 2.1.1 of an estimator for a certain given function τ is quite loose. As a consequence, it is up to us to find criteria which ensure that an estimator for τ is actually a good one.

Definition 2.1.3. *An estimator T for τ is called* unbiased ('erwartungstreu', 'unverfälscht') *if for each $\vartheta \in \Theta$, one has*

$$\mathbb{E}_\vartheta[T] = \tau(\vartheta).$$

Note that we need that $\Sigma \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$ in order to be able to compute $\mathbb{E}_\vartheta[T]$.[5]

Intuitively, an unbiased estimator does not tend to systematically under- or overestimate $\tau(\vartheta)$, and it seems to be a good property to ask for an estimator to fulfill.

Example 2.1.4. *(a) The estimator constructed in Example 2.0.1 in display* (2.1.3) *is unbiased. Indeed, using X_i to denote the state of health of the i-th person, i.e., $X_i(x_1,\dots,x_n) := x_i$ is the projection on the i-th coordinate,*[6] *we have*

$$\mathbb{E}_\vartheta[T] = \mathbb{E}_\vartheta\Big[\frac{\sum_{i=1}^n X_i}{n}\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta[X_i] = \vartheta.$$

(b) Consider the following setting: The sample space is given by $\mathfrak{X} := \mathbb{N}_0$, hence $\mathcal{A} := 2^{\mathbb{N}_0}$. Furthermore, set $\Theta := (0,1)$, and let $\mathbb{P}_\vartheta$ denote the Poisson distribution with parameter $-\frac{1}{2}\ln\vartheta$.

Claim 2.1.5.

$$T : \mathbb{N}_0 \ni n \mapsto (-1)^n \tag{2.1.4}$$

is the only unbiased estimator for $\tau(\vartheta) := \vartheta$.

Proof. We start with showing that the estimator is unbiased. Indeed, we get that

$$\mathbb{E}_\vartheta[T] = e^{\frac{1}{2}\ln\vartheta}\sum_{k=0}^\infty (-1)^k \frac{(-\frac{1}{2}\ln\vartheta)^k}{k!} = \sqrt{\vartheta}\sum_{k=0}^\infty \frac{(\frac{1}{2}\ln\vartheta)^k}{k!} = \vartheta,$$

which shows that the estimator is unbiased.

[5] In fact, we have seen how to compute $\mathbb{E}_\vartheta[T]$ for $k=1$. For $k \geqslant 2$, we define $\mathbb{E}_\vartheta[T]$ to be the vector containing the expectations of the projections of T onto its coordinates, i.e.,

$$\mathbb{E}_\vartheta[T] := (\mathbb{E}_\vartheta[\pi_1(T)],\dots,\mathbb{E}_\vartheta[\pi_k(T)]),$$

where for $i \in \{1,2,\dots,k\}$ we use the notation

$$\pi_i : \mathbb{R}^k \ni (x_1,\dots,x_k) \mapsto x_i \in \mathbb{R}$$

for the projection on the i-th coordinate.

[6] We will use this notation without further mention in the following.

On the other hand, it is the only estimator which is unbiased, for if $\widetilde{T}$ was another unbiased estimator we would obtain that for all $\vartheta \in (0,1)$

$$\vartheta = \mathbb{E}_\vartheta[\widetilde{T}] = e^{\frac{1}{2}\ln\vartheta} \sum_{k=0}^{\infty} \widetilde{T}(k) \frac{(-\frac{1}{2}\ln\vartheta)^k}{k!},$$

hence

$$\sqrt{\vartheta} = \sum_{k=0}^{\infty} \widetilde{T}(k) \frac{(-\frac{1}{2}\ln\vartheta)^k}{k!}.$$

On the other hand we have

$$\sqrt{\vartheta} = e^{\ln(\sqrt{\vartheta})} = \sum_{k=0}^{\infty} \frac{(\ln(\sqrt{\vartheta}))^k}{k!},$$

and using the uniqueness theorem for power series we deduce from the last two displays that $\widetilde{T}(k) = (-1)^k$ as desired. This shows that T as defined in (2.1.4) is the only unbiased estimator for $\tau(\vartheta) = \vartheta$. □

However, the estimator in (2.1.4) *can hardly be considered a good one since it does not even map into the set of parameters* Θ *(thus, note that in order to have that* T *does indeed define an estimator, we must have that* $\{-1,1\} \subset \Sigma$*, so choosing e.g.* $\Sigma := \mathbb{R}$ *would certainly do the job). Thus, we observe that on its own, the property that an estimator is unbiased does not yet guarantee that it is a reasonable estimator.*

Adding insult to injury, not only do we have 'bad' unbiased estimators as in the previous example, but there are also cases where no unbiased estimator exists, as the following example shows.

Example 2.1.6. *For* $n \in \mathbb{N}$*, let* $\mathfrak{X} := \{0,1,\dots,n\}$*,* $\mathcal{A} = 2^{\mathfrak{X}}$*, and for* $\Theta := (0,1)$ *and* $\vartheta \in (0,1)$*, let* $\mathbb{P}_\vartheta$ *denote* $\mathrm{Bin}_{n,\vartheta}$*. In addition, let the function* τ *be given by* $\tau(\vartheta) := \vartheta^{-1}$*, say.*

Claim 2.1.7. *There is no unbiased estimator for* τ*.*

Proof. Indeed, if T was an unbiased estimator for τ, for all $\vartheta \in \Theta$ we would have

$$\frac{1}{\vartheta} = \tau(\vartheta) = \mathbb{E}_\vartheta[T] = \sum_{k=0}^{n} T(k) \binom{n}{k} \vartheta^k (1-\vartheta)^{n-k},$$

which, multiplying by ϑ and subtracting 1 amounts to

$$0 = \sum_{k=0}^{n} T(k) \binom{n}{k} \vartheta^{k+1} (1-\vartheta)^{n-k} - 1.$$

Now we observe that on the right-hand side of the last display we have a non-trivial polynomial in ϑ of degree at most $n+1$. As a consequence it can have at most $n+1$ zeroes, which in particular implies that the last display can hold true for at most $n+1$ values of ϑ, but certainly not for all $\vartheta \in \Theta$. Therefore, T cannot be an unbiased estimator for τ.

□

The following serves as another standard example.

Example 2.1.8. *Assume that there is a random number generator that generates realizations of an i.i.d. sequence of random variables* $X_1,\dots,X_n$ *taking values in* $\{1,2,\dots,N\}$*. However, you are not told the value of* N*, but rather you are given a realization of* $X_1,\dots,X_n$ *generated by the random number generator and your task is to find a good estimator for* N*.*
Since we do not know N *a priori, we have to take care of the possibility that* X_i *may take any value in* $\mathbb{N}$*. Hence, for* $n \in \mathbb{N}$*, a corresponding statistical model can be defined as* $(\mathbb{N}^n, 2^{\mathbb{N}^n}, (\mathbb{P}_N^n)_{N\in\mathbb{N}})$*, where* $\Theta = \mathbb{N}$*, and* $\mathbb{P}_N^n$ *is the uniform distribution on the restriction of* $\mathbb{N}^n$ *to* $\{1,\dots,N\}^n$*, i.e.,* $\mathbb{P}_N^n(x_1,\dots,x_n) = \frac{1}{N^n}$*, for any* $(x_1,\dots,x_n) \in \{1,\dots,N\}^n$*. As before, we then have* $X_i(x_1,\dots,x_n) = x_i$ *for* $i \in \{1,\dots,n\}$*.*
There are different kinds of reasonings that suggest themselves in order to construct an estimator for N*.*

(a) We start with observing that for any choice of N, the probability under $\mathbb{P}_N^n$ that the maximum of $X_1, \dots, X_n$ is strictly smaller than N is given by

$$\Big(\frac{N-1}{N}\Big)^n.$$

Hence,

$$\lim_{n\to\infty} \mathbb{P}_N^n(\max\{X_1, \dots, X_n\} = N) = 1, \tag{2.1.5}$$

and it may seem reasonable to define the sequence of estimators

$$T_n(x_1, \dots, x_n) := \max\{x_1, \dots, x_n\}, \quad n \in \mathbb{N}. \tag{2.1.6}$$

(b) An alternative estimator can be constructed based on the following reasoning: If the X_i are realized according to $\mathbb{P}_N$, then the law of large numbers tells us that

$$\frac{\sum_{i=1}^n X_i}{n} \to \mathbb{E}_N^n[X_1] = \frac{1}{N}\sum_{i=1}^N i = \frac{N+1}{2} = \frac{N}{2} + \frac{1}{2}, \qquad \text{as } n \to \infty.$$

Hence, another sequence of estimators that looks reasonable is given by

$$\widetilde{T}_n := 2\frac{\sum_{i=1}^n X_i}{n} - 1 \quad \Big(\text{or, which is the same, } \widetilde{T}_n(x_1, \dots, x_n) := 2\frac{\sum_{i=1}^n x_i}{n} - 1\Big). \tag{2.1.7}$$

Comparing the above estimators we observe that the estimator in (a) is biased (for $N \geqslant 2$) since for any n and N, the estimator T_n never takes values above N, but it does take values smaller than N with strictly positive $\mathbb{P}_N$ probability. Nevertheless, from the above we can conclude that T_n is asymptotically unbiased in the sense that

$$\mathbb{E}_N^n[T_n] \to N \qquad \text{as } n \to \infty.$$

On the other hand, we observe that $\widetilde{T}_n$ is unbiased since

$$\mathbb{E}_N^n[\widetilde{T}_n] = \mathbb{E}_N^n\Big[2\frac{\sum_{i=1}^n X_i}{n} - 1\Big] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_N^n[X_i] - 1 = \frac{2}{n}\frac{n(N+1)}{2} - 1 = N.$$

Example 2.1.9 (Estimator for expectation and variance)**.** *Assume you know that the data you observe is the realization of an i.i.d. sequence $X_1, \dots, X_n$, and assume that you want to find unbiased estimators for the expectation and the variance of the underlying distribution.*
Due to the i.i.d. assumption we can consider the n-fold product model, which we denote as $(\mathfrak{X}^n, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ here.
It is not hard to check that

$$T_\mu = \frac{1}{n}\sum_{i=1}^n X_i \tag{2.1.8}$$

is an unbiased estimator for the unknown expectation $\tau(\vartheta) = \mathbb{E}_\vartheta[X_1] = \mathbb{E}_\vartheta[\frac{1}{n}\sum_{i=1}^n X_i]$.
To deal with the variance, if the expectation $\mathbb{E}_\vartheta[X_1]$ of the underlying random variables was known, a natural candidate for an estimator of the variance would then be given by

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}_\vartheta[X_1])^2, \tag{2.1.9}$$

which can be checked to be an unbiased estimator for the variance. Now since usually the expectation is not known either, we are tempted to replace it by its own unbiased estimator defined in (2.1.8)*. In this case we get due to $\mathbb{E}_\vartheta[X_i - T_\mu] = 0$ and Bienaymé's lemma that*

$$\begin{aligned}\mathbb{E}_\vartheta\Big[\frac{1}{n}\sum_{i=1}^n (X_i - T_\mu)^2\Big] &= \mathrm{Var}_\vartheta(X_1 - T_\mu) = \mathrm{Var}_\vartheta\Big(\frac{n-1}{n}X_1 - \frac{1}{n}\sum_{k=2}^n X_k\Big)\\ &= \Big(\Big(\frac{n-1}{n}\Big)^2 + \frac{n-1}{n^2}\Big)\mathrm{Var}_\vartheta(X_1) = \frac{n-1}{n}\mathrm{Var}_\vartheta(X_1).\end{aligned} \tag{2.1.10}$$

This means that this estimator given in (2.1.9) *is biased and we tend to underestimate the actual variance if we used it. Hence, on usually retreats to the estimator*

$$T_{\sigma^2} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - T_\mu)^2$$

for the variance, which is unbiased.

