MATH 2431

Honors Probability

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Maximilian Nitzschner
May 6, 2025

Disclaimer:

These are lecture notes for the postgraduate course *Honors Probability (MATH 2431)*, given at the Hong Kong University of Science and Technology in Spring 2025. They are intended for students at the Hong Kong University of Science and Technology and should *not* be made widely available without the author's consent.

The exposition does not strictly follow any particular textbook. The main references are [Geo12] and [GS01]. Other useful references include [Bré88, DS10, Dur09].

These notes are preliminary and may contain typos. If you see any mistakes or think that the presentation is unclear and could be improved, please send an email to: mnitzschner@ust.hk. All comments and suggestions are appreciated.

Contents

0. Motivation **5**

1. Outcomes, events, probability measures **6**

1.1. Sample spaces . . . 6

1.2. Elementary Combinatorics . . . 7

1.3. Events, σ-algebras . . . 11

1.4. Probability measures . . . 14

2. Conditional probability and stochastic independence **17**

2.1. Conditional probability . . . 17

2.2. The law of total probability and Bayes' theorem . . . 18

2.3. Stochastic independence . . . 20

3. Discrete distributions **23**

4. Continuous distributions **29**

4.1. Probability density functions . . . 29

4.2. Generated σ-algebras and the Borel-σ-algebra . . . 31

4.3. Properties of continuous distributions . . . 32

4.4. Distribution functions, Lebesgue-Stieltjes measures . . . 35

5. Random variables **44**

5.1. Definition of random variables . . . 44

5.2. Law and cumulative distribution of a real random variable . . . 45

5.3. Transformation of random variables . . . 46

6. Expectation, variance and higher moments of random variables **49**

6.1. Expectation . . . 49

6.2. Variance . . . 54

7. Joint distributions of random variables, covariance, extremes **59**

7.1. Joint distributions of random variables . . . 59

7.2. Independence of random variables . . . 67

7.3. Extremes . . . 71

7.4. Sums of independent random variables . . . 72

7.5. Covariance and correlation . . . 75

8. Jensen and Hölder inequalities **79**
8.1. Jensen's inequality 79
8.2. Hölder's inequality 80

9. Conditional distributions and conditional expectation **82**
9.1. Discrete conditional distributions 82
9.2. Continuous conditional distributions 84
9.3. Conditional expectation 84

10. Generating functions **89**

11. Convergence in probability, almost sure convergence and the law of large numbers **93**
11.1. Convergence in probability and the weak law of large numbers 93
11.2. Almost sure convergence and the strong law of large numbers 94
11.3. Application: Monte-Carlo integration 98

12. Weak convergence and the central limit theorem **100**
12.1. Convergence in distribution 100
12.2. The central limit theorem 101
12.3. More properties of convergence in distribution 104

13. An overview of selected random processes **107**
13.1. The Poisson process 107
13.2. Markov chains 110

A. Appendix **115**
A.1. Riemann integration 115
A.2. Multiple integrals 118
A.3. Alternative Proof of the central limit theorem 12.4 127

0. Motivation

The purpose of *probability theory* is to study and be able to make predictions about systems that involve *randomness.*

Consider as a simple example rolling a (fair) die multiple times, which is a process with a random outcome. A first objective will be to develop the mathematical description of characteristics of such a random experiment: This is the specification of a *stochastic model.* Loosely speaking:

(0.1) *Probability theory is concerned with the description of random phenomena using stochastic models.*

Here are some examples of phenomena calling for a probabilistic description:

- throwing a (fair) die or coin multiple times;
- describing the random movement of a particle in $\mathbb{Z}^d$, $d \geq 1$ (random walk): at every time step, the particle moves randomly to one of its neighboring sites, with “equal probability”;

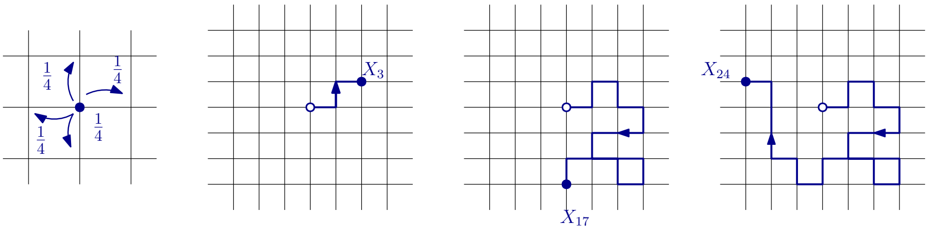

Figure 1.: Leftmost panel: Possible jumps for the walker in $\mathbb{Z}^2$ at the origin; next panels: Positions of the walker after 3, 17, or 24 steps, respectively.

Some typical simple questions we could ask in this context are for instance:

- What is the average of the numbers seen for the die after a large number of throws?
- Where will the random particle be after a large number of steps?
- What is the approximate probability that the sum of the numbers coming up when throwing the die 1000 times exceeds 5000?

In this course we develop rigorous techniques to answer some of these questions. Notably, we will see the *law of large numbers* and the *central limit theorem*, which address the three questions above.

1. Outcomes, events, probability measures

(Reference: [GS01, Sections 1.1-1.3], or [Geo12, Sections 1.1-1.2, 2.1-2.3])

Our primary objective is to construct a mathematical model for a *random experiment.* Conceptually, this involves the specification of three quantities:

- a *set of outcomes* or *sample space* $\Omega \neq \varnothing$; an element $\omega \in \Omega$ should be interpreted as a possible realization / measurement of the random experiment.
- a *class of events* $\mathscr{F} \subseteq \mathscr{P}(\Omega)$, called σ-algebra; an event $A \in \mathscr{F}$ is a subset of Ω, and we aim at specifying its probability.
- a *probability measure* $\mathbf{P}$, which is a map from $\mathscr{F}$ to $[0,1]$ that assigns a probability $\mathbf{P}[A]$ to any given event $A \in \mathscr{F}$.

The triple $(\Omega, \mathscr{F}, \mathbf{P})$ is called a *probability space.* In the following sections, we give precise definitions of these objects and present examples.

1.1. Sample spaces

Definition 1.1. A non-empty set Ω consisting of the possible realizations of a random experiment is called *set of outcomes* or *sample space.* An element $\omega \in \Omega$ is called an *outcome.*

Example 1.2. (i) Tossing a coin: The possible outcomes are heads and tails, which we denote by H and T respectively. In this case, we have

$$\Omega_1 = \{H, T\}. \tag{1.1}$$

(ii) Rolling a die: The outcomes are the integer numbers from 1 to 6, so

$$\Omega_2 = \{1,2,3,4,5,6\}. \tag{1.2}$$

(iii) Tossing a coin *and* rolling a die: We define the sample space as the *Cartesian product* of Ω_1 and Ω_2, namely

$$\begin{aligned} \Omega_3 &= \Omega_1 \times \Omega_2 \\ &= \{(\omega_1, \omega_2)\,;\, \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\} \\ &= \{(H,1),(H,2),\dots,(H,6),(T,1),(T,2),\dots,(T,6)\}. \end{aligned} \tag{1.3}$$

(iv) n-fold coin toss (where $n \in \mathbb{N} = \{1, 2, ...\}$): Here, we need to record the outcome as an n-tuple

$$\begin{aligned}\Omega_4 &= \{\underbrace{(H, H, ..., H, H)}_{n \text{ elements}}, (H, H, ..., H, T), (H, H, ..., T, H), ..., (T, T, ..., T, T)\} \\ &= \{(\omega_1, ..., \omega_n)\,;\, \omega_i \in \{H, T\} \text{ for } 1 \le i \le n\} \qquad (1.4) \\ &= \underbrace{\Omega_1 \times ... \times \Omega_1}_{n \text{ times}} = \Omega_1^n.\end{aligned}$$

(v) Tossing a coin infinitely many times: The natural choice for outcomes will be similar as in the previous example, but with sequences of infinite length rather than n-tuples. More precisely

$$\Omega_5 = \Omega_1^{\mathbb{N}} = \{(\omega_1, \omega_2, ...)\,;\, \omega_i \in \{H, T\} \text{ for } i \in \mathbb{N}\}. \qquad (1.5)$$

(vi) The number of customers in a shop during a given day:

$$\Omega_6 = \mathbb{N}_0 = \{0, 1, 2, ...\}. \qquad (1.6)$$

(vii) The lifetime of a light bulb:

$$\Omega_7 = \mathbb{R}_0^+ = [0, \infty). \qquad (1.7)$$

Let us point out that the sample spaces $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ and Ω_6 are countable[1], whereas Ω_5 and Ω_7 are uncountable.

1.2. Elementary Combinatorics

In many elementary cases, the assumption that *all (finitely many) outcomes are equally likely* is justified (think of rolling a die or flipping a coin multiple times). In this situation, the probability space will be uniquely characterized by the number $|\Omega| \in \mathbb{N}$. We want to develop effective methods to count the number of outcomes of Ω and events $A \subseteq \Omega$.

Remark 1.3. If $N \in \mathbb{N}$ random experiments with finite sample spaces $\Omega_1, \Omega_2, ..., \Omega_N$ are performed successively, an appropriate choice for the sample space of the combined experiment is given by the *Cartesian product*

$$\begin{aligned}\Omega = \prod_{j=1}^{N} \Omega_j &:= \Omega_1 \times \Omega_2 \times ... \times \Omega_N \qquad (1.8) \\ &= \{(\omega_1, ..., \omega_n)\,;\, \omega_j \in \Omega_j, 1 \le j \le N\}.\end{aligned}$$

[1] A set S is countable if it is empty or if there exists a surjective (onto) map $\rho : \mathbb{N} \to S$. This includes the case of finite S.

The cardinality of Ω is given by

$$|\Omega| = \prod_{j=1}^{N} |\Omega_j| = |\Omega_1| \cdot |\Omega_2| \cdot \ldots \cdot |\Omega_N|. \tag{1.9}$$

We already saw this in Example 1.2, (iii).

Example 1.4. Imagine a password contains four symbols, where the first two are (uppercase) roman letters, and the third and fourth are each a single digit (e.g. "$HP25$"). How many passwords can be formed with this set-up? Here we have:

$$\Omega_1 = \Omega_2 = \{A, B, \ldots, Z\},$$
$$\Omega_3 = \Omega_4 = \{0, 1, \ldots, 9\}.$$

A password is an element of $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3 \times \Omega_4$. Therefore:

$$|\Omega| = \prod_{j=1}^{4} |\Omega_j| = 26 \cdot 26 \cdot 10 \cdot 10 = 67600.$$

Proposition 1.5. *The number of choices of a sample of size $r \in \mathbb{N}$ out of $\{1, 2, \ldots, n\}$ is given as follows:*

	with repetitions	*without repetitions*
ordered	n^r	$\frac{n!}{(n-r)!}$
unordered	$\binom{n+r-1}{r}$	$\binom{n}{r}$

For the case without repetitions, we additionally require $r \leq n$. In the table above we used the notations $k! = k \cdot (k-1) \cdot \ldots \cdot 1$ for $k \in \mathbb{N}$ (and $0! = 1$), as well as $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ for $0 \leq k \leq n$.

Proof.

- ▶ Ordered samples, with repetitions: This is a special case of Remark 1.3. More precisely, we use

$$\Omega_1 = \{(\omega_1, \ldots, \omega_r)\,;\, \omega_j \in \{1, 2, \ldots, n\}\} = \{1, 2, \ldots, n\}^r, \tag{1.10}$$

with $|\Omega_1| = n^r$.

- ▶ Ordered samples, without repetitions: Here we use

$$\Omega_2 = \{(\omega_1, \ldots, \omega_r)\,;\, \omega_j \in \{1, 2, \ldots, n\}, \omega_i \neq \omega_j \text{ for } i \neq j\}, \tag{1.11}$$

with $|\Omega_2| = n \cdot (n-1) \cdot \ldots \cdot (n-r+1)$.

- Unordered samples, without repetitions: Here we use

  $$\Omega_3 = \{\{\omega_1, ..., \omega_r\}\,;\, \omega_j \in \{1, 2, ..., n\}, \omega_i \neq \omega_j \text{ for } i \neq j\}. \tag{1.12}$$

  Here $r!|\Omega_3| = |\Omega_2| = \frac{n!}{(n-r)!}$ holds: This is because for $r \in \{1, ..., n\}$ different elements $\omega_1, ..., \omega_r$, there are exactly r possibilities of reordering.

- Unordered samples, with repetitions: The sample space can be written as

  $$\Omega_4 = \{(\omega_1, ..., \omega_r)\,;\, \omega_j \in \{1, 2, ..., n\}, 1 \leq \omega_1 \leq \omega_2 \leq ... \leq \omega_r \leq n\}. \tag{1.13}$$

  We visualize an element of Ω_4 as follows: We separate the n numbers $1, ..., n$ by $n-1$ lines (|), and for each instance of one of these numbers within the sequence $(\omega_1, ..., \omega_r)$, we put a dot (•) in the respective bin.

  Example: Let $n = 6$, $r = 5$. The element $(1, 1, 3, 4, 6) \in \Omega_4$ corresponds to the string

  • • || • | • ||•,

  and the element $(2, 2, 2, 5, 5) \in \Omega_4$ corresponds to the string

  | • • • ||| • •|.

  The number of different strings corresponds therefore to the numbers of choices of a set of r elements (the dots) out of a set with $n + r - 1$ elements (the strings consisting of dots and lines), which is exactly $\binom{n+r-1}{r}$ by the previous step.

□

Example 1.6. A committee of 12 persons consists of 3 representatives of group A, 4 of group B and 5 of group C. We want to choose a subcommittee of 5 persons, with

- one member of group A,
- two members of group B,
- two members of group C?

Let S denote a set enumerating the different possible choices for the subcommittee. Note that we do not specify the order within the groups and obviously, there are no repetitions in the choice of the members. Thus we have

- $\binom{3}{1}$ choices for the member from group A,
- $\binom{4}{2}$ choices for the member from group B,
- $\binom{5}{2}$ choices for the member from group C,

and thus

$$|E| = \binom{3}{1} \cdot \binom{4}{2} \cdot \binom{5}{2} = 180. \tag{1.14}$$

In Proposition 1.5, we essentially considered all possible ways of choosing samples of size r out of a set with n elements. Suppose now that $n \in \mathbb{N}$ items are to be divided into k distinct groups of size $n_k \in \mathbb{N}$ (so that $n = n_1 + \ldots + n_k$). How many such choices are possible?

- There are $\binom{n}{n_1}$ choices for the first group,
- there are $\binom{n-n_1}{n_2}$ choices for the second group,
- there are $\binom{n-n_1-n_2}{n_3}$ choices for the third group, ...

and thus in total, there are

$$\begin{aligned} & \binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \binom{n-n_1-n_2}{n_3} \cdot \ldots \\ &= \frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdot \frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \cdot \ldots \\ &= \frac{n!}{\prod_{j=1}^{k} n_j!}. \end{aligned} \tag{1.15}$$

Example 1.7. Suppose a company has 10 employees. How many ways are there to assign tasks, if 5 employees are needed for task "A", 3 are needed for task "B" and 2 are needed for task "C"? In this set-up, we have $n = 10$ (the employees) and $n_1 = 5$, $n_2 = 3$, $n_3 = 2$, so:

$$\frac{10!}{5! \cdot 3! \cdot 2!} = 2520$$

possibilities. Incidentally, this is of course the same number as the number of reorderings of the "word" $AAAAABBBCC$.

The expression $\frac{n!}{n_1! \ldots n_k!}$ is called *multinomial coefficient*, and it generalizes the binomial coefficient $\binom{n}{n_1} = \frac{n!}{n_1! n_2!}$. Sometimes, the following abbreviation is used:

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{\prod_{j=1}^{k} n_j!}, \qquad n_1, \ldots, n_k \in \mathbb{N}_0, \sum_{j=1}^{k} n_j = n \in \mathbb{N}_0. \tag{1.16}$$

Remark 1.8. Using the multinomial coefficients, one can show the *multinomial theorem*: For $x_1, \ldots, x_k \in \mathbb{R}$ and $n \in \mathbb{N}$, one has

$$\left(\sum_{j=1}^{k} x_j \right)^n = \sum_{\substack{(n_1,\ldots,n_k) \in \mathbb{N}_0^k \\ n_1+\ldots+n_k=n}} \binom{n}{n_1, n_2, \ldots, n_k} \prod_{j=1}^{k} x_j^{n_j}. \tag{1.17}$$

This is a generalization of the well-known *binomial theorem*: For $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$, one has

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}. \tag{1.18}$$

End of Lecture 1

1.3. Events, σ-algebras

Suppose that we have fixed a sample space Ω. In general we are interested in the occurrence of *events* that consist of a certain selection of outcomes. For instance consider rolling a die once (recall from Example 1.2, (ii) that

$$\Omega_2 = \{1,2,3,4,5,6\}$$

is a reasonable choice for the sample space for this random experiment). The *event*

$$A = \text{“the upper face of the die shows an even number”} \tag{1.19}$$

can then be expressed as

$$A = \{2,4,6\} \subseteq \Omega_2. \tag{1.20}$$

$\rightsquigarrow$ **Naive definition:** An *event* is a subset $A \subseteq \Omega$ of the sample space.

This works in the case where Ω is countable (in particular, if Ω is finite), but leads to an important complication when Ω is uncountable (see Example 1.2, (v) and (vii)). It turns out that if we allow every subset $A \subseteq \Omega$ for an uncountable Ω, we cannot define a probability for A without running into problems. Fortunately, we can restrict out attention to smaller classes of subsets.

Definition 1.9. Let $\Omega \neq \varnothing$. The *power set* $\mathscr{P}(\Omega)$ is the set of all subsets of Ω, i.e.

$$\mathscr{P}(\Omega) = \{A\,;\ A \subseteq \Omega\}. \tag{1.21}$$

A *σ-algebra* on Ω is a subset $\mathscr{F} \subseteq \mathscr{P}(\Omega)$ that fulfills the following properties:

(S1) $\Omega \in \mathscr{F}$.

(S2) If $A \in \mathscr{F}$, then $A^c = \Omega \setminus A \in \mathscr{F}$.

(S3) If for every $j \in \mathbb{N}$, $A_j \in \mathscr{F}$, then $\bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup A_3 \cup \ldots \in \mathscr{F}$.

A set $A \in \mathscr{F}$ is called an *event.* If $\omega \in A$, we say that the event A *occurs* (for the outcome ω). If $\omega \notin A$, we say that A *does not occur* (for the outcome ω).

Remark 1.10. (i) The power set $\mathscr{P}(\Omega)$ itself is a σ-algebra (and we will usually use it if Ω is countable, in particular if it is finite).

(ii) The event Ω always occurs in a random experiment, since $\omega \in \Omega$ is always true. On the other hand, the event $\varnothing = \Omega^c$ never occurs, since $\omega \in \varnothing$ can never be true.

(iii) In the previous definition, (S2) should be understood as follows: If $A \in \mathscr{F}$ is an event, then A^c, which has the interpretation that A does not occur, should also be an event. Similarly (S3) means: If $A_1, A_2, A_3, \ldots$ are events, then $\bigcup_{j=1}^{\infty} A_j$, which has the interpretation that one of the A_j occurs, should also be an event.

(iv) Let $\Omega = \{1, 2, 3\}$ and consider the following subsets of $\mathscr{P}(\Omega)$:

$$\mathscr{F}_1 = \Big\{\varnothing, \{1\}, \{2, 3\}, \{1, 2, 3\}\Big\}$$

is a σ-algebra on Ω. In this case the set $\{2\}$ would not be an event (imagine an observe that cannot distinguish between the outcomes 2 and 3).

$$\mathscr{F}_2 = \Big\{\varnothing, \{1\}, \{2, 3\}\Big\}$$

is **not** a σ-algebra on Ω (it violates (S1)).

$$\mathscr{F}_3 = \Big\{\varnothing, \{1\}, \{1, 2, 3\}\Big\}$$

is **not** a σ-algebra on Ω (it fulfills (S1) and (S3), but violates (S2)).

$$\mathscr{F}_4 = \Big\{\varnothing, \{1\}, \{2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\Big\}$$

is **not** a σ-algebra on Ω (it fulfills (S1) and (S2), but violates (S3)).

We draw some simple conclusions from Definition 1.9.

Proposition 1.11. *Let $\Omega \neq \varnothing$ and $\mathscr{F} \subseteq \mathscr{P}(\Omega)$ a σ-algebra.*

(i) $\varnothing \in \mathscr{F}$.

(ii) If for every $j \in \mathbb{N}$, $A_j \in \mathscr{F}$, then $\bigcap_{j=1}^{\infty} A_j \in \mathscr{F}$.

(iii) If $A, B \in \mathscr{F}$, then $A \cup B \in \mathscr{F}$, $A \cap B \in \mathscr{F}$ and $A \setminus B \in \mathscr{F}$.

Proof. We first prove (i): Since $\Omega \in \mathscr{F}$ by (S1) and $\varnothing = \Omega^c = \Omega \setminus \Omega$, we have that $\varnothing \in \mathscr{F}$ by (S2).

We turn to (ii): By de Morgan's rules[2] we have that

$$\left(\bigcap_{j=1}^{\infty} A_j\right)^c = \bigcup_{j=1}^{\infty} \underbrace{A_j^c}_{\in\mathscr{F},\text{ by (S2)}} \in \mathscr{F}, \text{ by (S3)}. \tag{1.22}$$

Therefore, we have again by (S2) that

$$\bigcap_{j=1}^{\infty} A_j = \left(\left(\bigcap_{j=1}^{\infty} A_j\right)^c\right)^c \in \mathscr{F}. \tag{1.23}$$

[2] The *de Morgan rules* state that for any collection $\{U_i\}_{i\in I}$ of subsets $U_i \subseteq U$, one has

$$\left(\bigcup_{i\in I} U_i\right)^c = \bigcap_{i\in I} U_i^c, \qquad \left(\bigcap_{i\in I} U_i\right)^c = \bigcup_{i\in I} U_i^c$$

.

We now prove (iii): Set $A_1 = A = \widetilde{A}_1$, $A_2 = B = \widetilde{A}_2$ and $A_j = \varnothing$, $\widetilde{A}_j = \Omega$ for $j \geq 3$ (which are all in $\mathscr{F}$, using the assumption, (i) and (S1)). We then see that

$$A \cup B = A \cup B \cup \varnothing \cup \varnothing \cup \ldots = \bigcup_{j=1}^{\infty} A_j \in \mathscr{F}, \tag{1.24}$$

$$A \cap B = A \cap B \cap \Omega \cap \Omega \cap \ldots = \bigcup_{j=1}^{\infty} \widetilde{A}_j \in \mathscr{F}, \tag{1.25}$$

where we used (S2) and (ii), respectively. Finally, we have that

$$A \setminus B = A \cap \underbrace{B^c}_{\in \mathscr{F}\text{, by (S2)}} \in \mathscr{F}. \tag{1.26}$$

□

We illustrate the set operations using again the example of rolling a single die.

Example 1.12. We use $(\Omega, \mathscr{F}) = (\{1,2,3,4,5,6\}, \mathscr{P}(\{1,2,3,4,5,6\}))$ and consider the events

$$\begin{aligned} A &= \text{"the upper face of the die shows an even number"} = \{2,4,6\}, \\ B &= \text{"the upper face of the die shows a prime number"} = \{2,3,5\}, \\ C &= \text{"the upper face of the die shows an odd number"} = \{1,3,5\}. \end{aligned}$$

From this, we obtain

$$\begin{aligned} B^c &= \{1,4,6\}, \qquad & A \cup B &= \{2,3,4,5,6\}, \\ A \cap B &= \{2\}, & A \cap C &= \varnothing. \end{aligned}$$

We see that the set B^c describes the event that B does not occur, the set $A \cup B$ describes

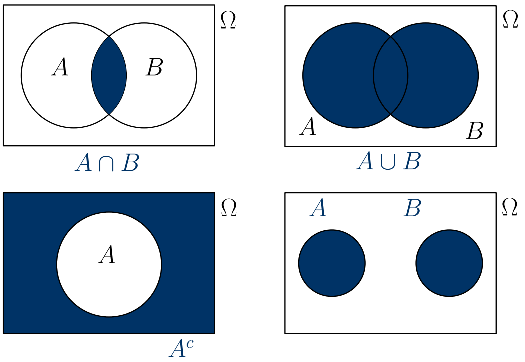

Figure 1.1.: Graphical representation of intersection, union and complement of sets (first three panels) and an example of two disjoint sets.

the event that A *or* B occurs[3] and $A \cap B$ describes the event that A *and* B occur (both). The fact that $A \cap C$ is the empty set corresponds to the fact that the events A and C are mutually exclusive.

1.4. Probability measures

Definition 1.13. Let $\Omega \neq \varnothing$ and $\mathscr{F} \subseteq \mathscr{P}(\Omega)$ a σ-algebra on Ω. A function $\mathbf{P} : \mathscr{F} \to [0,1]$ is called a *probability measure* (or simply a *probability*) if the following properties are fulfilled:

(P1) $\mathbf{P}[\Omega] = 1$ (*normalization*).