Although the property of being 'unbiased' is usually considered a good thing to have for an estimator, we should keep in mind the possible disadvantages we have discovered above:

- there might be no unbiased estimator (see Example 2.1.6);
- even if unbiased estimators do exist, they might not be useful (see Example 2.1.4 (b));
- the property of an estimator being unbiased is not transformation invariant in the sense that if T is an unbiased estimator for ϑ, then generally $\tau \circ T$ is not an unbiased estimator for $\tau(\vartheta)$.

2.1.2 Maximum likelihood estimators

As we have seen above, the notion of an estimator being unbiased does not necessarily lead to good estimators on its own. As a consequence, we introduce an additional idea for constructing estimators (which oftentimes leads to unbiased or at least consistent (see (2.1.20) below) sequences of estimators).

Definition 2.1.10. *Let* $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *be a statistical standard model. The function*

$$\begin{aligned}\varrho :\ & \mathfrak{X}\times\Theta \to [0,\infty),\\ & (x,\vartheta)\mapsto \varrho_\vartheta(x)\end{aligned}$$

is called the corresponding likelihood function ('Likelihoodfunktion') *(recall that* ϱ_ϑ *had been introduced in* (2.0.2) *and* (2.0.3)*).*
The mapping

$$\begin{aligned}\varrho_x :\ & \Theta \to [0,\infty),\\ & \vartheta \mapsto \varrho(x,\vartheta)\end{aligned}$$

is called likelihood function given the outcome $x \in \mathfrak{X}$ ('Likelihoodfunktion zum Beobachtungswert $x \in \mathfrak{X}$'). *In applications, an important role is played by the logarithms of the above functions which are referred to as the corresponding* log-likelihood functions*, see also Example 2.1.12 below.*

Definition 2.1.11. *An estimator* $T : \mathfrak{X} \to \Theta$ *for* $\tau(\vartheta) = \vartheta$ *is called a* maximum likelihood estimator ('Maximum-Likelihood Schätzer') *if for each* $x \in \mathfrak{X}$ *one has*

$$\varrho(x, T(x)) = \max_{\vartheta\in\Theta} \varrho(x,\vartheta).$$

As an abbreviation a maximum likelihood estimator is also denoted as MLE.

The idea behind maximum likelihood estimators is that they characterise the parameter ϑ such that under the corresponding $\mathbb{P}_\vartheta$ the observed data is the most likely; oftentimes, they provide us with estimators that are unbiased (or at least unbiased asymptotically). However, since we do not have any probability measure on the space Θ, there is nothing really compelling about them.

Example 2.1.12. *Consider the statistical model given by* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathbb{P}_\vartheta)_{\vartheta\in\mathbb{R}})$*, where* $\mathbb{P}_\vartheta$ *is the distribution on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *of a vector* $(X_1, \dots, X_n)$*, where the* X_i *are i.i.d.* $\mathcal{N}(\vartheta, \sigma^2)$*-distributed and* $\sigma^2 \in (0,\infty)$ *is supposed to be known for the sake of simplicity of exposition (otherwise, we have seen in Example 2.1.9 how to estimate the variance in an unbiased way).*
In order to find a maximum likelihood estimator for ϑ*, for a given observation* $(x_1, \dots, x_n) \in \mathbb{R}^n$ *we have to find that* $\vartheta \in \mathbb{R}$ *for which the density*

$$\varrho_\vartheta(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(x_i-\vartheta)^2}{2\sigma^2}\Big\}$$

is maximized for given x. Observing that a function attains a maximum at a certain argument if and only if the same applies to the logarithm of that function. Thus, we compute

$$\frac{\partial}{\partial\vartheta}(\ln \varrho(x,\vartheta)) = \sum_{i=1}^{n} \frac{\partial}{\partial\vartheta}\Big(- \frac{(x_i-\vartheta)^2}{2\sigma^2}\Big) = \sum_{i=1}^{n} \frac{x_i-\vartheta}{\sigma^2},$$

and we get that the maximizing ϑ is given by

$$\widehat{\vartheta}_{\mathrm{ML}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

In particular, we may check that this MLE is unbiased.

Example (Example 2.1.8 cont'd)**.** *For $(x_1,\dots,x_n) \in \{1,\dots,N\}$ and $\vartheta \in \mathbb{N}$ with $\vartheta \geqslant T_n(x) = \max\{x_1,\dots,x_n\}$ we have that*

$$\varrho(x,\vartheta) = \mathbb{P}^n_\vartheta((x_1,\dots,x_n)) = \Big(\frac{1}{\vartheta}\Big)^n.$$

Thus, we see that the argument $\vartheta \in \Theta$ maximising this expression is given by $\vartheta_{\max} := T_n(x) = \max\{x_1,\dots,x_n\}$ (since for $\vartheta < T_n(x)$ the probability of observing $(x_1,\dots,x_n)$ under $\mathbb{P}^n_N$ vanishes). Hence,

$$\varrho(x,T_n(x)) = \max_{\vartheta\in\Theta} \varrho(x,\vartheta),$$

and we see that T_n is an MLE for $\tau(\vartheta)=\vartheta$.

2.1.3 Fisher information and the Cramér-Rao inequality

This is yet another part of our endeavour to try to understand what it means for an estimator to be a good one. Although we have seen above that unbiased estimators do not necessarily provide us with what we want, they still do enjoy a lot of desirable properties. Therefore, and since it facilitates our investigations below, we will restrict ourselves to unbiased estimators in this subsection.

Yet another criterion to compare estimators is to consider their variance.

Definition 2.1.13. *Let $(\mathfrak{X},\mathcal{A},(\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical model and let a function $\tau:\Theta\to\Sigma$ be given. An unbiased estimator T for τ is called* (uniformly) minimum-variance unbiased estimator (UMVUE or MVUE) ('varianzminimierend' oder '(gleichmäßig) bester Schätzer') *if for any other unbiased estimator S for τ, and any $\vartheta\in\Theta$ we have*

$$\mathrm{Var}_\vartheta(T) \leqslant \mathrm{Var}_\vartheta(S).$$

Since for unbiased estimators S we have that

$$\mathrm{Var}_\vartheta(S) = \mathbb{E}_\vartheta[(S-\tau(\vartheta))^2],$$

in this case the variance is a measure for the fluctuations of T around τ. The function $(S-\tau(\vartheta))^2$ is an example of a *loss function ('Verlustfunktion')* which measures the 'loss' that we incur if we estimate the real parameter $\tau(\vartheta)$ by S. There are different kinds of loss functions, but in the same way that the variance plays a prominent role in probability theory, the loss function $(S-\tau(\vartheta))^2$ is important in statistics.

Example (Example 2.1.8 cont'd)**.** *We recall*

$$\mathrm{Var}_N(T_n) = \mathbb{E}_N[T_n^2] - \mathbb{E}_N[T_n]^2 \tag{2.1.11}$$

and thus compute

$$\begin{aligned}\mathbb{E}_N[T_n^2] &= \sum_{k=1}^{N} k^2 \cdot \mathbb{P}_N(T_n = k)\\ &= \sum_{k=1}^{N} k^2\big(\mathbb{P}_N(\max\{X_i \,:\, 1 \leqslant i \leqslant n\} \leqslant k) - \mathbb{P}_N(\max\{X_i \,:\, 1 \leqslant i \leqslant n\} \leqslant k-1)\big)\\ &= \sum_{k=1}^{N} k^2\big(\mathbb{P}_N(X_1 \leqslant k)^n - \mathbb{P}_N(X_1 \leqslant k-1)^n\big)\\ &= \sum_{k=1}^{N} k^2\left(\Big(\frac{k}{N}\Big)^n - \Big(\frac{k-1}{N}\Big)^n\right)\\ &= \frac{1}{N^n}\sum_{k=1}^{N} k^2\big(k^n - (k-1)^n\big).\end{aligned}$$

Note that for all $k \leqslant N-2$, $k^n \in o_n((N-1)^n)$, where for any functions f we say that an error term g 'is in $o_n(f)$' if we have $g(n)/f(n) \to 0$ as $n \to \infty$, and the subscript n is to emphasize which variable we are sending to infinity. We thus obtain

$$\begin{aligned}\mathbb{E}_N[T_n^2] &= \frac{1}{N^n}\left(N^{n+2} - N^2\cdot(N-1)^n + (N-1)^{n+2} + o_n\big((N-1)^n\big)\right)\\ &= N^2 + \left(\frac{N-1}{N}\right)^n (1 - 2N + o_n(1))\,.\end{aligned}$$

Similarly we obtain

$$\begin{aligned}\mathbb{E}_N[T_n] &= \frac{1}{N^n}\sum_{k=1}^{N} k(k^n - (k-1)^n)\\ &= \frac{1}{N^n}\left(N^{n+1} - N(N-1)^n + (N-1)^{n+1} + o_n\big((N-1)^n\big)\right)\\ &= N + \left(\frac{N-1}{N}\right)^n (-1 + o_n(1))\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}_N[T_n]^2 &= N^2\left(1 + \frac{1}{N}\left(\frac{N-1}{N}\right)^n (-1 + o_n(1))\right)^2\\ &= N^2\left(1 + \frac{2}{N}\left(\frac{N-1}{N}\right)^n (-1 + o_n(1))\right).\end{aligned}$$

As a consequence, in combination with (2.1.11) *we deduce that*

$$\mathrm{Var}_N(T_n) = \left(\frac{N-1}{N}\right)^n (1 - 2N + 2N + o_n(1)) = \left(\frac{N-1}{N}\right)^n (1 + o_n(1)).$$

On the other hand, for the estimator $\widetilde{T}_n$, we obtain in combination with Bienaymé's lemma that

$$\mathrm{Var}_N(\widetilde{T}_n) = \mathrm{Var}\left(2\frac{\sum_{i=1}^n X_i}{n} - 1\right) = \frac{4}{n^2} n\,\mathrm{Var}(X_1),$$

hence,

$$\mathrm{Var}_N(\widetilde{T}_n) = \frac{C}{n}$$

for some constant $C = C(N) \in (0,\infty)$. In particular, this implies that the variance of T_n decays significantly faster as $n \to \infty$ than that of $\widetilde{T}_n$. Note that T_n is however biased, but we have $\mathbb{E}_N[T_n] \to N$ as $n \to \infty$.

We have seen in the above example two sequences of unbiased estimators, and we have been able to show that the variances of one of the sequences decays asymptotically faster than that of the other. This immediately leads to the question of how small the variance of (unbiased) estimators can possibly be. For this purpose, we start with introducing some further information.