(P2) If $(A_j)_{j\in\mathbb{N}}$ is a sequence of events $A_j \in \mathscr{F}$ that are *pairwise disjoint*, namely $A_j \cap A_k = \varnothing$ for every $j, k \in \mathbb{N}$ with $j \neq k$, then

$$\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \mathbf{P}[A_j] \qquad (\sigma\text{-}\textit{additivity}). \tag{1.27}$$

The triple $(\Omega, \mathscr{F}, \mathbf{P})$ is called a *probability space.*

Example 1.14. A very natural class of examples is given by considering

$$\varnothing \neq \Omega \text{ finite}, \qquad \mathscr{F} = \mathscr{P}(\Omega), \tag{1.28}$$

and choosing the probability measure as follows:

$$\mathbf{P} : \mathscr{P}(\Omega) \to [0,1], \qquad \mathbf{P}[A] = \frac{|A|}{|\Omega|}, \tag{1.29}$$

where $|\cdot|$ denotes the cardinality (i.e. the number of elements) of a set. The probability measure $\mathbf{P}$ is the (discrete) *uniform distribution* on Ω. The resulting probability space $(\Omega, \mathscr{P}(\Omega), \mathbf{P})$ is sometimes called a *Laplace probability space.* It is characterized by the fact that

$$\mathbf{P}[\{\omega\}] = \frac{1}{|\Omega|}, \qquad \text{for every } \omega \in \Omega, \tag{1.30}$$

meaning that every outcome has the same probability.

Concrete example: We roll a die twice and are interested in the probability that the number 6 shows up at least once. Assuming that the die is fair, we set consider the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ given by

$$\begin{aligned} \Omega &= \{1,2,3,4,5,6\}^2 = \{(1,1),(1,2),...,(1,6),(2,1),...,(2,6),...,(6,6)\}, \\ \mathscr{F} &= \mathscr{P}(\Omega), \\ \mathbf{P}[A] &= \frac{|A|}{|\Omega|} = \frac{|A|}{36}, \qquad \text{for all } A \in \mathscr{P}(\Omega), \end{aligned} \tag{1.31}$$

[3] As always in mathematics, the word “or” has a non-exclusive meaning: it includes the case where A and B occur both.

and the event in question is given by

$$\begin{aligned} B = {} & \text{“At least one 6 shows up”} \\ & = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6),(6,5),(6,4),(6,3),(6,2),(6,1)\}. \end{aligned} \tag{1.32}$$

We clearly have that

$$\mathbf{P}[\text{“At least one 6 shows up”}] = \mathbf{P}[B] = \frac{11}{36}. \tag{1.33}$$

Let us now give some elementary but important properties of probabilities.

Proposition 1.15. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $A, B, A_j \in \mathcal{F}$ for $j \in \mathbb{N}$. Then the following properties hold:*

(i) $\mathbf{P}[\varnothing] = 0$,

(ii) $\mathbf{P}[A^c] = 1 - \mathbf{P}[A]$,

(iii) If $A \subseteq B$, *then* $\mathbf{P}[A] \leq \mathbf{P}[B]$,

(iv) $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]$,

(v) $\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j\right] \leq \sum_{j=1}^{\infty} \mathbf{P}[A_j]$,

(vi) If $A_1 \subseteq A_2 \subseteq \dots$ *(we say that* $(A_j)_{j\in\mathbb{N}}$ *is* increasing*), then* $\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \lim_{n\to\infty} \mathbf{P}[A_n]$,

(vii) If $A_1 \supseteq A_2 \supseteq \dots$ *(we say that* $(A_j)_{j\in\mathbb{N}}$ *is* decreasing*), then* $\mathbf{P}\left[\bigcap_{j=1}^{\infty} A_j\right] = \lim_{n\to\infty} \mathbf{P}[A_n]$.

Proof. We start with the proof of (i): Since $\varnothing = \varnothing \cup \varnothing \cup \varnothing \cup \dots$ (and clearly $\varnothing \cap \varnothing = \varnothing$), we see that

$$\mathbf{P}[\varnothing] \overset{\text{(P2)}}{=} \sum_{j=1}^{\infty} \mathbf{P}[\varnothing], \tag{1.34}$$

which can only be true if $\mathbf{P}[\varnothing] = 0$.

For (ii) note that A and A^c are disjoint and fulfill $A \cup A^c = \Omega$. We set $A_1 = A$, $A_2 = A^c$ and $A_j = \varnothing$ for $j \geq 3$, so that

$$\begin{aligned} 1 &\overset{\text{(P1)}}{=} \mathbf{P}[\Omega] = \mathbf{P}[A \cup A^c \cup \varnothing \cup \varnothing \cup \dots] \\ &\overset{\text{(P2)}}{=} \mathbf{P}[A] + \mathbf{P}[A^c] + \underbrace{\mathbf{P}[\varnothing] + \mathbf{P}[\varnothing] + \dots}_{=0 \text{ by (i)}} \\ &= \mathbf{P}[A] + \mathbf{P}[A^c]. \end{aligned} \tag{1.35}$$

For the proof of (iii), consider $\widetilde{B} = B \setminus A (= \{\omega \in B\,;\, \omega \notin A\})$, so that $A \cup \widetilde{B} = A \cup B = B$ and $A \cap \widetilde{B} = \varnothing$. We find by the same argument as for (ii):

$$\mathbf{P}[B] = \mathbf{P}[A \cup \widetilde{B} \cup \varnothing \cup \varnothing \cup \ldots] = \mathbf{P}[A] + \underbrace{\mathbf{P}[\widetilde{B}]}_{\geq 0} \geq \mathbf{P}[A]. \tag{1.36}$$

Note that this calculation shows the stronger statement

$$A \subseteq B \qquad \Rightarrow \qquad \mathbf{P}[B \setminus A] = \mathbf{P}[B] - \mathbf{P}[A]. \tag{1.37}$$

For (iv), we define $D = B \setminus (A \cap B)$ and note that $A \cap B \subseteq B$, $A \cup B \overset{(\star)}{=} A \cup D$ and $A \cap D = \varnothing$. Argument for $(\star)$:

$$\begin{aligned} \omega \in A \cup B \quad &\Leftrightarrow \quad \omega \in A \text{ or } \omega \in B \\ &\Leftrightarrow \quad \omega \in A \text{ or } \omega \in B \setminus (A \cap B) \quad \Leftrightarrow \quad \omega \in A \cup D. \end{aligned}$$

Thus, we see that

$$\begin{aligned} \mathbf{P}[A \cup B] = \mathbf{P}[A \cup D \cup \varnothing \cup \varnothing \cup \ldots] &\overset{\text{(P2)}}{=} \mathbf{P}[A] + \underbrace{\mathbf{P}[D]}_{\overset{(1.37)}{=} \mathbf{P}[B] - \mathbf{P}[A \cap B]} \\ &= \mathbf{P}[A] + \mathbf{P}[B] - \mathbf{P}[A \cap B]. \end{aligned} \tag{1.38}$$

Next, we prove (v). We define the sets

$$B_1 = A_1, \qquad B_n = A_n \setminus \bigcup_{j=1}^{n-1} A_j, \qquad n \geq 2. \tag{1.39}$$

The sets B_j, $j \in \mathbb{N}$, are pairwise disjoint and fulfill $\bigcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} B_j$ as well as $B_j \subseteq A_j$ for every $j \in \mathbb{N}$. Therefore we have that

$$\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \mathbf{P}\left[\bigcup_{j=1}^{\infty} B_j\right] \overset{\text{(P2)}}{=} \sum_{j=1}^{\infty} \mathbf{P}[B_j] \leq \sum_{j=1}^{\infty} \mathbf{P}[A_j]. \tag{1.40}$$

For (vi), we define again the same sets as in (1.39), but now note that $(\star\star)$ $\bigcup_{j=1}^{n} B_j = A_n$, $n \in \mathbb{N}$, and so

$$\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \mathbf{P}\left[\bigcup_{j=1}^{\infty} B_j\right] \overset{\text{(P2)}}{=} \sum_{j=1}^{\infty} \mathbf{P}[B_j] = \lim_{n \to \infty} \sum_{j=1}^{n} \mathbf{P}[B_j] = \lim_{n \to \infty} \mathbf{P}[A_n], \tag{1.41}$$

having used (iv) and $(\star\star)$ in the last step.
Finally, (vii) follows from (ii), (vi) and the fact that if $A_1 \supseteq A_2 \supseteq \ldots$, then $A_1^c \subseteq A_2^c \subseteq \ldots$ □

End of Lecture 2

2. Conditional probability and stochastic independence

(Reference: [GS01, Sections 1.4-1.5], or [Geo12, Sections 3.1, 3.3])

In this chapter we introduce the notion of *conditional probability*. Intuitively, the idea is that the existence of "partial knowledge" should influence how we determine the likelihood of a given outcome.

2.1. Conditional probability

Let us start with a very easy example.

Example 2.1. We throw two dice and ask for the probability that the sum of the numbers of both dice is smaller or equal to 7. We call this event A. Assuming that the dice are fair, this experiment is modelled by

$$(\Omega, \mathscr{F}, \mathbf{P}) = (\{1,2,3,4,5,6\}^2, \mathscr{P}(\Omega), \mathbf{P}), \qquad \mathbf{P}[\,\cdot\,] = \frac{|\cdot|}{36}. \tag{2.1}$$

Of course A is given by

$$\begin{aligned} A =& \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5), \\ &(3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(5,1),(5,2),(6,1)\}. \end{aligned} \tag{2.2}$$

So $\mathbf{P}[A] = \frac{21}{36} = \frac{7}{12}$. Now imagine we are given the information that one of the dice shows the number 6. We call this event B, i.e.

$$B = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6),(6,1),(6,2),(6,3),(6,4),(6,5)\}. \tag{2.3}$$

If we already *know* that B happens, how likely is A? Clearly, the only outcomes of A that can still have occured are

$$A \cap B = \{(1,6),(6,1)\}. \tag{2.4}$$

Thus, knowing that B occured, we should now estimate the probability that A occurs by restricting the sample space Ω to B, so:

$$\frac{\frac{|A\cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \cap B|}{|B|} = \frac{2}{11}. \tag{2.5}$$

We elevate the term on the left-hand side to a definition.

Definition 2.2. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Assume that the event $B \in \mathscr{F}$ has a positive probability $\mathbf{P}[B] > 0$. We define the *conditional probability of* $A \in \mathscr{F}$ *given* B by

$$\mathbf{P}[A|B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}. \tag{2.6}$$

Remark 2.3. (i) If the events A and B are mutually exclusive ($A \cap B = \varnothing$), then we always have $\mathbf{P}[A|B] = 0$, whenever the latter is defined.

(ii) One can rewrite the equation (2.6) as

$$\mathbf{P}[A \cap B] = \mathbf{P}[B] \cdot \mathbf{P}[A|B]. \tag{2.7}$$

This is sometimes called the *multiplication theorem.*

Proposition 2.4. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Assume that the event $B \in \mathscr{F}$ has a positive probability $\mathbf{P}[B] > 0$. Then $\mathbf{P}[\,\cdot\,|B]$ defines a probability distribution on $(\Omega, \mathscr{F})$ as well.*

Proof. We need to verify that $\mathbf{P}[\,\cdot\,|B]$ satisfies the axioms in Definition 1.13. First note that since $\mathbf{P}[B] > 0$, the expression $\mathbf{P}[A|B]$ in (2.6) is well defined for every $A \in \mathscr{F}$. Moreover, since $0 \le \mathbf{P}[A \cap B] \le \mathbf{P}[B]$ (using $A \cap B \subseteq B$ and Proposition 1.15 (iii)), we see that indeed

$$\mathbf{P}[A|B] \in [0,1] \qquad \text{for every } A \in \mathscr{F}. \tag{2.8}$$

Furthermore, we have $\Omega \cap B = B$, and thus

$$\mathbf{P}[\Omega|B] = \frac{\mathbf{P}[\Omega \cap B]}{\mathbf{P}[B]} = \frac{\mathbf{P}[B]}{\mathbf{P}[B]} = 1, \tag{2.9}$$

verifying (P1). Finally, consider $A_j \in \mathscr{F}$, $j \in \mathbb{N}$, pairwise disjoint. Then also the sets $A_j \cap B$ are in $\mathscr{F}$ (using Proposition 1.11 (iii)), and are pairwise disjoint. This implies

$$\begin{aligned}\mathbf{P}\left[\bigcup_{j=1}^{\infty} A_j \middle| B\right] &= \frac{\mathbf{P}\left[\left(\bigcup_{j=1}^{\infty} A_j\right) \cap B\right]}{\mathbf{P}[B]} \\ &= \frac{\mathbf{P}\left[\bigcup_{j=1}^{\infty} (A_j \cap B)\right]}{\mathbf{P}[B]} \overset{\text{(P2)}}{=} \frac{\sum_{j=1}^{\infty} \mathbf{P}[A_j \cap B]}{\mathbf{P}[B]} \\ &= \sum_{j=1}^{\infty} \frac{\mathbf{P}[A_j \cap B]}{\mathbf{P}[B]} = \sum_{j=1}^{\infty} \mathbf{P}[A_j|B]\end{aligned} \tag{2.10}$$

concluding the proof of (P2). □

2.2. The law of total probability and Bayes' theorem

In this section, we fix a probability space $(\Omega, \mathscr{F}, \mathbf{P})$.
The following result is known as the *law of total probability.*

Theorem 2.5. *Let $B_1, ..., B_n \in \mathscr{F}$ with $\mathbf{P}[B_j] > 0$ for all $1 \leq j \leq n$ and $\bigcup_{j=1}^n B_j = \Omega$ with $B_j \cap B_k = \varnothing$ for every $j \neq k$. Then we have for all $A \in \mathscr{F}$ that*

$$\mathbf{P}[A] = \sum_{j=1}^{n} \mathbf{P}[B_j]\mathbf{P}[A|B_j]. \tag{2.11}$$

Proof. Note that

$$\begin{aligned} \mathbf{P}[A] &= \mathbf{P}\left[A \cap \Omega\right] = \mathbf{P}\left[A \cap \bigcup_{j=1}^{n} B_j\right] = \mathbf{P}\left[\bigcup_{j=1}^{n} (A \cap B_j)\right] \\ &\overset{\text{Prop. 1.15, (iv)}}{=} \sum_{j=1}^{n} \mathbf{P}[A \cap B_j] = \sum_{j=1}^{n} \mathbf{P}[B_j]\mathbf{P}[A|B_j]. \end{aligned} \tag{2.12}$$

□

We can right away combine the previous theorem with the definition of the conditional probability to obtain the following result, called *Bayes' theorem*:

Theorem 2.6. *Under the same assumptions as Theorem 2.5, we have for every $1 \leq k \leq n$, that*

$$\mathbf{P}[B_k|A] = \frac{\mathbf{P}[B_k]\mathbf{P}[A|B_k]}{\sum_{j=1}^{n} \mathbf{P}[B_j]\mathbf{P}[A|B_j]}. \tag{2.13}$$

Proof.

$$\mathbf{P}[B_k|A] \overset{(2.6)}{=} \frac{\mathbf{P}[B_k \cap A]}{\mathbf{P}[A]} \overset{(2.7),(2.11)}{=} \frac{\mathbf{P}[B_k]\mathbf{P}[A|B_k]}{\sum_{j=1}^{n} \mathbf{P}[B_j]\mathbf{P}[A|B_j]}. \tag{2.14}$$

□

Example 2.7. Biochemical tests for a certain marker / antigen / disease / ... within a population are never absolutely reliable. We consider a test with the following properties:

Let T denote the event "the test is positive". Let M be the event "a given individual has the marker". We assume that the test in question satisfies:

$$\begin{aligned} \mathbf{P}[T|M] &= 0.99 \qquad \textit{(sensitivity)}, \\ \mathbf{P}[T^c|M^c] &= 0.99 \qquad \textit{(specificity)}, \end{aligned} \tag{2.15}$$

and the marker we are looking for is such that for the population which is considered,

$$\mathbf{P}[M] = 0.01 \qquad \textit{(prevalence)}. \tag{2.16}$$

What is the probability that someone who tests positive actually has the maker / antigen / disease?

Observe that $\mathbf{P}[T|M^c] = 1 - 0.99 = 0.01$. We use Bayes' theorem 2.6

$$\begin{aligned}\mathbf{P}[M|T] &= \frac{\mathbf{P}[T|M] \cdot \mathbf{P}[M]}{\mathbf{P}[T|M] \cdot \mathbf{P}[M] + \mathbf{P}[T|M^c] \cdot \mathbf{P}[M^c]} \\ &= \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.01 \cdot 0.99} \\ &= \frac{1}{2}.\end{aligned} \tag{2.17}$$

We see that because the trait under consideration is so rare, we find that half of those who test positive do not actually admit this trait, even though the test is fairly reliable!

Remark 2.8. Both the law of total probability and Bayes' theorem are valid if we have countably many pairwise disjoint $B_1, B_2, \ldots \in \mathscr{F}$ with $\mathbf{P}[B_j] > 0$ for all $j \in \mathbb{N}$ and $\bigcup_{j=1}^{\infty} B_j = \Omega$. In this case, (2.11) and (2.13) become

$$\mathbf{P}[A] = \sum_{j=1}^{\infty} \mathbf{P}[B_j]\mathbf{P}[A|B_j], \text{ and} \tag{2.18}$$

$$\mathbf{P}[B_k|A] = \frac{\mathbf{P}[B_k]\mathbf{P}[A|B_k]}{\sum_{j=1}^{\infty} \mathbf{P}[B_j]\mathbf{P}[A|B_j]}, \tag{2.19}$$

respectively.

End of Lecture 3

2.3. Stochastic independence

We now introduce the notion of *stochastic independence* of events, which is one of the central concepts in probability theory. Again, we fix a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ in this section.

Heuristics: The events $A, B \in \mathscr{F}$ should be independent, if the occurence of A has no influence on the occurence of B, and vice versa. Specifically, if A happens, it should neither be more, nor less likely that B occurs, and vice versa, so

$$\begin{aligned}\mathbf{P}[A] = \mathbf{P}[A|B] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[B]}, \\ \mathbf{P}[B] = \mathbf{P}[B|A] = \frac{\mathbf{P}[A \cap B]}{\mathbf{P}[A]},\end{aligned} \tag{2.20}$$

where we implicitly assumed that $\mathbf{P}[A], \mathbf{P}[B] > 0$. We turn this reasoning into a definition.

Definition 2.9. (i) The events $A, B \in \mathscr{F}$ are called *(stochastically) independent*, if

$$\mathbf{P}[A \cap B] = \mathbf{P}[A] \cdot \mathbf{P}[B]. \tag{2.21}$$

(ii) Let I be a set. The events $(A_i)_{i \in I} \subseteq \mathscr{F}$ are called *jointly (stochastically) independent*, if for every $\{i_1, \ldots, i_m\} \subseteq I$ with $i_1, \ldots, i_m$ pairwise distinct,

$$\mathbf{P}[A_{i_1} \cap \ldots \cap A_{i_m}] = \mathbf{P}[A_{i_1}] \cdot \ldots \cdot \mathbf{P}[A_{i_m}]. \tag{2.22}$$

Remark 2.10. (i) Note that both definitions include the case that a given event has probability zero. If we assume that $\mathbf{P}[A] > 0$ and $\mathbf{P}[B] > 0$, then (2.20) and (2.21) are equivalent.

(ii) The events $\varnothing$ and Ω are independent from any other given event. Intuitively, they contain "no additional information" on the probability.

(iii) Equation (2.22) means that the occurence of any subset of the events $A_1, \ldots, A_n$ does not give additional information on the occurence of the others. For instance,

$$\mathbf{P}[A_1 | A_2 \cap \ldots \cap A_n] = \frac{\mathbf{P}[A_1 \cap A_2 \cap \ldots \cap A_n]}{\mathbf{P}[A_2 \cap \ldots \cap A_n]} = \frac{\prod_{j=1}^n \mathbf{P}[A_j]}{\prod_{j=2}^n \mathbf{P}[A_j]} = \mathbf{P}[A_1], \tag{2.23}$$

provided that $\mathbf{P}[A_2 \cap \ldots \cap A_n] > 0$.

(iv) We stress that stochastic independence of two events A and B has nothing to do with them being disjoint as sets! In fact, If A and B are disjoint, then

$$0 = \mathbf{P}[\varnothing] = \mathbf{P}[A \cap B] = \mathbf{P}[A] \cdot \mathbf{P}[B], \tag{2.24}$$

so unless $\mathbf{P}[A] = 0$ or $\mathbf{P}[B] = 0$, disjoint events A and B are not independent.

We illustrate the concept of independence with a number of examples.

Example 2.11. (i) We draw a card randomly from a standard card deck[1], with

$$\Omega = \big\{(i,j)\,;\, i \in \{\clubsuit, \spadesuit, \blacklozenge, \heartsuit\}, j \in \{1, 2, \ldots, 13\}\big\}, \tag{2.25}$$

equipped with the discrete uniform distribution. Consider the events

$$\begin{aligned} A &= \big\{(\heartsuit, j)\,;\, j \in \{1, \ldots, 13\}\big\} = \text{ drawing a } \heartsuit\text{-card}, \\ B &= \big\{(i, 1)\,;\, i \in \{\clubsuit, \spadesuit, \blacklozenge, \heartsuit\}\big\} = \text{ drawing an ace}. \end{aligned} \tag{2.26}$$

Clearly, we have that

$$A \cap B = \big\{(\heartsuit, 1)\big\}. \tag{2.27}$$

With this, we see that

$$\begin{aligned} \mathbf{P}[A] = \frac{|A|}{|\Omega|} = \frac{13}{52} = \frac{1}{4}, \qquad \mathbf{P}[B] = \frac{|B|}{|\Omega|} = \frac{4}{52} = \frac{1}{13} \\ \mathbf{P}[A \cap B] = \frac{|A \cap B|}{|\Omega|} = \frac{1}{52} = \mathbf{P}[A] \cdot \mathbf{P}[B]. \end{aligned} \tag{2.28}$$

So A and B are independent.

[1] With 52 *French-suited playing cards*.

(ii) Consider tossing a fair coin twice:

$$\begin{aligned}
\Omega &= \{(H,H),(H,T),(T,H),(T,T)\},\\
A &= \text{“Heads” comes up in the first round}\\
&= \{(H,H),(H,T)\}, & \mathbf{P}[A] &= \frac{1}{2},\\
B &= \text{“Heads” comes up in the second round}\\
&= \{(H,H),(T,H)\}, & \mathbf{P}[B] &= \frac{1}{2},\\
& & \mathbf{P}[A\cap B] &= \frac{1}{4},\\
C &= \text{“Heads” comes up exactly once}\\
&= \{(H,T),(T,H)\}, & \mathbf{P}[C] &= \frac{1}{2},\\
& & \mathbf{P}[A\cap C] = \frac{1}{4} &= \mathbf{P}[B\cap C],
\end{aligned}$$

However: $\mathbf{P}[A\cap B\cap C] = \mathbf{P}[\varnothing] = 0 \neq \mathbf{P}[A]\cdot\mathbf{P}[B]\cdot\mathbf{P}[C]$.

This shows that the events A, B and C are *pairwise* independent (this means every two events out of $\{A,B,C\}$ are independent), but not jointly independent.

We finish this section with the following result:

Theorem 2.12. *Let $A_1, A_2, ..., A_n \in \mathscr{F}$ be jointly independent. Then also $B_1, B_2, ..., B_n$ with $B_i \in \{A_i, A_i^c\}$, for $1 \le i \le n$, are jointly independent.*

Proof. We only show the case $n = 2$ (the general case follows by induction).

$$\begin{aligned}
&(A_1\cap A_2)\cup(A_1\cap A_2^c) = A_1 \qquad \text{(disjoint union)}\\
\Rightarrow\quad &\underbrace{\mathbf{P}[A_1\cap A_2]}_{=\mathbf{P}[A_1]\cdot\mathbf{P}[A_2]} + \mathbf{P}[A_1\cap A_2^c] = \mathbf{P}[A_1]\\
\Rightarrow\quad &\mathbf{P}[A_1\cap A_2^c] = \mathbf{P}[A_1]\cdot(1-\mathbf{P}[A_2]) = \mathbf{P}[A_1]\cdot\mathbf{P}[A_2^c].
\end{aligned} \tag{2.29}$$

By changing the roles of A_1 and A_2, we also have

$$\mathbf{P}[A_1^c\cap A_2] = \mathbf{P}[A_1^c]\cdot\mathbf{P}[A_2]. \tag{2.30}$$

We can finally use the same argument as in (2.29) (which implied the independence of A_1 and A_2^c from the independence of A_1 and A_2) to infer the independence of A_1^c and A_2^c from the independence of A_1^c and A_2. □

3. Discrete distributions

(Reference: [GS01, Section 3.1], or [Geo12, Sections 2.1-2.5.1])

In the present chapter the most important discrete distributions are defined. We need to start with a short reminder on countable sets and sums over countable sets.