Definition 2.1.14. *Let a one parameter standard model $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be given. The model is called* regular *if the following conditions are satisfied:*

(a) $\Theta \subset \mathbb{R}$ is an open interval;

(b) the likelihood function ϱ is strictly positive on $\mathfrak{X} \times \Theta$, and continuously differentiable in $\vartheta \in \Theta$. In particular the so-called score function *U_ϑ defined via*

$$U_\vartheta(x) := \frac{\partial}{\partial\vartheta} \ln \varrho(x, \vartheta) = \frac{\varrho'_x(\vartheta)}{\varrho_x(\vartheta)} \tag{2.1.12}$$

exists.

(c) For each $\vartheta \in \Theta$ the variance

$$I(\vartheta) := \mathrm{Var}_\vartheta(U_\vartheta) \tag{2.1.13}$$

exists in $(0, \infty)$, and the following interchange of differentiation and integration is valid:

$$\int \frac{\partial}{\partial\vartheta} \varrho(x, \vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta} \int \varrho(x, \vartheta)\,\mathrm{d}x; \tag{2.1.14}$$

(this is for the case of a continuous model; in the discrete case the integrals in (2.1.14) *have to be replaced by sums).*

The function I defined in (2.1.13) *is called the* Fisher information *(English statistician Sir R. Fisher (1890–1962)). It is one way to measure the information that a random variable X distributed according to $\mathbb{P}_\vartheta$ contains about ϑ.*

Remark 2.1.15. *(a) You might have seen rules of when an equality of the type* (2.1.14) *is certainly valid; for example and essentially due to the bounded convergence theorem, it holds true when ϱ'_x is continuous and for some $\varepsilon > 0$ as well as all $\vartheta \in \Theta$ we have*

$$\int \sup_{\substack{\vartheta'\in\Theta \\ |\vartheta'-\vartheta|\leqslant\varepsilon}} |\varrho'_x(\vartheta')|\,\mathrm{d}x < \infty.$$

(b) (2.1.14) *implies that the random variable U_ϑ is centred. Indeed, in the continuous case, since we are in the setting of a standard model,* (2.0.3) *in combination with the definition of U_ϑ in* (2.1.12) *implies*

$$\mathbb{E}_\vartheta[U_\vartheta] = \int \frac{\varrho'_x(\vartheta)}{\varrho_x(\vartheta)} \cdot \varrho_x(\vartheta)\,\mathrm{d}x \overset{(2.1.14)}{=} \frac{\mathrm{d}}{\mathrm{d}\vartheta} \underbrace{\int \varrho(x, \vartheta)\,\mathrm{d}x}_{=1} = 0. \tag{2.1.15}$$

The discrete case is left as an exercise.

Definition 2.1.16. *An unbiased estimator T for τ is called* regular ('regulär') *if for all $\vartheta \in \Theta$,*

$$\int T(x) \frac{\partial\varrho}{\partial\vartheta}(x, \vartheta)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}\vartheta} \int T(x)\varrho(x, \vartheta)\,\mathrm{d}x.$$

Exercise 2.1.17. *The Fisher-information is additive in the following sense: Given a standard statistical model $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ with Fisher information I, then the Fisher-information of the product model $(\mathfrak{X}^n, \mathcal{A}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ is given by $I^{\otimes n} = nI$.*

We are now in the position to prove the following result, which provides a bound for how small the variance of an unbiased estimator can get.

Theorem 2.1.18 (Information inequality / Cramér-Rao inequality (Swedish mathematician Harald Cramér (1893–1985) and Indian-American statistician Calyampudi Radhakrishna Rao (born 1920))**.** *Assume a regular statistical model $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ to be given. Furthermore, let $\tau : \Theta \to \mathbb{R}$ differentiable be given and let T be an unbiased regular estimator for τ. Then for all $\vartheta \in \Theta$:*

$$\operatorname{Var}_\vartheta(T) \geqslant \frac{\tau'(\vartheta)^2}{I(\vartheta)}. \tag{2.1.16}$$

Proof. Using that $\mathbb{E}_\vartheta[U_\vartheta] = 0$ (see (2.1.15)) we obtain the first equality in

$$\begin{aligned}\operatorname{Cov}_\vartheta(T, U_\vartheta) = \mathbb{E}_\vartheta[TU_\vartheta] &= \int_{\mathfrak{X}} T(x)\frac{\varrho'_x(\vartheta)}{\varrho_x(\vartheta)} \cdot \varrho(x,\vartheta)\,\mathrm{d}x = \int_{\mathfrak{X}} T(x)\frac{\partial}{\partial\vartheta}\varrho(x,\vartheta)\,\mathrm{d}x \\ &= \frac{\mathrm{d}}{\mathrm{d}\vartheta}\underbrace{\int_{\mathfrak{X}} T(x)\varrho(x,\vartheta)\,\mathrm{d}x}_{=\mathbb{E}_\vartheta[T]=\tau(\vartheta)} = \tau'(\vartheta),\end{aligned}$$

where the last equality follows from the assumption that T is an unbiased estimator for τ. Applying Hölder's inequality to $|\operatorname{Cov}_\vartheta(T, U_\vartheta)|$ we deduce in combination with the above that

$$\operatorname{Var}_\vartheta(T)^{\frac{1}{2}} I(\vartheta)^{\frac{1}{2}} \geqslant |\tau'(\vartheta)|,$$

which implies (2.1.16). □

Theorem 2.1.18 gives a lower bound for the variance of certain unbiased estimators. In fact, in the case of a product model satisfying the assumptions of Theorem 2.1.18, we obtain in combination with Exercise 2.1.17 that

$$\operatorname{Var}_\vartheta(T_n) \geqslant \frac{1}{n}\frac{\tau'(\vartheta)^2}{I(\vartheta)},$$

where T_n is any unbiased estimator for the product model, and I is the Fisher information for any of the n factors constituting the product model.
In particular, we cannot find any better such estimators, and therefore the ones that attain this bound deserve their own name.

Definition 2.1.19. *If for an estimator T of τ one has equality in* (2.1.16)*, then T is called* (Cramér-Rao) efficient ('(Cramér-Rao) effizient').

2.1.4 Consistency of estimators

Already in Example 2.1.8 we have observed the situation that a sequence (T_n) of estimators consisted of biased estimators, but that in at least an asymptotic sense one could observe an unbiased limiting behavior. In this section we are going to formalize this concept.
Consider the case of a statistical product model as described in Example 2.0.3. We want to have a sequence of estimators (T_n) for τ such that $T_n : \mathbb{R}^n \to \Sigma$, and one criterion that on the one hand would be desirable to have, and on the other hand is reasonable to demand, too, would be that $T_n(X_1, \dots, X_n)$ converges to $\tau(\vartheta)$ as $n \to \infty$ (where, as before, the $X_1, \dots, X_n$ denote the different coordinates of the entire observation, i.e. $X_i : \mathbb{R}^{\mathbb{N}} \ni (x_n)_{n\in\mathbb{N}} \mapsto x_i$). As seen before, there are different kinds of convergence that are of importance to us.
When introducing the next definitions it turns out to be useful to consider all the X_n to be defined on the product space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ (recall that we argued at the end of Section 1.14.3 that we are actually allowed to do so). If (X_n) is an i.i.d. sequence such that each X_n has law $\mathbb{P}_\vartheta$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then we denote by $\mathbb{P}_\vartheta^\infty$ the probability measure on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ such that $\mathbb{P}_\vartheta^\infty((X_{k_1}, \dots, X_{k_m}) \in \cdot)$ is the law of the vector $(X_{k_1}, \dots, X_{k_m})$ for any admissible choice of indices $k_1, \dots, k_m$ (if you feel uncomfortable with the infinite product measure $\mathbb{P}_\vartheta^\infty$ you can also most of the time just continue working with $\mathbb{P}_\vartheta^n$, where you have to keep adjusting the n according to the context).

Definition 2.1.20. *(a) A sequence of estimators (T_n) for τ is called* strongly consistent *if for all $\vartheta \in \Theta$ we have that $\mathbb{P}_\vartheta^\infty$-almost surely,*

$$T_n(X_1, \dots, X_n) \to \tau(\vartheta) \quad \text{as } n \to \infty.$$

(b) A sequence of estimators (T_n) *for* τ *is called* weakly consistent *if for all* $\vartheta \in \Theta$ *we have that for all* $\varepsilon > 0$,

$$\mathbb{P}_\vartheta^\infty\big(|T_n(X_1,\dots,X_n) - \tau(\vartheta)| \geqslant \varepsilon\big) \to 0 \quad \text{as } n \to \infty.$$

It is apparent that the above two definitions are inspired by the weak and strong laws of large numbers, in terminology as well as in their very definitions.
.

Theorem 2.1.21. *If a sequence* (T_n) *of estimators is strongly consistent, then it is also weakly consistent.*

Proof. Exercise □

Example 2.1.22. *We revisit the setting of Example 2.1.8. Although we have seen that the* (T_n) *of Example 2.1.8 (a) defined in* (2.1.6) *are biased, they (as well as their unbiased version* T^* *defined in* (2.1.7)*) form a strongly consistent sequence for* $\tau(N) = N$. *Indeed, as seen before in* (2.1.5)*, we have for any* $\varepsilon > 0$ *that*

$$\mathbb{P}_N^\infty(|T_n(X_1,\dots,X_n) - N| > \varepsilon) \leqslant \mathbb{P}_N^\infty(\max\{X_1,\dots,X_n\} < N) = \Big(\frac{N-1}{N}\Big)^n.$$

Since the RHS of this display is summable in $n \in \mathbb{N}$, *Theorem 1.13.1 (d) tells us that* $\mathbb{P}_N$*-a.s.,*

$$\lim_{n\to\infty} T_n(X_1,\dots,X_n) = \tau(N) = N.$$

Furthermore, Theorem 2.1.21 supplies us with the fact that the sequence is weakly consistent as well. Also the $(\widetilde{T}_n)$ *of Example 2.1.8 (a) form a strongly consistent sequence, for the strong law of large numbers supplies us with*

$$\widetilde{T}_n(X_1,\dots,X_n) = 2\frac{\sum_{i=1}^n X_i}{n} - 1 \overset{n\to\infty}{\longrightarrow} 2\frac{N}{2} = N, \quad \mathbb{P}_N^\infty - a.s.$$

As above, we can use Theorem 2.1.21 to argue that the sequence $(\widetilde{T}_n)$ *is weakly consistent,*

In fact, we are able to establish consistency for much more general sequences of estimators.