Definition 3.1. A set Ω is called *countable* if it is empty or there is a surjective map $\rho : \mathbb{N} \to \Omega$ (i.e. for every $\omega \in \Omega$, there exists $j \in \mathbb{N}$ with $\rho(j) = \omega$).[1]
For a countable set Ω and a function $f : \Omega \to [0, \infty)$, we define the *sum of f over Ω* by

$$\sum_{\omega \in \Omega} f(\omega) = \sup_{\substack{F \subseteq \Omega \\ F \text{finite}}} \sum_{\omega \in F} f(\omega). \tag{3.1}$$

The expression on the right-hand side is an element of $[0, \infty] = [0, \infty) \cup \{+\infty\}$. In (3.1) and in the following, we use the convention that $\sum_{\omega \in \varnothing} f(\omega) = 0$ for an empty sum. By slight abuse of notation, we also write $\sum_{\omega \in A} f(\omega)$ instead of $\sum_{\omega \in A} f_{|A}(\omega)$ for $A \subseteq \Omega$ countable and infinite.

Remark 3.2. (i) The definition is consistent with finite sets: Indeed, if $\Omega = \{\omega_1, ..., \omega_N\}$ is a finite set, we immediately have that

$$\sum_{\omega \in \Omega} f(\omega) = \sum_{i=1}^{N} f(\omega_i). \tag{3.2}$$

(ii) Suppose that Ω is countable, but infinite (such as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}^2$, ...). In this case, we can write $\Omega = \{\omega_1, \omega_2, ...\}$ by constructing a (not necessarily unique) bijective function $\rho_* : \mathbb{N} \to \Omega$, $\omega_j = \rho_*(j)$, which we call *enumeration*. In this case, one can show that

$$\sum_{\omega \in \Omega} f(\omega) = \sum_{j=1}^{\infty} f(\omega_j) = \lim_{N \to \infty} \sum_{j=1}^{N} f(\omega_j), \tag{3.3}$$

and the limit on the right-hand side does not depend on the choice of the enumeration.

(iii) If A_j, $j \in \mathbb{N}$ are pairwise disjoint ($A_j \cap A_k = \varnothing$ for $j \neq k$) with $\Omega = \bigcup_{j=1}^{\infty} A_j$, then one can show the *rearrangement theorem*:

$$\sum_{\omega \in \Omega} f(\omega) = \sum_{j=1}^{\infty} \sum_{\omega \in A_j} f(\omega). \tag{3.4}$$

[1] In particular, finite sets are countable by this convention.

End of Lecture 4

Definition 3.3. Let $(\Omega, \mathscr{P}(\Omega), \mathbf{P})$ be a probability space with a countable sample space Ω. The probability measure $\mathbf{P}$ is then called a *discrete distribution.* The function

$$p : \Omega \to [0,1], \qquad p(\omega) = \mathbf{P}[\{\omega\}] \tag{3.5}$$

is the *probability mass function* of the distribution.

Obviously, if we are given a discrete distribution on $(\Omega, \mathscr{P}(\Omega))$, the probability mass function is uniquely determined. Conversely, every set $(p(\omega))_{\omega \in \Omega}$ of non-negative numbers with $\sum_{\omega \in \Omega} p(\omega) = 1$ determines a unique probability measure on $(\Omega, \mathscr{P}(\Omega))$, which is the statement of the following proposition.

Proposition 3.4. *Let Ω be countable and $p : \Omega \to [0,1]$ a map fulfilling*

$$\sum_{\omega \in \Omega} p(\omega) = 1. \tag{3.6}$$

Then the map

$$\mathbf{P} : \mathscr{P}(\Omega) \to [0,1], \qquad A \mapsto \mathbf{P}[A] = \sum_{\omega \in A} p(\omega) \tag{3.7}$$

defines a probability measure on $(\Omega, \mathscr{P}(\Omega))$.

Proof. We first remark that since Ω is countable and the real numbers $(p(\omega))_{\omega \in \Omega}$ are non-negative, we have

$$\sum_{\omega \in \Omega} p(\omega) = \begin{cases} \sum_{i=1}^N p(\omega_i), & \Omega = \{\omega_1, ..., \omega_N\} \text{ finite,} \\ \lim_{N \to \infty} \sum_{i=1}^N p(\omega_i), & \Omega = \{\omega_1, \omega_2, ...\} \text{ infinite,} \end{cases} \tag{3.8}$$

and the value of the series does not depend on the choice of the enumeration (see Remark 3.2). Moreover, since $A \subseteq \Omega$ is also countable, the expression for $\mathbf{P}[A]$ in (3.7) is well-defined and in $[0,1]$.

The condition (P1) is immediate by (3.6). For (P2), we consider $A_j \in \mathscr{P}(\Omega)$ for $j \in \mathbb{N}$ pairwise disjoint and use

$$\begin{aligned} \mathbf{P}\left[\bigcup_{j=1}^\infty A_j\right] &= \sum_{\omega \in \bigcup_{j=1}^\infty A_j} p(\omega) \overset{(3.4)}{=} \sum_{j=1}^\infty \sum_{\omega \in A_j} p(\omega) \\ &= \sum_{j=1}^\infty \mathbf{P}[A_j]. \end{aligned} \tag{3.9}$$

□

We will now present some of the most important discrete distributions.

The discrete uniform distribution $\mathscr{U}(\Omega)$

$$\Omega \text{ finite }, \qquad p(\omega) = \frac{1}{|\Omega|}. \tag{3.10}$$

This is just giving a name for the distribution considered already multiple times, see Example 1.14.

The Bernoulli distribution $Ber(p)$

$$\Omega = \{0, 1\}, \qquad p(1) = p \in [0, 1], \qquad p(0) = 1 - p. \tag{3.11}$$

The Bernoulli distribution models random experiments in which a “success” occurs with probability p, and a “failure” occurs with probability $1 - p$ (for instance, tossing a biased coin). Such experiments are also called *Bernoulli experiments.*

The Binomial distribution $Bin(n, p)$

$$\Omega = \{0, 1, ..., n\}, \qquad p(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad p \in [0, 1], n \in \mathbb{N}. \tag{3.12}$$

Note that

$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (p + 1 - p)^n = 1. \tag{3.13}$$

The binomial distribution is extending the Bernoulli distribution in the following way: It models how many attempts out of n independent experiments with the same success parameter $p \in [0, 1]$ are successful. To explain it, consider the auxiliary probability space

$$(\{0, 1\}^n, \mathscr{P}(\{0, 1\}^n), \mathbf{Q}), \qquad \mathbf{Q}[\{(\omega_1, ..., \omega_n)\}] = \underbrace{p^{\sum_{j=1}^n \omega_j}}_{p^{\#\text{ of successes}}} \underbrace{(1 - p)^{n - \sum_{j=1}^n \omega_j}}_{(1-p)^{\#\text{ of failures}}}. \tag{3.14}$$

Here, the string $(\omega_1, ..., \omega_n) \in \{0, 1\}^n$ stands for the successes and failures of the experiment in the order observed, i.e.

$$\omega_j = \begin{cases} 1, & \text{if the } j\text{th experiment is a success,} \\ 0, & \text{if the } j\text{th experiment is a failure.} \end{cases} \tag{3.15}$$

For instance, the string $(1, 0, 0, 1)$ means that the first and last of four experiments are successes, whereas the second and third experiments are failures. By the product structure in (3.14), the experiments are independent. Now consider the event (for $0 \leq k \leq n$)

$$E_k = \Big\{(\omega_1, ..., \omega_n) \in \{0, 1\}^n \,;\, \sum_{j=1}^{n} \omega_j = k\Big\} = \text{“exactly } k \text{ sucesses”.} \tag{3.16}$$

We have

$$\mathbf{Q}[E_k] = |E_k| p^k (1 - p)^{n-k} = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{3.17}$$

Example 3.5. We throw a die 4 times and we are interested in the number of times that the number six shows up. This is modelled by the binomial distribution $Bin(4, \frac{1}{6})$. In the description (3.14) "0" stands for the occurence of a number other than six (failure), whereas "1" stands for the occurence of a six (success). In this example, we have

Probability	Outcomes in $\{0, 1\}^4$
$p(0) = \left(\frac{5}{6}\right)^4$	$(0, 0, 0, 0)$
$p(1) = \binom{4}{1} \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3,$	$(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)$
$p(2) = \binom{4}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^2$	$(1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 1, 1, 0), (0, 1, 0, 1), (0, 0, 1, 1)$
$p(3) = \binom{4}{3} \cdot \left(\frac{1}{6}\right)^3 \cdot \frac{5}{6}$	$(0, 1, 1, 1), (1, 0, 1, 1), (1, 1, 0, 1), (0, 0, 0, 1)$
$p(4) = \left(\frac{1}{6}\right)^4$	$(1, 1, 1, 1)$

We can also order the outcomes in the form of a "tree diagram" as follows:

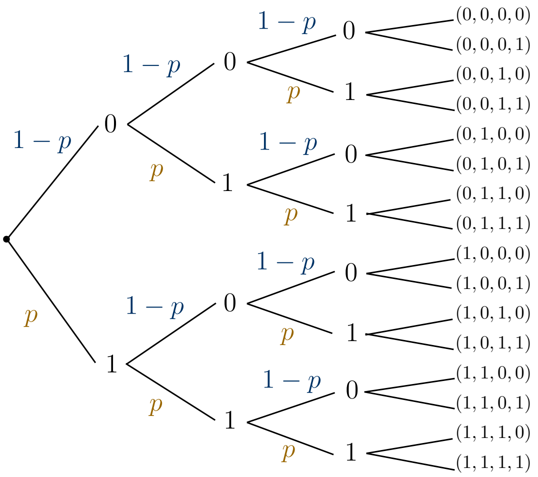

Figure 3.1.: Tree diagram of 4 successive independent Bernoulli experiments.

Remark 3.6. The above example shows that the same question "How likely is it that the number six comes up exactly twice when rolling a die four times?" is treated much more efficiently on the probability space

$$(\Omega = \{0, 1, 2, 3, 4\}, \mathscr{P}(\Omega), \mathbf{P}), \qquad \mathbf{P} = Bin(4, \tfrac{1}{6}),$$

where we simply have

$$\mathbf{P}[\text{“2 sixes”}] = \mathbf{P}[\{2\}] = \binom{4}{2} \cdot \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2,$$

than on the probability space

$$(\widetilde{\Omega} = \{0, 1\}^n, \mathscr{P}(\widetilde{\Omega}), \mathbf{Q}), \qquad \mathbf{Q}[\{(\omega_1, ..., \omega_n)\}] = p^{\sum_{j=0}^n \omega_j}(1-p)^{n-\sum_{j=0}^n \omega_j},$$

where

$$\begin{aligned}\mathbf{Q}[\text{“2 sixes”}] &= \mathbf{Q}[\{(1,1,0,0),(1,0,1,0),(1,0,0,1),(0,1,1,0),(0,1,0,1),(0,0,1,1)\}]\\ &= \binom{4}{2}\cdot\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2.\end{aligned}$$

The information we need is already contained in the space $(\Omega = \{0,1,2,3,4\}, \mathscr{P}(\Omega), \mathbf{P})$. This concept of a *reduction in complexity* will motivate the study of random variables later.

The Geometric distribution $Geo(p)$

$$\Omega = \mathbb{N}, \qquad p(k) = (1-p)^{k-1}p, \qquad p \in (0,1). \tag{3.18}$$

Note that

$$\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = \frac{p}{1-(1-p)} = 1. \tag{3.19}$$

The interpretation of the geometric distribution is the number of repetitions of a Bernoulli experiment (with success parameter $p \in (0,1)$) until the first success.

The Hypergeometric distribution $H(N, M, n)$

$$\Omega = \{0, ..., n\}, \qquad p(k) = \frac{\binom{M}{k}\cdot\binom{N-M}{n-k}}{\binom{N}{n}}, \qquad N, m, n \in \mathbb{N}, 0 \leq n, M \leq N. \tag{3.20}$$

The hypergeometric distribution should be understood as follows: Out of a set of N elements, M subelements have a certain favorable property. We choose uniformly at random an unordered sample of $0 \leq n \leq N$ elements out of the large set without repetitions. Then $p(k)$ denotes the probability that exactly $0 \leq k \leq n$ have the the favorable property (this probability is always zero if $M < k$, which can happen if $n > M$).

Example 3.7. In an urn there are 10 balls, three are green and seven are red. We draw (at once) four balls from the urn. Here

$$N = 10 \text{ (\# of balls in urn)}, \qquad M = 3 \text{ (\# of green balls)}, \qquad n = 4 \text{ (\# of balls drawn)}. \tag{3.21}$$

The probability that exactly two of the balls drawn are green is

$$p(2) = \frac{\binom{3}{2}\cdot\binom{7}{2}}{\binom{10}{4}}.$$

The Poisson distribution $Pois(\lambda)$

$$\Omega = \mathbb{N}_0, \qquad p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \lambda > 0. \tag{3.22}$$

Note that

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1. \tag{3.23}$$

The Poisson distribution is a natural distribution for modelling events that in principle can occur infinitely often (for instance the number of goals in a football game, or the number of raindrops falling in a given area during a given time, ...).

An important application is the following approximation of the Binomial distribution by the Poisson distribution:

Proposition 3.8. *Let $\lambda > 0$ be fixed and $p_n = \frac{\lambda}{n}$ for $n \in \mathbb{N}$. Then, for every $k \in \mathbb{N}_0$, it holds that*

$$\lim_{n \to \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{3.24}$$

Proof.

$$\begin{aligned} \binom{n}{k} p_n^k (1 - p_n)^{n-k} &= \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\to \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda} \cdot 1, \qquad \text{as } n \to \infty. \end{aligned} \tag{3.25}$$

□

End of Lecture 5

4. Continuous distributions

(Reference: [GS01, Section 4.1], or [Geo12, Sections 1.2, 2.5.2, 2.6])

In this section, we introduce continuous distributions on $\Omega \subseteq \mathbb{R}$. Specifically, we want to be able to talk about probability spaces like $(\mathbb{R}, \mathscr{F}, \mathbf{P})$ or $([0,1], \mathscr{F}, \mathbf{P})$ with appropriate choices of $\mathscr{F}$ and $\mathbf{P}$. This requires some more details about σ-algebras, which we will present (mostly) without proofs.

4.1. Probability density functions

Example 4.1. Consider the arrival of a minibus with delay. We assume that its arrival is "uniformly distributed" between 1 PM and 1:15 PM. How can we model this? Suppose $0 \widehat{=}$ 1 PM and $1 \widehat{=}$ 1:15 PM. We split $[0,1)$ in half-open intervals of equal length $\frac{1}{n}$, $n \in \{2,3,4,\dots\}$ whose leftmost points are

$$\left\{\tfrac{j}{n}\,;\, 0 \le j \le n-1\right\} = \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \dots, 1 - \tfrac{1}{n}\right\} \subseteq [0,1). \tag{4.1}$$

The probability for the minibus to arrive within one of these intervals $\Delta_j = [\frac{j}{n}, \frac{j+1}{n})$, $0 \le j \le n-1$ should be $\frac{1}{n}$. For $0 \le a < b < 1$, we should have approximately

$$\mathbf{P}\big[[a,b)\big] \approx \sum_{a \le \frac{j}{n} < b} \frac{1}{n} = \frac{1}{n} \sum_{a \le \frac{j}{n} < b} 1 \longrightarrow \int_a^b \underbrace{1}_{=:f(x)} \,\mathrm{d}x, \qquad \text{as } n \to \infty. \tag{4.2}$$

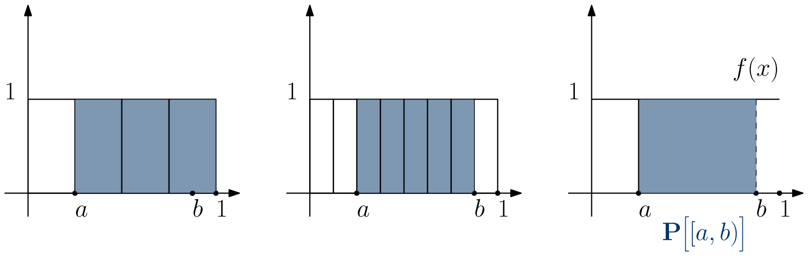

Figure 4.1.: The first and second panel represent the sum expression in (4.2) above for $n = 4$ and $n = 8$ respectively. The third panel represents the integral expression in the same equation.

Suppose now that the minibus has the highest chance to arrive around 1:15 PM. More specifically, let us assume that the probability to arrive within the interval $\Delta_j = [\frac{j}{n}, \frac{j+1}{n})$ is

$\frac{2j}{n(n-1)}$ for $0 \le j \le n-1$. Note that

$$\sum_{j=0}^{n-1} \frac{2j}{n(n-1)} = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} j = \frac{2}{n(n-1)} \cdot \frac{n(n-1)}{2} = 1. \tag{4.3}$$

This suggests that for $0 \le a < b < 1$, we should have

$$\mathbf{P}\big[[a,b)\big] \approx \sum_{a \le \frac{j}{n} < b} \frac{2j}{n(n-1)} = \frac{1}{n-1} \sum_{a \le \frac{j}{n} < b} 2\frac{j}{n} \longrightarrow \int_a^b \underbrace{2x}_{=:f(x)} \,\mathrm{d}x, \qquad \text{as } n \to \infty. \tag{4.4}$$

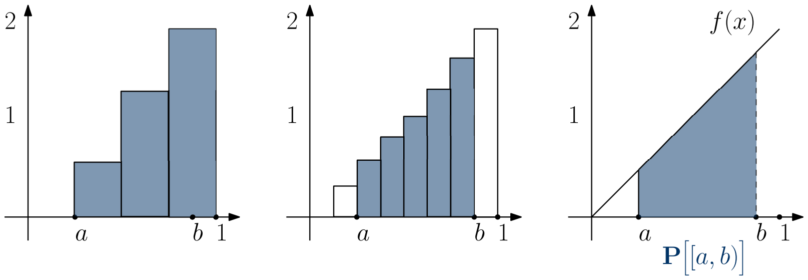

Figure 4.2.: The first and second panel represent the sum expression in (4.4) above for $n = 4$ and $n = 8$ respectively. The third panel represents the integral expression in the same equation.

The above considerations show that similarly to the probability mass function $(p(\omega))_{\omega\in\Omega}$ for countable Ω, and probabilities being defined as sums, we may want to define probabilities of intervals $[a,b) \subseteq [0,1)$ (or more generally $[a,b) \subseteq \mathbb{R}$) as integrals over a function f. Let us turn this into a definition.

Definition 4.2. A piecewise continuous function $f : \mathbb{R} \to [0,\infty)$ is called *probability density function* if

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \lim_{r,s\to\infty} \int_{-r}^{s} f(x)\mathrm{d}x = 1. \tag{4.5}$$

We can then define for any interval $[a,b) \subseteq \mathbb{R}$ witb $a < b$:

$$\mathbf{P}\big[[a,b)\big] = \int_a^b f(x)\mathrm{d}x \in [0,1]. \tag{4.6}$$

To calculate the probability of intervals, (4.6) can be taken as a definition.

4.2. Generated σ-algebras and the Borel-σ-algebra

A problem arises now if we want to define the probability $\mathbf{P}[A]$ for sets $A \subseteq \mathbb{R}$ which are *not* intervals. We would like our probability space to be $([0,1), \mathscr{P}([0,1)), \mathbf{P})$ or $(\mathbb{R}, \mathscr{P}(\mathbb{R}), \mathbf{P})$ with $\mathbf{P}$ fulfilling (4.6). Something like this is however impossible, as the following deep result shows:

Theorem 4.3. *There is no probability measure* $\mathbf{P}$ *on* $([0,1), \mathscr{P}([0,1)))$ *such that* $\mathbf{P}[[a,b)] = b - a$ *for every* $0 \leq a < b < 1$.

Proof. Measure Theory – Omitted. □

The solution to this obstacle is to use a smaller σ-algebra than $\mathscr{P}([0,1))$ or $\mathscr{P}(\mathbb{R})$. We need the following preparations.

Proposition 4.4. *Let* Ω *a non-empty set and* $\mathscr{E} \subseteq \mathscr{P}(\Omega)$. *The class defined by*

$$\sigma(\mathscr{E}) = \bigcap_{\substack{\mathscr{E} \subseteq \mathscr{F} \\ \mathscr{F}\ \sigma\text{-algebra}}} \mathscr{F} \subseteq \mathscr{P}(\Omega) \tag{4.7}$$

is a σ*-algebra on* Ω*, called the* σ-algebra generated by $\mathscr{E}$.

Proof. This follows from the fact that the intersection of (arbitrarily) many σ-algebras on Ω is again a σ-algebra. □

The σ-algebra $\sigma(\mathscr{E})$ is the smallest σ-algebra on Ω that contains $\mathscr{E}$.

Example 4.5. (i) For any non-empty set Ω we have $\sigma(\{\Omega\}) = \{\varnothing, \Omega\}$. Indeed, the right-hand side is a σ-algebra containing $\{\Omega\}$, and every σ-algebra must contain $\{\varnothing, \Omega\}$.

(ii) Consider $\Omega = \{1,2,3,4,5,6\}$. What is $\sigma\big(\{\{1\}\}\big)$? Since this is a σ-algebra that contains $\{1\}$ as an element, it must also contain $\{1\}^c = \{2,3,4,5,6\}$ as an element. We then have

$$\sigma\big(\{\{1\}\}\big) = \big\{\varnothing, \{1\}, \{2,3,4,5,6\}, \Omega\big\},$$

since the expression on the right-hand side of the above equation is a σ-algebra itself.

The generated σ-algebra can also be understood as follows: Suppose specify the sets $A \in \mathscr{E}$ to be observable events. Then $\sigma(\mathscr{E})$ is precisely the "class of *all* events that can be observed within the model". We use this idea to *define* a σ-algebra on $\mathbb{R}$ (or on $[a,b)$, $[a,b]$, (a,b) and $(a,b]$ for $a < b$).

Definition 4.6. The *Borel* σ*-algebra* $\mathscr{B}(\mathbb{R})$ on $\mathbb{R}$ is defined by

$$\mathscr{B}(\mathbb{R}) = \sigma\left(\{[a,b)\,;\ a,b \in \mathbb{R}, a < b\}\right). \tag{4.8}$$

For $a < b$ we also define the *Borel* σ*-algebras* on $[a,b]$, $[a,b)$, $(a,b]$, (a,b) by

$$\begin{aligned}
\mathscr{B}([a,b]) &= \{A \cap [a,b]\,;\ A \in \mathscr{B}(\mathbb{R})\},\\
\mathscr{B}([a,b)) &= \{A \cap [a,b)\,;\ A \in \mathscr{B}(\mathbb{R})\},\\
\mathscr{B}((a,b]) &= \{A \cap (a,b]\,;\ A \in \mathscr{B}(\mathbb{R})\},\\
\mathscr{B}((a,b)) &= \{A \cap (a,b)\,;\ A \in \mathscr{B}(\mathbb{R})\}.
\end{aligned} \tag{4.9}$$

By definition, $\mathscr{B}(\mathbb{R})$ contains all intervals $[a,b)$, $a<b$, so these sets are events. Are intervals like $[a,b]$ or points also events?

Example 4.7. Let $a,b\in\mathbb{R}$, $a<b$. Then $\{a\},[a,b],(a,b],(a,b)\in\mathscr{B}(\mathbb{R})$. If A is a countable subset of $\mathbb{R}$, then $A\in\mathscr{B}(\mathbb{R})$.

Indeed, we have for instance that

$$\{a\}=\bigcap_{n=1}^{\infty}\underbrace{[a,a+\tfrac{1}{n})}_{\in\mathscr{B}(\mathbb{R})}\in\mathscr{B}(\mathbb{R}) \tag{4.10}$$

by Proposition 1.11, (ii). Then we have that

$$(a,b)=[a,b)\setminus\{a\}=\underbrace{[a,b)}_{\in\mathscr{B}(\mathbb{R})}\cap\underbrace{\{a\}^c}_{\in\mathscr{B}(\mathbb{R})}\in\mathscr{B}(\mathbb{R}), \tag{4.11}$$

again by Proposition 1.11, and the other claims about intervals follow similarly. If $A\subseteq\mathbb{R}$ is countable, then

$$A=\{a_1,a_2,a_3,...\}=\bigcup_{n=1}^{\infty}\underbrace{\{a_n\}}_{\in\mathscr{B}(\mathbb{R})}\in\mathscr{B}(\mathbb{R}), \tag{4.12}$$

by (S3).