Theorem 2.1.23. *Assume that the sequence* (X_n) *denotes the coordinates in the infinite product model (as before, the generalization of Example 2.0.3 to the infinite context as introduced at the beginning of Section 2.1.4, and with the same notation* $\mathbb{P}_\vartheta^\infty$*). In addition, assume that each factor in the product model is of the form* $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B}(\mathbb{R}))$ *and that for each* $\vartheta \in \Theta$ *we have that* $\mathbb{E}_\vartheta[X_1]$ *as well as* $\mathrm{Var}_\vartheta(X_1)$ *exist. We denote by* $T_n := \frac{1}{n}\sum_{i=1}^n X_i$ *and* $\widetilde{T}_n := \frac{1}{n-1}\sum_{i=1}^n (X_i - T_n)^2$ *the two sequences of canonical unbiased estimators for the expectation* $\mathbb{E}_\vartheta[X_1]$ *(or, more precisely, for the function* $\tau : \Theta \to \mathbb{R}$, $\vartheta \mapsto \mathbb{E}_\vartheta[X_1]$*) as well as for the variance* $\mathrm{Var}_\vartheta(X_1)$ *(analogously; cf. also Examples 2.1.4 (a) and 2.1.9)*
Then (T_n) *and* $(\widetilde{T}_n)$ *form strongly consistent sequences of estimators.*

Proof. The strong law of large numbers implies that for all $\vartheta \in \Theta$ we have $\mathbb{P}_\vartheta^\infty$-a.s.,

$$T_n = \frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}_\vartheta[X_1], \quad \text{as } n \to \infty,$$

and hence (T_n) is a strongly consistent sequence.
To prove that $(\widetilde{T}_n)$ defines a strongly consistent sequence for the variance, we rewrite

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^n (X_i - T_n)^2 &= \frac{1}{n-1}\sum_{i=1}^n \Big((X_i - \mathbb{E}_\vartheta[X_i])^2 + 2(X_i - \mathbb{E}_\vartheta[X_i])(\mathbb{E}_\vartheta[X_1] - T_n) - (\mathbb{E}_\vartheta[X_1] - T_n)^2\Big)\\ &= \underbrace{\frac{1}{n-1}\sum_{i=1}^n (X_i - \mathbb{E}_\vartheta[X_i])^2}_{\to \mathrm{Var}_\vartheta(X_1)\ \mathbb{P}_\vartheta^\infty - a.s.\ \text{as } n\to\infty} + \underbrace{\frac{1}{n-1}\sum_{i=1}^n \Big(2(X_i - \mathbb{E}_\vartheta[X_i])(\mathbb{E}_\vartheta[X_1] - T_n) - (\mathbb{E}_\vartheta[X_1] - T_n)^2\Big)}_{\to 0\ \mathbb{P}_\vartheta^\infty - a.s.\ \text{as } n\to\infty},\end{aligned}$$

where the first convergence follows from the strong law of large numbers, and the second from the fact that $(\mathbb{E}_\vartheta[X_1] - T_n) \to 0$ $\mathbb{P}_\vartheta^\infty$-a.s. in combination with the fact that (T_n) is a strongly consistent sequence and the strong law of large numbers.

□

Yet another important class of estimators is that of maximum likelihood estimators.

Theorem 2.1.24. *Let a one parameter standard model as in Theorem 2.1.23 be given. Furthermore, assume the following conditions to be fulfilled:*

(a) $\Theta \subset \mathbb{R}$ *is an open interval;*

(b)

$$\text{for all } \vartheta_1, \vartheta_2 \in \Theta \text{ with } \vartheta_1 \neq \vartheta_2 : \quad \mathbb{P}_{\vartheta_1} \neq \mathbb{P}_{\vartheta_2}; \tag{2.1.17}$$

(c) for each $n \in \mathbb{N}$ *and all* $x \in \mathbb{R}^n$*, the function*

$$\Theta \ni \vartheta \mapsto \varrho^{\otimes n}(x, \vartheta) := \prod_{i=1}^{n} \varrho(x_i, \vartheta)$$

is unimodal;[7] *in particular, there is a unique maximum likelihood estimator* $T_n(x) := \vartheta(x) \in \Theta$ *for* $\tau(\vartheta) = \vartheta$ *such that for all* $\vartheta \in \Theta$ *with* $\vartheta \neq T_n(x)$ *we have*

$$\varrho^{\otimes n}(x, \vartheta) < \varrho^{\otimes n}(x, T_n(x)).$$

Then the sequence (T_n) *is weakly consistent for* $\tau(\vartheta) = \vartheta$*.*

Remark 2.1.25. *It follows from the computations of Example 2.1.12 that the conditions of the Theorem (and in particular the unimodality) are fulfilled in the case of the* X_n *being i.i.d. Gaussian.*

In order to prove this result, we have to introduce another fundamental concept of statistics, namely that of relative entropy, and some of its properties.

Definition 2.1.26. *Let* P *and* Q *be two discrete or continuous probability measures on* $(\mathbb{R}, \mathcal{B})$ *(with densities* p *and* q*; or in the discrete case, these functions denote the one point probabilities). Their* relative entropy *or* Kullback-Leibler distance[8] *is defined as*

$$H(P \,|\, Q) := \begin{cases} \mathbb{E}_P\Big[\ln\Big(\frac{p}{q}\Big)\Big], & \text{if } P(q = 0) = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

A very nice and simple introduction to the concept of relative entropy, which also highlights its importance in statistics, is given in [Geo03].

Claim 2.1.27. *In the context of Definition 2.1.26,*

(a) If $P(q = 0) = 0$*, then*

$$H(P \,|\, Q) = \mathbb{E}_Q\Big[\frac{p}{q} \ln\Big(\frac{p}{q}\Big)\Big] = \int p(x) \ln\Big(\frac{p(x)}{q(x)}\Big)\, \mathrm{d}x,$$

with $0 \ln 0 := 0$*.*

Exercise 2.1.28. *Do we necessarily have* $H(P \,|\, Q) < \infty$ *in this case? (non-trivial)*

(b) For all probability measures P *and* Q *on* $(\mathbb{R}, \mathcal{B})$ *as in Definition 2.1.26,*

$$H(P \,|\, Q) \geqslant 0;$$

(c) $H(P \,|\, Q) = 0$ *if and only if* $P = Q$*.*

Proof sketch. (a) This is essentially a consequence of the change of variable formula Proposition 1.9.10.

[7] A function $f : \Theta \to \mathbb{R}$ is called 'unimodal' if there exists $m \in \Theta$ such that f is increasing on $(-\infty, m)$, decreasing on $(-\infty, m)$, and $f(m)$ is the unique global maximum. The value of m is referred to as the *mode*.

[8] Note, however, that d defined via $d(P, Q) := H(P \,|\, Q)$ does not denote a metric on the set of probability measures (on a particular space); in fact, it is not symmetric, and even the symmetrisation $\widetilde{H}(P \,|\, Q) := (H(P \,|\, Q) + H(Q \,|\, P))/2$ does not satisfy the triangle inequality.

(b) Jensen's inequality applied with the convex function $\varphi : [0,\infty) \ni x \mapsto x \ln x$ implies

$$\mathbb{E}_Q\Big[\frac{p}{q}\ln\Big(\frac{p}{q}\Big)\Big] \geqslant \ln\Big(\mathbb{E}_Q\Big[\frac{p}{q}\Big]\Big) = \ln(\mathbb{E}_P[1]) = 0.$$

(c) If $P = Q$, then P-a.s. we have $p = q$, so $p/q = 1$, hence $H(P\,|\,Q) = 0$ by definition. If, on the other hand, $P \neq Q$, then $Q(\{x \in \mathbb{R} : p(x) \neq q(x)\}) > 0$. Using the representation $H(P\,|\,Q) = \mathbb{E}_Q\Big[\frac{p}{q}\ln\Big(\frac{p}{q}\Big)\Big]$ established before, we can now apply the strict version of Jensen's inequality (cf. Theorem 1.12.9). Indeed, we check (e.g. by differentiating twice to get that its second derivative is strictly positive) that the function $(0,\infty) \ni x \mapsto x\ln x$ is strictly convex and the underlying probability measure Q is not concentrated in a single point. Thus, Jensen's inequality is strict and we infer $H(P\,|\,Q) > 0$ in this case.

□

Proof of Theorem 2.1.24. Let the fixed parameter $\vartheta \in \Theta$ be given and choose $\varepsilon > 0$ sufficiently small such that $[\vartheta-\varepsilon, \vartheta+\varepsilon] \subset \Theta$.
Using (2.1.17) in combination with Claim 2.1.27 (c) we deduce that there exists $\delta > 0$ such that

$$H(\mathbb{P}_\vartheta\,|\,\mathbb{P}_{\vartheta-\varepsilon}) \wedge H(\mathbb{P}_\vartheta\,|\,\mathbb{P}_{\vartheta+\varepsilon}) > \delta. \tag{2.1.18}$$

We start with observing that if

$$\varrho^{\otimes n}(x, \vartheta-\varepsilon) < \varrho^{\otimes n}(x,\vartheta) \tag{2.1.19}$$

as well as

$$\varrho^{\otimes n}(x, \vartheta+\varepsilon) < \varrho^{\otimes n}(x,\vartheta), \tag{2.1.20}$$

then, since $\varrho^{\otimes n}(x,\cdot)$ is unimodal, its maximum must be attained within the interval $(\vartheta-\varepsilon, \vartheta+\varepsilon)$. In particular, since $T_n(x)$ is a maximum-likelihood estimator, this implies $T_n(x) \in (\vartheta-\varepsilon, \vartheta+\varepsilon)$.
In particular, (2.1.19) and (2.1.20) are fulfilled for

$$x \in \Big\{y \in \mathbb{R}^n : \frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}(y)}{\varrho_{\vartheta+\varepsilon}^{\otimes n}(y)} > 0\Big\} \bigcap \Big\{y \in \mathbb{R}^n : \frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}(y)}{\varrho_{\vartheta-\varepsilon}^{\otimes n}(y)} > 0\Big\}, \tag{2.1.21}$$

whence it is sufficient to show that the probability of the right-hand side of the previous display under $\mathbb{P}_\vartheta^\infty$ (or $\mathbb{P}_\vartheta^n$ if you prefer) converges to 1 as $n \to \infty$.
The strong law of large numbers tells us that $\mathbb{P}_\vartheta^\infty$-a.s.,

$$\frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}}{\varrho_{\vartheta+\varepsilon}^{\otimes n}} \to \mathbb{E}_\vartheta\Big[\ln\Big(\frac{\varrho_\vartheta}{\varrho_{\vartheta+\varepsilon}}\Big)\Big] = H(\mathbb{P}_\vartheta\,|\,\mathbb{P}_{\vartheta+\varepsilon}) > \delta \quad \text{if } \mathbb{P}_\vartheta(\varrho_{\vartheta+\varepsilon} = 0) = 0,$$

and similarly

$$\frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}}{\varrho_{\vartheta-\varepsilon}^{\otimes n}} \to \mathbb{E}_\vartheta\Big[\ln\Big(\frac{\varrho_\vartheta}{\varrho_{\vartheta-\varepsilon}}\Big)\Big] = H(\mathbb{P}_\vartheta\,|\,\mathbb{P}_{\vartheta-\varepsilon}) > \delta \quad \text{if } \mathbb{P}_\vartheta(\varrho_{\vartheta-\varepsilon} = 0) = 0.$$

In the case $\mathbb{P}_\vartheta(\varrho_{\vartheta+\varepsilon} = 0) > 0$ we obtain

$$\mathbb{P}_\vartheta^n\Big(\frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}}{\varrho_{\vartheta+\varepsilon}^{\otimes n}} = \infty\Big) = 1 - (1 - \mathbb{P}_\vartheta(\varrho_{\vartheta+\varepsilon} = 0))^n \to 1,$$

and similarly for $\mathbb{P}_\vartheta^n\Big(\frac{1}{n}\ln\frac{\varrho_\vartheta^{\otimes n}}{\varrho_{\vartheta-\varepsilon}^{\otimes n}} = \infty\Big)$.
Hence, in combination with (2.1.18) we deduce that the $\mathbb{P}_\vartheta^n$-probability of the right-hand side of (2.1.21) converges to 1, which implies the desired weak consistency of the sequence (T_n). □

We will see an application of the above results in the exercise classes.