Every set that can be obtained from intervals or points by taking unions, intersections and complements is an element of $\mathscr{B}(\mathbb{R})$. Somewhat informally, we can say that "every set of our natural imagination is in $\mathscr{B}(\mathbb{R})$". However, $\mathscr{B}(\mathbb{R})\neq\mathscr{P}(\mathbb{R})$, and this is what gives us a chance to define a probability measure $\mathbf{P}$ on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$:

Proposition 4.8. *Let $f:\mathbb{R}\to\mathbb{R}$ be a probability density function. Then there exists a unique probability measure $\mathbf{P}$ on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ with*

$$\mathbf{P}\big[[a,b)\big]=\int_a^b f(x)\mathrm{d}x. \tag{4.13}$$

Such a probability measure $\mathbf{P}$ is called a continuous distribution with density f.

Proof. Measure Theory – Omitted. □

End of Lecture 6

4.3. Properties of continuous distributions

Let us record the following simple observations:

Lemma 4.9. *Let $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbf{P})$ be a probability space with a continuous distribution $\mathbf{P}$ and let $a, b \in \mathbb{R}$ with $a < b$. Then*

$$\mathbf{P}\big[\{a\}\big] = 0, \tag{4.14}$$

$$\mathbf{P}\big[[a,b]\big] = \mathbf{P}\big[(a,b]\big] = \mathbf{P}\big[(a,b)\big] = \int_a^b f(x)\mathrm{d}x, \tag{4.15}$$

$$\mathbf{P}\big[(-\infty,a]\big] = \mathbf{P}\big[(-\infty,a)\big] = \int_{-\infty}^a f(x)\mathrm{d}x, \tag{4.16}$$

$$\mathbf{P}\big[[a,\infty)\big] = \mathbf{P}\big[(a,\infty)\big] = \int_a^\infty f(x)\mathrm{d}x \tag{4.17}$$

Proof. Let $\varepsilon > 0$. Then $\{a\} \subseteq [a, a+\varepsilon)$, so

$$\mathbf{P}\big[\{a\}\big] \leq \mathbf{P}\big[[a,a+\varepsilon)\big] = \int_a^{a+\varepsilon} f(x)\mathrm{d}x. \tag{4.18}$$

which tends to zero as $\varepsilon \to 0$.[1] Thus we have (4.14). The other claims follow easily. □

Let us give some of the most important continuous distributions.

The uniform distribution $\mathscr{U}([a,b])$

$$f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b], \end{cases} \qquad a < b. \tag{4.19}$$

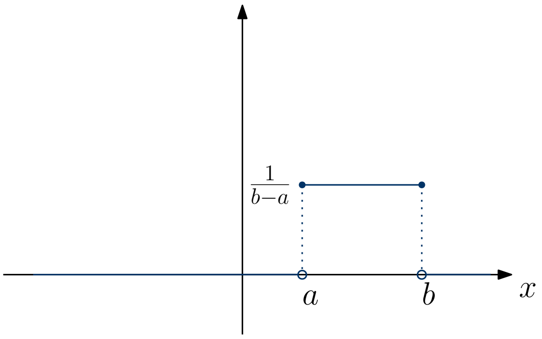

Figure 4.3.: Probability density function for $\mathscr{U}([0,1])$.

[1] This is immediately clear if f is continuous in a. However, we do not exclude a case where $\lim_{\varepsilon\downarrow 0} f(a+\varepsilon) = \infty$, as long as f is still integrable (for instance $f(x) = \frac{1}{2\sqrt{x}}\mathbb{1}_{(0,1]}(x)$ at $a = 0$). In this case, at the point a one has to argue that $\int_a^{a+\varepsilon} f(x)\mathrm{d}x + \int_{a+\varepsilon}^{a+1} f(x)\mathrm{d}x = c \leq 1$, and since $\int_a^{a+1} f(x)\mathrm{d}x = \lim_{\varepsilon\downarrow 0}\int_{a+\varepsilon}^{a+1} f(x)\mathrm{d}x = c$, we must have $\int_a^{a+\varepsilon} f(x)\mathrm{d}x \to 0$, since $f \geq 0$. See also the Appendix A for a review on improper Riemann integrals.

The exponential distribution $\mathscr{E}(\lambda)$

$$f(x) = \lambda e^{-\lambda x}\mathbb{1}_{[0,\infty)}(x) = \begin{cases}\lambda e^{-\lambda x}, & x \geq 0,\\ 0, & x < 0\end{cases} \qquad \lambda > 0. \tag{4.20}$$

Note that this is indeed a probability density function, since

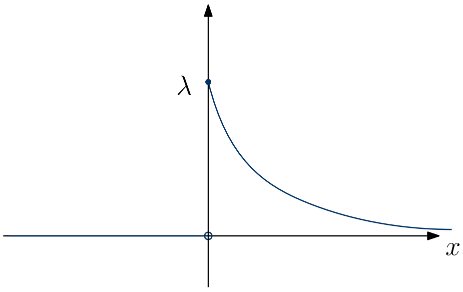

Figure 4.4.: Probability density function for $\mathscr{E}(\lambda)$.

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_0^{\infty} \lambda e^{-\lambda x}\mathrm{d}x = \left[-e^{-\lambda x}\right]_0^{\infty} = 1. \tag{4.21}$$

Typical application: The lifetime of radioactive isotopes is $\mathscr{E}(\lambda)$-distributed. An important property of the exponential distribution is the fact that it is “memoryless”. Indeed, one has

$$\mathbf{P}\big[(s+t,\infty)|(s,\infty)\big] = \mathbf{P}\big[(t,\infty)\big], \tag{4.22}$$

for $s, t > 0$. This is mimicking a similar property of the geometric distribution in the discrete case.

The normal distribution $\mathscr{N}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \mu \in \mathbb{R}, \sigma > 0. \tag{4.23}$$

Let us explain why this is indeed a probability density function:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x \overset{y=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\mathrm{d}y =: I. \tag{4.24}$$

Now we see that

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\mathrm{d}y\right)^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\mathrm{d}x\mathrm{d}y = \int_0^{2\pi}\int_0^{\infty} re^{-\frac{1}{2}r^2}\mathrm{d}r\mathrm{d}\varphi\\ &= 2\pi\left[-e^{-\frac{1}{2}r^2}\right]_0^{\infty} = 2\pi,\end{aligned} \tag{4.25}$$

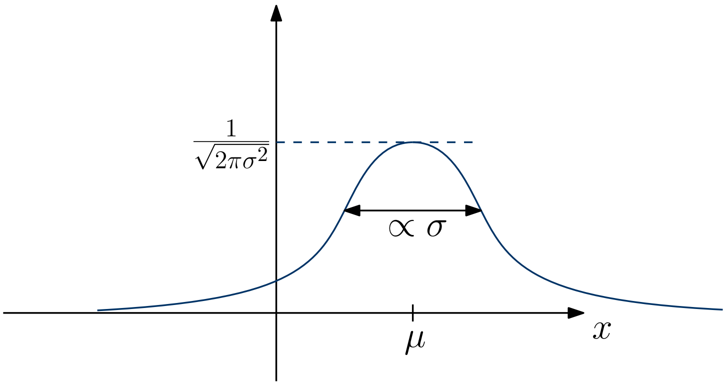

Figure 4.5.: Probability density function for $\mathcal{N}(\mu, \sigma^2)$.

therefore $I = 1$. We will use the notation

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \qquad \Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t \tag{4.26}$$

for the probability density and distribution function of a *standard normal distribution* $\mathcal{N}(0, 1)$. The function f in (4.23) or φ in (4.26) is often also called a *Gaussian (bell) curve.*

End of Lecture 7

4.4. Distribution functions, Lebesgue-Stieltjes measures

We have seen in Proposition 4.8 that every probability density function $f : \mathbb{R} \to [0, \infty)$ induces a unique probability measure $\mathbf{P}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. On the other hand, *not* every probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ has a probability density function. We now introduce a concept that will in fact characterize probability measures completely, namely that of distribution functions.

Definition 4.10. Let $\mathbf{P}$ be a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. We define

$$F_{\mathbf{P}}(x) = \mathbf{P}\big[(-\infty, x]\big], \qquad x \in \mathbb{R}. \tag{4.27}$$

The function $F_{\mathbf{P}}$ is called the *cumulative distribution function* of the probability measure $\mathbf{P}$.

The most important cases for us will be distribution functions for discrete or continuous distributions.

- Suppose that $\mathbf{P}$ is a discrete distribution on $(\Omega', \mathscr{P}(\Omega'))$ with $\Omega' \subseteq \mathbb{R}$ countable. By Example 4.7, $\Omega' \in \mathscr{B}(\mathbb{R})$. We can extend $\mathbf{P}$ to a distribution $\widetilde{\mathbf{P}}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ by setting

  $$\widetilde{\mathbf{P}}[A] = \mathbf{P}[A \cap \Omega'], \qquad A \in \mathscr{B}(\mathbb{R}). \tag{4.28}$$

  We typically still write $\widetilde{\mathbf{P}}$ as $\mathbf{P}$. In this case, we have

  $$F_{\mathbf{P}}(x) = \sum_{y \le x} \mathbf{P}[\{y\}] = \sum_{y \le x} p(y), \tag{4.29}$$

  where p is the probability mass function of $\mathbf{P}$.

- If $\mathbf{P}$ is a continuous probability measure characterized by the density f, then we have

$$F_{\mathbf{P}}(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t, \tag{4.30}$$

Since f is assumed to be piecewise continuous, we have (by the fundamental theorem of Calculus) the important identity

$$f(x) = F_{\mathbf{P}}'(x), \qquad \text{at all points of continuity of } f. \tag{4.31}$$

We also often abbreviate $F = F_{\mathbf{P}}$ if no confusion arises.

Example 4.11. (i) Consider the distribution $Bin(2, \frac{1}{2})$. In this case, we have

$$\begin{aligned} F_{Bin(2,\frac{1}{2})}(x) &= \begin{cases} 0, & x < 0, \\ \mathbf{P}[\{0\}], & x \in [0,1), \\ \mathbf{P}[\{0\}] + \mathbf{P}[\{1\}], & x \in [1,2), \\ 1, & x \geq 2, \end{cases} \\ &= \begin{cases} 0, & x < 0, \\ \frac{1}{4}, & x \in [0,1), \\ \frac{3}{4}, & x \in [1,2), \\ 1, & x \geq 2. \end{cases} \end{aligned} \tag{4.32}$$

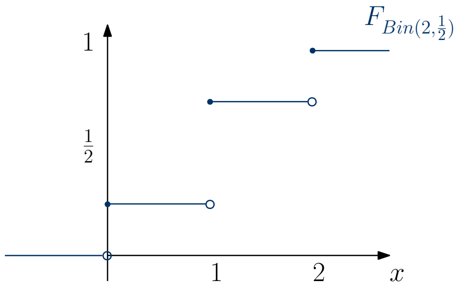

Figure 4.6.: Cumulative distribution function of $Bin(2, \frac{1}{2})$

(ii) Now consider $\mathcal{E}(\lambda)$, $\lambda > 0$. Then the cumulative distribution function is given by

$$F_{\mathcal{E}(\lambda)}(x) = \int_{-\infty}^{x} \lambda e^{-\lambda t} \mathbb{1}_{[0,\infty)}(t)\mathrm{d}t = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases} \tag{4.33}$$

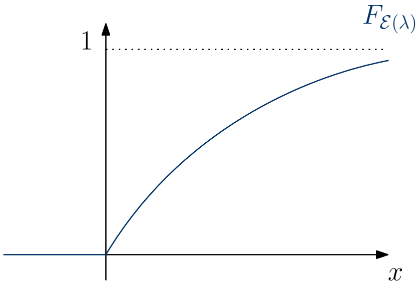

Figure 4.7.: Cumulative distribution function of $\mathcal{E}(\lambda)$

(iii) In the previous examples (i) and (ii) the distributions are either discrete or continuous. Let us stress that this is not a dichotomy. For instance, consider the probability measure

$$\mathbf{P}[A] = \frac{1}{2}\delta_0[A] + \frac{1}{2}\mathcal{N}(0,1)[A \cap (0,\infty)], \qquad A \in \mathcal{B}(\mathbb{R}), \tag{4.34}$$

where $\delta_0[A] = 1$ if $1 \in A$ and $\delta_0[A] = 0$ if $0 \notin A$. Here we have

$$F_{\mathbf{P}}(x) = \begin{cases} 0, & x < 0, \\ \Phi(x) = \frac{1}{2} + \frac{1}{2}\int_{-\infty}^{x} \varphi(t)\mathrm{d}t, & x \geq 0. \end{cases} \tag{4.35}$$

Note that $\mathbf{P}$ is neither continuous ($\mathbf{P}[\{0\}] = \frac{1}{2}$), nor discrete (it is not supported on a countable set).