2.2 Confidence regions

In the previous section we have mainly been concerned with trying to find estimators (or sequences of estimators) for $\tau(\vartheta)$ with more or less desirable properties such as being unbiased, having small variance, or being consistent. The hope was that these estimators gave us the 'correct' value of $\tau(\vartheta)$.
In most situations, however, an estimator which is a function of finitely many observations will not provide the correct value anyway. Indeed, recall the estimator of Example 2.0.1 defined in (2.1.3). If we had chosen a different sample of the population, the estimator would presumably have given us a value different from that of the first sample. However, the underlying ϑ (or the number of sick persons in the entire population for that matter) would still have been the same.
Therefore, given some observed data, oftentimes, rather than trying to find the one right value $\tau(\vartheta)$, one is interested in finding a subset (depending on the observed data) of Σ such that (in some sense) one can be confident that the correct $\tau(\vartheta)$ is actually contained in this subset.
In fact, Example 2.1.17 even tells us that under suitable assumptions any unbiased estimator in the n-fold product model has variance bounded below by $\frac{c}{n}$, some constant $c \in (0,\infty)$, as $n \to \infty$.
This leads us to the following definition.

Definition 2.2.1. *Let $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ be a statistical model, $\tau : \Theta \to \Sigma$ an arbitrary function, and $\alpha \in (0,1)$. A* confidence region for τ with error level α ('Konfidenzbereich für τ zum Fehlerniveau α') *is a mapping $C : \mathfrak{X} \to 2^\Sigma$ such that*[9]

$$\inf_{\vartheta\in\Theta} \mathbb{P}_\vartheta\big(x \in \mathfrak{X} : \tau(\vartheta) \in C(x)\big) \geqslant 1-\alpha. \tag{2.2.1}$$

If for all $x \in \mathfrak{X}$, the set $C(x)$ is actually an interval, then C is called 'confidence interval'('Konfidenzintervall') instead of 'confidence region'.

Remark 2.2.2. *(a) The condition in* (2.2.1) *is certainly fulfilled if we choose the constant confidence region $C(x) \equiv \Sigma$. This, however, would not be of too much use to us. In fact, we aim for choosing $C(x)$ as small as possible as this gives us more information on the real value of $\tau(\vartheta)$ (with high probability).*

(b) It is important to interpret (2.2.1) *in the right way: If we choose e.g. $\alpha := 0.05 = 5\%$,*[10] *then, given an outcome $x \in \mathfrak{X}$,* (2.2.1) *does not tell us that in 95% of the cases ϑ is contained in $C(x)$ (in fact, ϑ is not known, it is fixed nevertheless, and we do not have a probability distribution on Θ either; so it does not make sense to say that in 95% of the cases ϑ has a certain property).*

Rather, (2.2.1) *is to be interpreted in the following way: No matter what the real value of ϑ is, in at least 95% of the realizations x that we observe under $\mathbb{P}_\vartheta$, we will have that $\tau(\vartheta) \in C(x)$ (i.e., the probability of those observations x for which we have $\tau(\vartheta) \notin C(x)$ is bounded from above by 5%).*

2.2.1 One recipe for constructing confidence regions

While there are of course many ways to choose confidence regions, there is a sort of canonical procedure that suggests itself in the standard model. We will illustrate it for

$$\tau(\vartheta) = \vartheta \tag{2.2.2}$$

for the sake of simplicity:
Let $\alpha \in (0,1)$ be given.

(a) For any $\vartheta \in \Theta$, choose $C_\vartheta \in \mathcal{A}$ in such a way that

$$\mathbb{P}_\vartheta(C_\vartheta) \geqslant 1-\alpha.$$

In the case of a standard model where we have (2.0.3) and a density ϱ_ϑ at our disposal, one possibility is to choose

$$C_\vartheta := \varrho_\vartheta^{-1}([s^*,\infty)),$$

where

$$s^* := \sup\Big\{s \in \mathbb{R} : \underbrace{\int \mathbb{1}_{\varrho_\vartheta(x)\geqslant s}(x)\cdot\varrho_\vartheta(x)\,\mathrm{d}x}_{\mathbb{P}_\vartheta(x\in\mathfrak{X}\,:\,\varrho_\vartheta(x)\geqslant s)} \geqslant 1-\alpha\Big\}$$

[9] In particular, for the probability in (2.2.1) to be well-defined, we need $\{x \in \mathfrak{X} : \tau(\vartheta) \in C(x)\} \in \mathcal{A}$.
[10] typical values for α are e.g. 0.05 and 0.01;

(and where in the discrete case the integral is as always replaced by a sum); this choice of C_ϑ means that we choose C_ϑ in such a way that

- it consist of the highest possible values of ϱ_ϑ, and
- at the same time is as small as possible (in terms of the $\mathbb{P}_\vartheta$-probability of subsets of $\mathfrak{X}$). (However, we could as well have chosen any other $C_\vartheta \in \mathcal{A}$ such that $\mathbb{P}_\vartheta(C_\vartheta) \geqslant 1-\alpha$.)

(b) We now define

$$C(x) := \{\vartheta \in \Theta \,:\, x \in C_\vartheta\}, \tag{2.2.3}$$

which does the job.

Remark 2.2.3. *The Reader may convince herself that if τ is not of the form as in* (2.2.2)*, then the mapping*

$$\begin{aligned} \mathfrak{X} &\to 2^\Sigma \\ x &\mapsto \tau(C(x)) \end{aligned}$$

with $C(x)$ as in (2.2.3) *still provides us with a confidence region for τ with error level α. However, one might at times be able to devise 'better' confidence regions which take into consideration the specific structure of τ.*

In the literature you find formulae and computations for confidence intervals for common distributions and different error levels. We will not go into those details but rather present another approach for computing confidence intervals of the median (which is independent of the underlying distribution!).

2.2.2 Order statistics

Although we have seen that the Central Limit Theorem (Theorem 1.15.1) implies that the normal distribution takes a pre-eminent role under distributions, it might not always be the case that it is reasonable to assume that certain data are obtained as realization of normally distributed random variables, or of any other preselected distribution: We might want to be able to obtain some reasonable statements without the assumption that the data observed has been obtained according to a fixed distribution.
For the purpose of introducing order statistics, it seems reasonable to get acquainted to the following definition. For the sake of simplicity it suggests itself to

$$\text{assume that the law according to the observations } X_1, \ldots, X_n \text{ on } (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \tag{2.2.4}$$

$$\text{is realized has a continuous distribution.} \tag{2.2.5}$$

This is not really necessary but keeps overly technical details out of our focus.

Definition 2.2.4. *The* order statistics ('Ordnungsstatistik') $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ *of the random variables $X_1, \ldots, X_n$ is defined recursively as*

$$X_{1:n} := \min\{X_1, \ldots, X_n\},$$

and given $X_{1:n}, \ldots, X_{j:n}$ for some $j \in \{1, \ldots, n-1\}$, we define

$$X_{j+1:n} := \min\{X_k \,:\, X_k > X_{j:n}\}.$$

The assumption of (2.2.4) ensure that almost surely with respect to the underlying probability measure,

$$X_{1:n} < X_{2:n} < \ldots < X_{n:n}.$$

As you might imagine, without any further assumptions (in particular on the tails of the distribution of the X_i) one might have a hard time trying to reasonably estimate central parameters such as the expectation with a guaranteed probability (e.g. in the sense of confidence regions – which is hard to realise here since we do not have a parametric family of potential probability measures at our disposal); even more so, it is not even clear whether the expectation of the underlying distribution exists at all. There is, however, another characteristic quantity of the underlying distribution which is of significant relevance and which is less prone to the occurrence of outliers in sample data: the median.

Definition 2.2.5. *Let a probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be given. For $\alpha \in (0,1)$, any real number q with the property that*

$$\mathbb{P}((-\infty, q]) \geqslant \alpha \quad \text{and simultaneously} \quad \mathbb{P}([q, -\infty)) \geqslant 1-\alpha$$

is called an α-quantile of $\mathbb{P}$ ('α-Quantil von $\mathbb{P}$'). *An $\frac{1}{2}$-quantile is called* median ('Median') *(often denoted $m(\mathbb{P})$) and the quantiles at level $\frac{1}{4}$ and $\frac{3}{4}$ are called* lower ('unteres') *and* upper quantile ('oberes Quantil'), *respectively.*
For a real random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\alpha \in (0,1)$, the α-quantile of X *is defined as the α-quantile of its distribution $\mathbb{P} \circ X^{-1}$. The median of X is often denoted by $m(X)$.*

Exercise 2.2.6. *Show that in the setting of Definition 2.2.5, for any $\alpha \in (0,1)$, the α-quantile exists.*

Example 2.2.7. *(a) For $p \in (0,1)$ consider the Bernoulli distribution with parameter p. If $p > \frac{1}{2}$, then its median is given by 1, whereas for $p < \frac{1}{2}$ its median is 0. For $p = \frac{1}{2}$, any number in $[0,1]$ is a median.*

(b) Quantiles of a $\mathcal{N}(\mu, \sigma^2)$ distributed random variable play important roles in statistics (e.g. in confidence regions or tests). Since they are non-trivial to compute there are quantile tables which containing the quantiles for wide ranges of parameters. The median, however, is easy to compute since it just coincides with is expectation (and is uniquely determined) μ.

The very definition of the median and the fact that we have a total order on $\mathbb{R}$ ensure that it can be estimated nicely using the binomial distribution lurking behind. For this purpose we denote

$$q_n(\alpha) := \max\left\{k \in \{1, \dots, n\} \,:\, \mathrm{Bin}_{n,\frac{1}{2}}(\{0, \dots, k-1\}) \leqslant \alpha\right\}.$$

Theorem 2.2.8. *Let (X_n) be a sequence of i.i.d. real random variables with a continuous distribution $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then for $n \in \mathbb{N}$ and $\alpha \in (0,1)$ we have that*

$$[X_{q_n(\frac{\alpha}{2}):n}, X_{n-q_n(\frac{\alpha}{2}):n}]$$

is a confidence interval for the median $m(X_1)$ of X_1 with error level α.

Proof. Using the standard product model notation, from the very definition of the median we obtain that

$$\mathbb{P}^n\big(m(X) < X_{q_n(\frac{\alpha}{2}):n}\big) = \mathrm{Bin}_{n,\frac{1}{2}}(\{0, \dots, q_n(\alpha/2) - 1\}) \leqslant \alpha/2,$$

where the last inequality follows from the definition of $q_n(\alpha/2)$. Similarly for $\mathbb{P}^n\big(m(X) > X_{n-q_n(\frac{\alpha}{2}):n}\big)$, which completes the proof. □

In a slightly more complicated manner than what is done in the proof of Theorem 2.2.8, we can obtain confidence intervals not only for the median but also for other quantiles of a distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (and again, these confidence intervals are independent of the underlying distribution $\mathbb{P}$).

2.3 Tests

In the previous sections we have devoted our investigations to estimating parameters as well as finding confidence regions (depending on certain observed data) such that if a value $\tau(\vartheta) \in \Sigma$ was not contained in the corresponding confidence region, then the observed data was unlikely to be observed under $\mathbb{P}_\vartheta$.
In this section, instead of trying to understand the underlying distribution $\mathbb{P}_\vartheta$, we are just interested in accepting or rejecting certain hypotheses. Again, as we have seen before e.g. in the context of estimators, a statistician does have some freedom in how to choose the tests she applies, and usually different choices will have different pros and cons. It is part of the statistician's job to try to figure out what's suited best to her current needs.
In order to be able to accept or reject hypotheses, we have to fix a model as well as a couple of parameters. This is illustrated in the following example.

Example 2.3.1. *Assume you are a retailer and have ordered a sizeable amount of light bulbs. The contract you signed with the manufacturer demands that at most* 1% *of the bulbs delivered may be dysfunctional. If the proportion of dysfunctional light bulbs in the batch exceeds* 1%*, then the manufacturer is obliged to pay you a penalty for non-performance. Therefore, the question you might want to resolve could be of the following type: What is the number x of broken light bulbs that you do have to accept in a small sample of size n of the entire batch (the latter having size $N \gg n$), such that if you see more than x dysfunctional light bulbs in this sample and you claim the non-performance penalty, you will be right in at least* 95% *of the cases (i.e., in at least* 95% *of the realizations with at least x bad light bulbs in the sample, the proportion of defect light bulbs in the entire batch is more than* 1%*).*

(a) We can fit this context into the following statistical model: $\mathfrak{X} = \{0, 1, \dots, n\}$ (where n is the size of the sample, $N \geqslant n$ is the size of the delivery batch) and $\Theta = \{0, 1, \dots, N\}$ is the set in which the number of defect light bulbs contained in the batch can take its values. For $\vartheta \in \Theta$ the probability $\mathbb{P}_\vartheta$ can then be defined via

$$\mathbb{P}_\vartheta := \mathrm{H}_{N,\vartheta,n},$$

where the hypergeometric distribution $\mathrm{H}_{N,\vartheta,n}$ has been introduced in Example 1.8.10 in such a way that for $k \in \{0 \vee n + \vartheta - N, \dots, n \wedge \vartheta\}$ one has

$$\mathrm{H}_{N,\vartheta,n}(k) = \frac{\binom{N-\vartheta}{n-k}\binom{\vartheta}{k}}{\binom{N}{n}}.$$

(b) We now partition the space Θ in two components Θ_0 and Θ_1 in such a way that if $\vartheta \in \Theta_0$, then we should accept the batch of bulbs as is (which is deemed the standard case, cf. Remark 2.3.5), whereas if $\vartheta \in \Theta_1$ we could claim a penalty from the manufacturer (the exceptional case). In our case, due to the contract outlined above, we immediately arrive at

$$\Theta_0 := \{0, 1, \dots, \lfloor N/100 \rfloor\},$$

and

$$\Theta_1 := \{\lfloor N/100 \rfloor + 1, \lfloor N/100 \rfloor + 2, \dots, N\}.$$

Once we have this decomposition, the usual wording is to refer to the case $\vartheta \in \Theta_0$ as the null hypothesis ('Nullhypothese') *, whereas the case $\vartheta \in \Theta_1$ is referred to as* alternative ('Alternative')*.*

(c) We now have to choose a significance level ('Signifikanzniveau') *$\alpha \in (0, 1)$ (this is very much in spirit of the confidence regions introduced in the previous section). In our example it translates to the following: We want to keep the probability reasonably small (say smaller than some level $\alpha \in (0, 1)$) that after having inspected our sample of size n, we do claim a penalty from the manufacturer, but that at the same time the real ϑ is contained in Θ_0 (i.e., we are claiming a penalty although we are not entitled to do so since the total proportion of dysfunctional light bulbs does not exceed* 1%*). Such an error of a false rejection of the null hypothesis is called a* type I error*. The other possible type of error would be a false acceptance of the null hypothesis which is called a* type II error*.*

(d) The last task is to find a rule which tells us in dependence on the observed data $x \in \mathfrak{X}$ whether or not to accept the null hypothesis. Such a rule is usually devised in terms of a statistic $T : \mathfrak{X} \to [0, 1]$, which is to be interpreted such that if for the observed data $x \in \mathfrak{X}$ we have $T(x) = 0$, then the null hypothesis would be accepted, whereas if $T(x) = 1$, then the null hypothesis should be rejected in favour of the alternative. If $T(x) \in (0, 1)$, then an additional experiment should be performed which with probability $T(x)$ tells you to go for the alternative.

The following setting has been shortly looked at in Homework sheet 8, exercise 1 c). We can now perform a more structured analysis of the situation.

Example 2.3.2. *Consider a sequence of i.i.d. coin flips $X_1, X_2, \dots, X_n$ with a coin that has probability $\vartheta \in [0, 1] =: \Theta$ to show heads (corresponding to* 1*, and tails corresponding to* 0*). We are interested in whether the coin is fair ($\vartheta = \frac{1}{2}$) or not. Therefore, we can choose $\mathfrak{X} = \{0, 1\}^n$, $\mathcal{A} = 2^{\mathfrak{X}}$, and*

$$\mathbb{P}_\vartheta(x_1, \dots, x_n) = \vartheta^{\sum_{i=1}^n x_i}(1 - \vartheta)^{n - \sum_{i=1}^n x_i}$$

(see Example 1.8.2).

As null hypothesis we choose the case that the coin is fair, hence $\Theta_0 := \{\frac{1}{2}\}$ *and* $\Theta_1 := (0, 1/2) \cup (1/2, 1)$, *and for the significance level we could e.g. choose* $\alpha = 0.05 = 5\%$ *as an upper bound for committing a type I error.*
Using a normal approximation and the fact that $\mathrm{Var}_{\frac{1}{2}}(\sum_{i=1}^n X_i) = n \mathrm{Var}_{\frac{1}{2}}(X_1) = n(\frac{1}{2} - \frac{1}{4}) = \frac{n}{4}$, *we can approximate*

$$\mathbb{P}_{\frac{1}{2}}\Big(\Big|\sum_{i=1}^n X_i - \frac{n}{2}\Big| \geqslant c_n\Big) = \mathbb{P}_{\frac{1}{2}}\Big(\Big|\sum_{i=1}^n X_i - \frac{n}{2}\Big|\Big/\sqrt{\frac{n}{4}} \geqslant \frac{c_n}{\sqrt{\frac{n}{4}}}\Big) \approx 2\Phi\Big(-\frac{c_n}{\sqrt{\frac{n}{4}}}\Big),$$

with Φ *denoting the cumulative distribution function of a* $\mathcal{N}(0, 1)$ *distributed random variable. Solving for*

$$2\Phi\Big(-\frac{c_n}{\sqrt{\frac{n}{4}}}\Big) \stackrel{!}{=} 0.05,$$

e.g. by using a quantile table for the normal distribution which supplies us with $\Phi(1.96) \approx 0.975$, *we obtain* $c_n \approx 0.98\sqrt{n}$. *Thus, if we have a sample of* n *coin tosses of a fair coin, then we only see 'extreme realizations' for which*

$$|\sum_{i=1}^n X_i - n/2| \geqslant 0.98\sqrt{n} \tag{2.3.1}$$

with a probability of less than 5% *(modulo the errors incurred in the above approximations). Therefore, if we have a significance level* $\alpha = 5\%$, *we might reject the null hypothesis if we observe such realizations. For the sake of example, in the case of* $n = 100$, (2.3.1) *corresp either more than* 59 *or less than* 41 *heads.*

Formalizing the above we arrive at the following.

Definition 2.3.3. *Let a statistical model* $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ *and a partition of* Θ *into the null hypothesis* Θ_0 *and the alternative* Θ_1 *be given.*

(a) A statistic $T : \mathfrak{X} \to [0, 1]$ *is called a test of* Θ_0 *against* Θ_1. *The subset* $T^{-1}(\{1\})$ *of* $\mathfrak{X}$ *is called* rejection region ('Ablehnungsbereich, Verwerfungsbereich, kritischer Bereich'). *On the other hand, if an observation* x *is contained in* $T^{-1}(\{0\})$, *the null hypothesis should be accepted.*

If $x \in T^{-1}((0, 1))$, *an additional random experiment (independent from everything else) should lead to a rejection of the null hypothesis with probability* $T(x)$.

(b) Among all $\vartheta \in \Theta_0$, *the smallest upper bound on the probability of committing a type I error is given by*

$$\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta[T];$$

the latter quantity is also called the size ('Umfang') *or* effective level ('effektives Niveau') *of* T. *A test* T *is said to be a* test of significance level ('Signifikanzniveau') $\alpha \in (0, 1)$ *if*

$$\sup_{\vartheta\in\Theta_0} \mathbb{E}_\vartheta[T] \leqslant \alpha. \tag{2.3.2}$$

(c) The power function ('Macht, Stärke') *of a test* T *at* $\vartheta \in \Theta_1$ *is defined as* $\mathbb{E}_\vartheta[T]$. *The function*

$$\begin{aligned} G_T : \Theta &\to [0, 1] \\ \vartheta &\mapsto \mathbb{E}_\vartheta[T] \end{aligned}$$

is called the power function of T ('Gütefunktion von T').

Thus, for $\vartheta \in \Theta_0$ *the quantity* $G_T(\vartheta)$ *is the probability of a type I error, whereas for* $\vartheta \in \Theta_1$ *it provides us with the probability of a (correct) rejection of the hypothesis.*

Remark 2.3.4. *One reason for having* T *map to* $[0, 1]$ *instead of just* $\{0, 1\}$ *in Definition 2.3.3 (a) is the following. Assume that you not only want to construct a test with significance level* $\alpha \in (0, 1)$, *but rather you want* equality *in* (2.3.2) *to hold. I.e., you are really prepared to possibly be embarrassed in a proportion exactly* α *of cases (if the underlying* ϑ *happens to be the one for which the supremum in* (2.3.2) *is attained for the test you choose to apply).*

For instance, in Example 2.3.1 we see that Θ_0 is a finite set and therefore $\mathbb{E}_\vartheta[T]$, $\vartheta \in \Theta_0$, takes finitely many values only. In particular, for all but finitely many choices of α the inequality in (2.3.2) *would be strict* for all of the finitely many tests $T : \mathfrak{X} \to \{0,1\}$. *Thus, one way to achieve equality in* (2.3.2) *is to admit so-called* randomised tests $T : \mathfrak{X} \to [0,1]$.
In fact, assume that in Example 2.3.1 we do have

$$\sup_T \sup_{\vartheta \in \Theta_0} \mathbb{E}_\vartheta[T] < \alpha, \tag{2.3.3}$$

where the first supremum is taken over all tests $T : \mathfrak{X} \to \{0,1\}$ with significance level α. We now construct a randomised test T that satisfies equality in (2.3.2). *For this purpose, we choose the minimal $k \in \mathfrak{X} = \{0,1,\dots,n\}$ such that*[11]

$$c := \sup_{\vartheta \in \Theta_0 = \{0,1,\dots,\lfloor \frac{N}{100} \rfloor\}} \mathbb{P}_\vartheta(\{k, k+1, \dots, n\}) \leqslant \alpha,$$

which, as you may convince yourself, amounts to

$$c = \mathbb{P}_{\lfloor \frac{N}{100} \rfloor}(\{k, k+1, \dots, n\}) \leqslant \alpha.$$

and furthermore (2.3.3) *then implies that the inequality in the above display is strict. Thus, we can define the test*

$$T(x) = \begin{cases} 1, & \text{if } x \in \{k, k+1, \dots, n\}, \\ \frac{\alpha - c}{\mathbb{P}_{\lfloor \frac{N}{100} \rfloor}(\{k-1\})} > 0, & \text{if } x = k-1, \\ 0, & \text{if } x \in \{0,1,\dots,k-2\}. \end{cases}$$

Hence, if one observes $x = k-1$, then an additional independent experiment is conducted that leads to rejection of the null hypothesis with probability exactly $\alpha - c$. We check that indeed we have

$$\begin{aligned} \sup_{\vartheta \in \Theta_0} \mathbb{E}_\vartheta[T] &= \mathbb{E}_{\lfloor \frac{N}{100} \rfloor}[T] = \mathbb{P}_{\lfloor \frac{N}{100} \rfloor}(\{k, k+1, \dots, n\}) + \mathbb{P}_{\lfloor \frac{N}{100} \rfloor}(\{k-1\}) \frac{\alpha - c}{\mathbb{P}_{\lfloor \frac{N}{100} \rfloor}(\{k-1\})} \\ &= c + \alpha - c = \alpha. \end{aligned}$$

Remark 2.3.5. *Given a problem such as the above, oftentimes a partition $\Theta = \Theta' \dot\cup \Theta''$ suggests itself (in this case one of the two sets Θ' and Θ'' should consist of the 'acceptable' values for the numbers of defect bulbs, the other one should consist of those values that are not acceptable and hence lead to a breach of contract). Slightly more difficult is the task to determine which of the two sets Θ' and Θ'' should correspond to the null hypothesis, and which one should be the alternative.*
Given the nature of a test (i.e., that the null hypothesis is discarded only if the data one has as an observation is reasonably unlikely with a probability smaller than α, the latter one often being chosen as 0.01 *or* 0.05*) it suggests itself to choose those values of Θ for the null hypothesis that imply acceptance of the batch; otherwise, if we chose the null hypothesis to consist of those values of Θ that would make us claim a penalty for non-performance, we would generally claim a non-performance fee and only accept the delivered batch as is if the observed data had a probability less than α under all $\mathbb{P}_{N,\vartheta,n}$, where $\vartheta \in \Theta_1$, for which the batch would be rejected. In particular, you would most probably very quickly loose your credibility toward the manufacturer since you'd often claim a non-performance fee although the ratio of dysfunctional bulbs in the batch did not exceed* 1%.

We recall here that in Definition 2.1.13 we had introduced the concept of best unbiased estimators for some function τ; i.e., among all unbiased estimators for τ, we considered those the best for which the variance was minimised. In a similar fashion we now define what we consider best tests of Θ_0 against Θ_1 (where the latter two sets are assumed to be given, in the same way that the function τ to estimate was given for defining best estimators).

Definition 2.3.6. *Let a statistical model $(\mathfrak{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$, $\alpha \in (0,1)$, and a partition $\Theta_0 \dot\cup \Theta_1 = \Theta$ be given. A test T of Θ_0 against Θ_1 is called* uniformly most powerful (UMP) test at level α ('gleichmäßig bester Test zum Niveau α'), *if T is a test of Θ_0 against Θ_1 at level α, and at the same time for any other test $\widetilde{T}$ of Θ_0 against Θ_1 at level α, one has*

$$G_T(\vartheta) \geqslant G_{\widetilde{T}}(\vartheta) \quad \forall \vartheta \in \Theta_1. \tag{2.3.4}$$

[11] There are of course many tests T which satisfy (2.3.2), and our approach here is not mandatory, i.e., we could as well choose a rejection region that is not of the form $\{k, k+1, \dots, n\}$; however, choosing a rejection region of the form $\{k, k+1, \dots, n\}$ corresponds to the heuristics that 'the higher the observed number of dysfunctional bulbs in the sample, the higher the tendency to discard the null hypothesis'.

Now even though we can define the notion of a best test of Θ_0 against Θ_1 at level α, it might still not be clear how to actually define a useful test. For example, we should certainly choose a reasonable value of α, and the smaller we choose α, the smaller the power of the test gets – we have to strike a balance here!

2.4 Testing for alternatives ('Alternativtests')

A reasonably simple setting is that of observing data of which you know that it has been obtained as a realization according to one of two possible candidate distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ – it is your task to determine whether to accept the null hypothesis that the data has been realized according to $\mathbb{P}_0$, or whether to reject the null hypothesis and opt for the alternative $\mathbb{P}_1$.
We will deal with the case of a standard model, i.e., there exist corresponding likelihood functions $\varrho_0, \varrho_1 : \mathfrak{X} \to [0, \infty)$ for $\mathbb{P}_0$ and $\mathbb{P}_1$. Following the maximum-likelihood heuristics that led us to define maximum likelihood estimators in Definition 2.1.11, one may arrive at the idea to consider the ratio

$$R(x) = \begin{cases} \frac{\varrho_1(x)}{\varrho_0(x)}, & \text{if } \varrho_0(x) > 0, \\ \infty, & \text{if } \varrho_0(x) = 0. \end{cases} \tag{2.4.1}$$

of the two densities for a given observation $x \in \mathfrak{X}$. Intuitively, the larger this quotient, the more one should tend to reject the null hypothesis that the underlying distribution is $\mathbb{P}_0$.

Definition 2.4.1. *A test T of the form*[12]

$$T(x) = \begin{cases} 1, & \text{if } R(x) > c, \\ 0, & \text{if } R(x) < c, \end{cases} \tag{2.4.2}$$

for some constant $c \in (0, \infty)$ and with R as in (2.4.1)*, is called a* Neyman-Pearson test *(Polish-US statistician Jerzy Neyman (1894–1981), Egon S. Pearson (1895–1980)).*

The following result tells us that in the above context of two alternative hypotheses, Neyman-Pearson tests are in fact the best you can get.

Theorem 2.4.2 (Neyman-Pearson lemma)**.** *Assume a statistical standard model $(\mathfrak{X}, \mathcal{A}, \mathbb{P}_0, \mathbb{P}_1)$ as well as a level $\alpha \in (0, 1)$ be given ($\mathbb{P}_0$ will be the probability distribution under the null hypothesis, $\mathbb{P}_1$ the probability distribution of the alternative).*

(a) There exists a Neyman-Pearson test T such that

$$\mathbb{E}_0[T] = \alpha. \tag{2.4.3}$$

(b) Any such Neyman-Pearson test T satisfying (2.4.3) *is a best test at level α.*

Proof. (a) Let $\alpha \in (0, 1)$ be given. The only tasks we have to solve is to find the right c and then possibly randomise on the set $\{R = c\}$ in order to get the desired equality in (2.4.3).

Define

$$c := \inf\big\{s \in \mathbb{R} \,:\, \mathbb{P}_0(R \in (s, \infty)) \leqslant \alpha\big\}.$$

Thus, for any $s < c$, by definition we have

$$\mathbb{P}_0 \circ R^{-1}((s, \infty)) > \alpha.$$

Since probability measures are continuous from above (recall Proposition 1.3.9 (g)) we deduce that

$$\mathbb{P}_0 \circ R^{-1}([c, \infty)) \geqslant \alpha.$$

as well as

$$\mathbb{P}_0 \circ R^{-1}((c, \infty)) \leqslant \alpha.$$

Therefore, if $\mathbb{P}_0(R = c) = 0$, then in combination with the two previous displays we deduce

$$\mathbb{P}_0(R \geqslant c) = \alpha, \tag{2.4.4}$$

[12] The case $R(x) = c$ is not specified here in order to leave sufficient leeway to construct Neyman-Pearson tests that attain any given significance level $\alpha \in (0, 1)$, i.e., corresponding to equality in (2.3.2).

and hence we define

$$T(x) = \begin{cases} 1, & \text{if } R(x) \geqslant c, \\ 0, & \text{if } R(x) < c, \end{cases}$$

which due to (2.4.4) is a Neymann-Pearson test at level α as desired.

If, on the other hand, $\mathbb{P}_0(R = c) > 0$, then we set

$$T(x) = \begin{cases} 1, & \text{if } R(x) > c, \\ \frac{\alpha - \mathbb{P}_0(R>c)}{\mathbb{P}_0(R=c)}, & \text{if } R(x) = c, \\ 0, & \text{if } R(x) < c, \end{cases}$$

which again is a Neymann-Pearson test at level α as desired.

(b) Let T be a Neyman-Pearson test with $\mathbb{E}_0[T] = \alpha$ as in the assumptions, and let $\widetilde{T}$ be any other test at level α. We have to show that

$$G_T(1) \geqslant G_{\widetilde{T}}(1) \quad \forall \vartheta \in \Theta_1. \tag{2.4.5}$$

Denote as before by c the constant from the definition of the Neyman-Pearson test T as in (2.4.4). We have

$$\begin{aligned} G_T(1) - G_{\widetilde{T}}(1) = \mathbb{E}_1[T] - \mathbb{E}_1[\widetilde{T}] = &\int_{\mathfrak{X}} \mathbb{1}_{T(x)-\widetilde{T}(x)>0}(x)\,(T(x) - \widetilde{T}(x)) \cdot \varrho_1(x)\,\mathrm{d}x \\ &+ \int_{\mathfrak{X}} \mathbb{1}_{T(x)-\widetilde{T}(x)<0}(x)\,(T(x) - \widetilde{T}(x)) \cdot \varrho_1(x)\,\mathrm{d}x. \end{aligned}$$

Now since T is a Neyman-Pearson test with constant c, on $\{T - \widetilde{T} > 0\}$ we must have $T > 0$ and hence $\varrho_1 \geqslant c\varrho_0$. On the other hand, on $\{T - \widetilde{T} < 0\}$ we must have $T < 1$ and therefore in particular $\varrho_1 \leqslant c\varrho_0$. As a consequence, we get

$$\mathbb{E}_1[T] - \mathbb{E}_1[\widetilde{T}] \geqslant c \int_{\mathfrak{X}} (T(x) - \widetilde{T}(x)) \cdot \varrho_0(x)\,\mathrm{d}x = c(\mathbb{E}_0[T] - \mathbb{E}_0[\widetilde{T}]) \geqslant 0.$$

This shows that T is a best test at level α.

□

Exercise 2.4.3. *Convince yourself that there might be cases that satisfy the assumptions of Theorem 2.4.2, and for which we have $\mathbb{P}_0(R = c) > 0$.*

Remark 2.4.4. *The previous Theorem 2.4.2 tells us how to construct UMP tests under the assumption that we do have two one-element sets Θ_0 and Θ_1 via Neyman-Pearson tests. Under certain monotonicity assumptions on the underlying distributions, this technique extends to obtain UMP tests for more interesting sets Θ_0 and Θ_1 of parameters.*
For instance, have another look at Example 2.3.1 and recall that $\Theta_0 = \{0, 1, \dots, \lfloor \frac{N}{100} \rfloor\}$ and $\Theta_1 = \{\lfloor \frac{N}{100} \rfloor, \lfloor \frac{N}{100} \rfloor + 1, \dots, N\}$. In particular, we want to show that the test T constructed in Remark 2.3.4 is a UMP test of Θ_0 against Θ_1 at level α.
The principal observation now is that for $\vartheta_1 > \vartheta_0$ the quotients

$$R(x) := \frac{\mathrm{H}_{N,\vartheta_1,n}(x)}{\mathrm{H}_{N,\vartheta_0,n}(x)} \quad \textit{are increasing in } x \in \{0, 1, \dots, n\} \tag{2.4.6}$$

(with $R(x) := \infty$ for all $x > \vartheta_0$), as is easy to check; indeed, it follows from

$$\frac{\mathrm{H}_{N,\vartheta,n}(x)}{\mathrm{H}_{N,\vartheta+1,n}(x)} = \frac{(\vartheta + 1)(N - \vartheta - n + x)}{(\vartheta + 1 - x)(N - \vartheta)},$$

for x from the corresponding adequate range.
Display (2.4.6) implies that T is a Neyman-Pearson test of $\{\vartheta_0\}$ (with $\vartheta_0 = \lfloor \frac{N}{100} \rfloor$) against $\{\vartheta_1\}$ at level α, for any $\vartheta_1 \in \{\lfloor \frac{N}{100} \rfloor + 1, \dots, N\}$. As a consequence, Theorem 2.4.2 tells us that T is a UMP test of $\{\lfloor \frac{N}{100} \rfloor\}$ against $\{\vartheta_1\}$ at level α; hence, it satisfies (2.3.4) for $\Theta_1 = \{\vartheta_1\}$; but since $\vartheta \in \{\lfloor \frac{N}{100} \rfloor + 1, \dots, N\}$ was chosen arbitrarily, T also is a UMP test of $\{\lfloor \frac{N}{100} \rfloor\}$ against $\Theta_1 := \{\lfloor \frac{N}{100} \rfloor + 1, \dots, N\}$ at level α.

The only thing we still have to show is that even as a test of Θ_0 against Θ_1, the test T still has level α. For this purpose it is sufficient to prove that

$$\mathbb{E}_\vartheta[T] \leqslant \mathbb{E}_{\lfloor \frac{N}{100} \rfloor}[T] \tag{2.4.7}$$

for all $\vartheta \in \{0, 1, \ldots, \lfloor \frac{N}{100} \rfloor - 1\}$. But using (2.4.6) again, we get that T is a Neyman-Pearson test of $\{\vartheta\}$ against $\lfloor \frac{N}{100} \rfloor$, hence it is a UMP test of $\{\vartheta\}$ against $\lfloor \frac{N}{100} \rfloor$ at level $\beta := \mathbb{E}_\vartheta[T]$. Thus, it is at least as powerful as the constant test $\widehat{T} := \beta$, and therefore we obtain the last inequality in

$$\mathbb{E}_\vartheta[T] = \beta = \mathbb{E}_{\lfloor \frac{N}{100} \rfloor}[\widehat{T}] \leqslant \mathbb{E}_{\lfloor \frac{N}{100} \rfloor}[T],$$

which establishes (2.4.7) and hence finishes the proof.
Note that this procedure of extending the UMP property from Neyman-Pearson tests to more general null hypotheses and alternatives is not limited to this example. Once you have a property of the type (2.4.6), you're in business to extend Θ_1 to sets of more than one element. Furthermore, you would also need to show that $\sup_{\vartheta \in \Theta_0} \mathbb{E}_\vartheta[T] = \mathbb{E}_{\vartheta_0}[T]$ in order to extend $\{\vartheta_0\}$ to Θ_0.

2.4.1 Chi-square test of goodness of fit ('Chiquadrat-Anpassungstest')

In Example 2.3.2 we have seen an easy way to construct a test of the null hypothesis of a fair coin (against the alternative of an unfair coin) at an arbitrary level α. While this was a pretty simple case, in more general state spaces it is usually harder to check whether or not some given data has been realized according to a certain distribution (which will be the null hypothesis) or not (alternative). For the time being we will focus on (discrete) distributions in this setting.
As in the setting of the coin flips we will assume that the data observed has been produced as a realization of i.i.d. experiments, so we will work in the infinite product model, where each coordinate takes a value in a finite set E. I.e., the model is given by

$$(E^{\mathbb{N}}, (2^E)^{\otimes \mathbb{N}}, (\mathbb{P}_\vartheta^{\otimes \mathbb{N}})_{\vartheta \in \Theta}),$$

where $\mathbb{P}_\vartheta$ is the distribution of any of the coordinates.
For notational convenience, we denote by Θ the space of probability measures on $(E, 2^E)$. We are going to test the null hypothesis that the coordinates have a certain distribution given by $\varrho \in \Theta$ against the alternative $\Theta_1 := \Theta \backslash \{\varrho\}$. (In the context of a presumably fair coin we would have taken $E = \{0, 1\}$ and $\varrho(0) = \varrho(1) = \frac{1}{2}$).
To the first n coordinates (i.e., observations) we can associate a probability measure as follows. For $x \in E$ denote

$$\ell_n(x) := |\{1 \leqslant k \leqslant n \,:\, X_k = x\}|$$

the frequency of x during the first n observations. We then define the probability measures

$$L_n(A) := \frac{1}{n} \sum_{x \in A} \ell_n(x), \quad A \in 2^E,$$

on $(E, 2^E)$. It is often referred to as the *empirical distribution*.
As done several times before already, again we are guided by the maximum likelihood heuristics and define for any probability measure $\vartheta \in \Theta$ the quotient

$$R_n(\vartheta) := \frac{\prod_{x \in E} \vartheta(x)^{\ell_n(x)}}{\prod_{x \in E} \varrho(x)^{\ell_n(x)}}.$$

Taking logarithms, we obtain with $\vartheta = L_n$ that

$$\ln R_n(L_n) = n \sum_{x \in E} L_n(x) \ln \frac{L_n(x)}{\varrho(x)} = nH(L_n \,|\, \varrho),$$

where H is the relative entropy that we have introduced in Definition 2.1.26. Since we have seen that the relative entropy is always non-negative as well as that $H(Q \,|\, Q) = 0$ (cf. Proposition 2.1.27), it suggests itself to introduce tests of the form[13]

$$T_n(L_n) = \begin{cases} 1, & \text{if } nH(L_n \,|\, \varrho) > c, \\ 0, & \text{if } nH(L_n \,|\, \varrho) \leqslant c. \end{cases}$$

[13] While it suggests itself to discard the null hypothesis if the quantity $H(L_n \,|\, \varrho)$ takes large values, there is some seemingly arbitrary element in choosing $nH(L_n \,|\, \varrho)$ as the quantity to test against c. This can, however be motivated e.g. by the use of so-called 'large deviation principles', which show that in quite general situations the probability of observing L_n roughly decays like $e^{-nH(L_n \,|\, \varrho)(1+o(1))}$.

(Also recall that $H(L_n \,|\, Q)$ was called the Kullback-Leibler distance, so in some sense we will discard the null hypothesis if the distance of L_n to Q is 'large'.) Denote

$$D_{n,\varrho} := n \sum_{x \in E} \varrho(x) \Big(\frac{L_n(x)}{\varrho(x)} - 1 \Big)^2. \tag{2.4.8}$$

Definition 2.4.5. *Any test of the null hypothesis $H_0 : \vartheta = \varrho$ against the alternative $H_1 : \vartheta \neq \varrho$ with rejection region of the form $\{D_{n,\varrho} > c\}$, some $c \in (0, \infty)$, is called a* χ^2-test of goodness of fit ('χ^2-Anpassungstest') *after n observations.*

The reason for this terminology will be given in Theorem 2.4.8 below.

Theorem 2.4.6.

$$nH(L_n \,|\, \varrho) - D_{n,\varrho}/2$$

converges to 0 *in distribution under $\mathbb{P}_\vartheta$ (and hence in $\mathbb{P}_\vartheta$-probability as well, since the limit is a constant).*

Proof. The proof can be found as the proof of Proposition 11.10 in [Geo09]. □

Definition 2.4.7. *Let $X_1, \ldots, X_n$ be i.i.d. independent $\mathcal{N}(0,1)$ distributed random variables. The distribution of the random variable $\sum_{i=1}^n X_i^2$ is called the* χ^2-distribution with n degrees of freedom ('χ^2-Verteilung mit n Freiheitsgraden').

The following result is due to Karl Pearson (1857–1936), the father of Egon S. Pearson whom we encountered in the context of the Neyman-Pearson tests.

Theorem 2.4.8. *Under $\mathbb{P}_\vartheta$, the sequence $(D_{n,\varrho})_{n \in \mathbb{N}}$ converges to a $\chi^2_{|E|-1}$ distributed random variable.*

Proof. See proof of Theorem 11.12 in [Geo09]. □

Note by the author. It is my sincere hope that you did learn some mathematics during this class. I would also be happy if you enjoyed it as much as I did; and if you did not, and feel that there is anything else I could do to improve your learning experience in this class (except for modifying the exercise class system, which we will be doing anyways), then please do let me know (drewitz@math.uni-koeln.de)!
Even though I might not always have treated you with kid gloves, I hope this ultimately helps growing your independence and resilience. And if you are willing to take one last bit of advice, heed the following: `https://www.youtube.com/watch?v=D1R-jKKp3NA` (see `https://www.youtube.com/watch?v=DpMwWaxoI4Y` for a German version).
Acknowledgment: I would like to thank Alexis Prévost and Lars Schmitz for pointing out various mistakes and suggestions which led to an improved version of these notes.

Bibliography

[Ber13] J. Bernoulli. *Ars conjectandi*. Landmarks of science. Impensis Thurnisiorum, fratrum, 1713.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and {ISDN} Systems*, 30(1?7):107 – 117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[Das88] Lorraine Daston. *Classical probability in the Enlightenment*. Princeton University Press, Princeton, NJ, 1988.

[Dur10] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[Ete81] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrsch. Verw. Gebiete*, 55(1):119–122, 1981.

[Geo03] Hans-Otto Georgii. Probabilistic aspects of entropy. In *Entropy*, Princeton Ser. Appl. Math., pages 37–54. Princeton Univ. Press, Princeton, NJ, 2003.

[Geo09] Hans-Otto Georgii. *Stochastik*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, expanded edition, 2009. Einführung in die Wahrscheinlichkeitstheorie und Statistik. [Introduction to probability and statistics].

[GK99] G. Gigerenzer and C. Krüger. *Das Reich des Zufalls: Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen*. Spektrum, Akad. Verlag, 1999.

[Hal90] Anders Hald. *A history of probability and statistics and their applications before 1750*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1990. A Wiley-Interscience Publication.

[Kle14] Achim Klenke. *Probability theory*. Universitext. Springer, London, second edition, 2014. A comprehensive course.

[Kol33] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

[Kre05] U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik: Für Studium, Berufspraxis und Lehramt*. vieweg studium; Aufbaukurs Mathematik. Vieweg+Teubner Verlag, 2005.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[Ren69] A. Renyi. *Briefe über die Wahrscheinlichkeit*. Wissenschaft und Kultur. Birkhäuser Basel, 1969.

[Sch88] I. Schneider. *Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis 1933: Einf. u. Texte*. Akad.-Verlag, 1988.

[Sen06] E. Seneta. *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York, 2006. Revised reprint of the second (1981) edition [Springer-Verlag, New York; MR0719544].

[vdWB75] B.L. van der Waerden and J. Bernoulli. *Die Werke von Jakob Bernoulli: Bd. 3: Wahrscheinlichkeitsrechnung*. Die Werke von Jakob Bernoulli. Birkhäuser Basel, 1975.