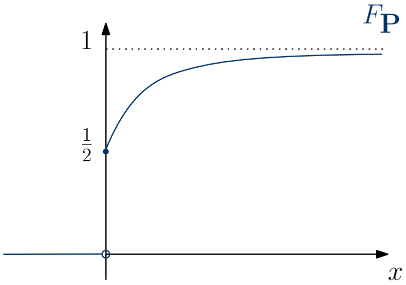

Figure 4.8.: Cumulative distribution function of $\mathbf{P}$ as defined in (4.34).

We will now collect some general facts about cumulative distribution functions.

Lemma 4.12. *Let* $\mathbf{P}$ *be a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*. Its cumulative distribution function* $F = F_{\mathbf{P}}$ *satisfies the following properties:*

(i) $F(x) \in [0,1]$ *for all* $x \in \mathbb{R}$*.*

(ii) F *is non-decreasing.*

(iii) F *is right continuous, i.e.*

$$\lim_{\varepsilon\downarrow 0} F(x+\varepsilon) = F(x). \tag{4.36}$$

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

Proof. Claim (i) follows since $F(x) = \mathbf{P}\big[(-\infty, x]\big] \in [0,1]$.

For claim (ii), we use the fact that for $x \le x'$, we have $(-\infty, x] \subseteq (-\infty, x']$ and so

$$F(x) = \mathbf{P}\big[(-\infty, x]\big] \le \mathbf{P}\big[(-\infty, x']\big] = F(x'). \tag{4.37}$$

For claim (iii) we apply Proposition 1.15, (vii), to the probability measure $\mathbf{P}$ and the sets $A_n = (-\infty, x_n]$, where $x_n \downarrow x$ for some $x \in \mathbb{R}$ (here we mean $x_n \ge x_{n+1}$ and $x_n \ge x$ for every $n \in \mathbb{N}$ and $x_n \to x$. Then

$$F(x_n) = \mathbf{P}\big[(-\infty, x_n]\big] \to \mathbf{P}\big[(-\infty, x]\big] = F(x), \qquad \text{as } n \to \infty, \tag{4.38}$$

since $\bigcap_{n=1}^{\infty}(-\infty, x_n] = (-\infty, x]$.

For (iv), we consider a sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ with $a_n \uparrow \infty$ as $n \to \infty$. Then

$$F(a_n) = \mathbf{P}\big[(-\infty, a_n]\big] \to \mathbf{P}[\mathbb{R}] = 1, \qquad \text{as } n \to \infty, \tag{4.39}$$

since $\mathbb{R} = \bigcup_{n=1}^{\infty}(-\infty, a_n]$ and $(-\infty, a_n] \subseteq (-\infty, a_{n+1}]$ for all $n \in \mathbb{N}$, upon using Proposition 1.15, (vi). The other claim follows similarly, again by using Proposition 1.15, (vii). □

Let us briefly explain the role of the discontinuity points of a cumulative distribution function F. If we look at (i), (iii) in Example 4.11, we see that a jump of the cumulative distribution function at a point $x \in \mathbb{R}$ correspond to the probability $\mathbf{P}[\{x\}]$. More formally:

Lemma 4.13. *Let* $\mathbf{P}$ *be a probability measure on* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ *with cumulative distribution function* $F = F_{\mathbf{P}}$. *Then, for every* $x \in \mathbb{R}$, *we have*

$$\mathbf{P}[\{x\}] = F(x) - F(x-), \tag{4.40}$$

where $F(x-)$ *(the left limit of* F *at* x*) is defined as*

$$F(x-) = \lim_{\varepsilon\downarrow 0} F(x-\varepsilon). \tag{4.41}$$

Proof. Note that since F is non-decreasing, the limit in (4.41) is well defined and equal to $\lim_{n\to\infty} F(x - \frac{1}{n})$. The claim (4.40) then follows from Proposition 1.15, (vii), the fact that

$$\{x\} = \bigcap_{n=1}^{\infty} (x - \tfrac{1}{n}, x] \tag{4.42}$$

and

$$F(x) - F(x - \tfrac{1}{n}) = \mathbf{P}\big[(x - \tfrac{1}{n}, x]\big], \tag{4.43}$$

by taking the limit $n \to \infty$. □

End of Lecture 8

In the remaining part of this section, we will show that functions fulfilling (i) – (iv) of Lemma 4.12 uniquely characterize *all* probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. To this end, we need some measure-theoretic preparations.

Definition 4.14. Let Ω be a non-empty set.

(i) A non-empty collection $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ is called π*-system* (or simply $\cap$-stable) if $A, B \in \mathscr{E}$ implies $A \cap B \in \mathscr{E}$.

(ii) A collection $\mathscr{D} \subseteq \mathscr{P}(\Omega)$ is called *Dynkin system* or λ*-system* if the following hold:

(D1) $\Omega \in \mathscr{D}$,

(D2) $A \in \mathscr{D} \Rightarrow A^c \in \mathscr{D}$,

(D3) If $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{D}$ with $A_i \cap A_j = \varnothing$ for $i \neq j$, then $\bigcup_{n=1}^\infty A_n \in \mathscr{D}$.

Note that for a Dynkin system $\mathscr{D}$ we have that

$$A, B \in \mathscr{D} \text{ with } A \subseteq B \qquad \Rightarrow \qquad B \setminus A = B \cap A^c = (B^c \cup A)^c \in \mathscr{D}, \tag{4.44}$$

since $B^c \cap A = \varnothing$, and thus (D1), (4.44) and (D3) are an equivalent definition of Dynkin systems. We also record the following observation.

Proposition 4.15. *A collection $\mathscr{D} \subseteq \mathscr{P}(\Omega)$ is a σ-algebra if and only if it is a $\cap$-stable Dynkin system.*

Proof. Suppose $\mathscr{D}$ is a σ-algebra, then $\mathscr{D}$ is clearly a Dynkin system (the requirements (S1), (S2) and (S3) include (D1), (D2) and (D3) and also $\Omega \in \mathscr{D}$) and by Proposition 1.11 it is also $\cap$-stable.

For the other direction, (S1) and (S2) (coinciding with (D1), (D2)) are guaranteed by definition, so we are left proving (S3). Let $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{D}$ and consider $B_1 = A_1$ and $B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right) = A_n \cap \bigcap_{i=1}^{n-1} A_i^c$ for $n \geq 2$. We first notice that by (D2) and the fact that $\mathscr{D}$ is $\cap$-stable, we have $B_n \in \mathscr{D}$ for every $n \in \mathbb{N}$. Moreover, $(B_n)_{n\in\mathbb{N}}$ consists of pairwise disjoint members. By (D3) we thus see that $\bigcup_{n=1}^\infty A_n = \bigcup_{n=1}^\infty B_n \in \mathscr{D}$. □

Similarly as generated σ-algebras, we can introduce the notion of generated Dynkin systems.

Definition 4.16. Take again $\varnothing \neq \mathscr{E} \subseteq \mathscr{P}(\Omega)$. We define the *Dynkin system generated by* $\mathscr{E}$ as

$$\delta(\mathscr{E}) = \bigcap_{\substack{\mathscr{E} \subseteq \mathscr{D} \\ \mathscr{D} \text{Dynkin system}}} \mathscr{D} \subseteq \mathscr{P}(\Omega). \tag{4.45}$$

This is again a Dynkin system (the smallest Dynkin system containing $\mathscr{E}$), since any intersection of Dynkin systems is again a Dynkin system. Note that since every σ-algebra is a Dynkin system by Proposition 4.15, we see that

$$\delta(\mathscr{E}) \subseteq \sigma(\mathscr{E}) \qquad \text{for all } \varnothing \neq \mathscr{E} \subseteq \mathscr{P}(\Omega). \tag{4.46}$$

Dynkin systems are an indispensable tool for uniqueness proofs, and we will see applications of the following deep result frequently.

Theorem 4.17 (Dynkin's π-λ theorem)**.** *Let $\mathscr{D}$ be a Dynkin system and $\mathscr{E}$ a π-system (i.e. $\cap$-stable) on Ω. Suppose that $\mathscr{E} \subseteq \mathscr{D}$. Then*

$$\sigma(\mathscr{E}) \subseteq \mathscr{D}. \tag{4.47}$$

Before proving Theorem 4.17, we give its major application which pertains to uniqueness of probability.

Corollary 4.18. *Let $\mathscr{E} \subseteq \mathscr{P}(\Omega)$ be $\cap$-stable and $\mathscr{F} = \sigma(\mathscr{E})$. If $\mathbf{P}$ and $\mathbf{Q}$ are probability measures fulfilling*

$$\mathbf{P}[A] = \mathbf{Q}[A] \qquad \textit{for all } A \in \mathscr{E}, \tag{4.48}$$

we have $\mathbf{P} = \mathbf{Q}$ on $\mathscr{F}$.

Proof. Consider the set

$$\mathscr{D} = \{A \in \mathscr{F} \ : \ \mathbf{P}[A] = \mathbf{Q}[A]\}. \tag{4.49}$$

We now prove that $\mathscr{D}$ is a Dynkin system. Indeed, we have $\Omega \in \mathscr{D}$ since $\mathbf{P}[\Omega] = \mathbf{Q}[\Omega] = 1$, so (D1) holds. Moreover, we see that if $A \in \mathscr{D}$, then

$$\begin{aligned} \mathbf{P}[A^c] &= 1 - \mathbf{P}[A] \\ &\overset{(4.49)}{=} 1 - \mathbf{Q}[A] = \mathbf{Q}[A^c], \end{aligned} \tag{4.50}$$

thus $\mathscr{D}$ fulfills (D2). As for (D3), let $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{D}$ be made of mutually disjoint members with $A = \bigcup_{n=1}^{\infty} A_n$. We then have

$$\mathbf{P}[A] = \sum_{n=1}^{\infty} \mathbf{P}[A_n] \overset{(4.49)}{=} \sum_{n=1}^{\infty} \mathbf{Q}[A_n] = \mathbf{Q}[A], \tag{4.51}$$

thus (D2) holds. Since $\mathscr{D}$ is a Dynkin system containing $\mathscr{E}$ and $\mathscr{E}$ is $\cap$-stable, we can use Dynkin's π-λ theorem (Theorem 4.17) so

$$\mathscr{F} = \sigma(\mathscr{E}) \subseteq \mathscr{D} \subseteq \mathscr{F}. \tag{4.52}$$

Hence $\mathscr{D} = \mathscr{F}$, which can be translated to

$$\mathbf{P}[A] = \mathbf{Q}[A] \qquad \text{for all } A \in \mathscr{F}, \tag{4.53}$$

which completes the proof. □

We now give the proof of Theorem 4.17, see also [Geo12, Lemma 1.13].

Proof of Theorem 4.17. Recall the definition (4.45). We need to show that $\sigma(\mathscr{E}) \subseteq \mathscr{D}$, but since $\delta(\mathscr{E}) \subseteq \mathscr{D}$ by definition, it is enough to show that $\sigma(\mathscr{E}) \subseteq \delta(\mathscr{E})$. By the definition of $\sigma(\mathscr{E})$, see (4.7), this is done once we show that

$$\delta(\mathscr{E}) \text{ is a } \sigma\text{-algebra}. \tag{4.54}$$

Now, by Proposition 4.15, (4.54) follows from the claim

$$\delta(\mathscr{E}) \text{ is } \cap\text{-stable.} \tag{4.55}$$

The remainder of the proof is thus to prove the claim (4.55), which is done in two steps.

Step 1: We show that $A \in \delta(\mathscr{E})$, $B \in \mathscr{E} \Rightarrow A \cap B \in \delta(\mathscr{E})$. To that end, define for fixed $B \in \mathscr{E}$ the set

$$\mathscr{D}_B = \{A \subseteq \Omega \, : \, A \cap B \in \delta(\mathscr{E})\}. \tag{4.56}$$

We now show that $\mathscr{D}_B$ is a Dynkin system. Indeed:

(D1) $\Omega \in \mathscr{D}_B$ is immediate.

(D2) Suppose $E \in \mathscr{D}_B$, then

$$E^c \cap B = \underbrace{B}_{\in \mathscr{E} \subseteq \delta(\mathscr{E})} \setminus \underbrace{(E \cap B)}_{\in \delta(\mathscr{E})} \in \delta(\mathscr{E}), \tag{4.57}$$

and it follows that $E^c \in \mathscr{D}_B$.

(D3) Let $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{D}_B$ be pairwise disjoint. Then

$$\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B = \bigcup_{n=1}^{\infty} \underbrace{(A_n \cap B)}_{\in \delta(\mathscr{E}),\text{ pairwise disjoint}} \in \delta(\mathscr{E}). \tag{4.58}$$

Also, it is clear that $\mathscr{E} \subseteq \mathscr{D}_B$: Indeed, if $A \in \mathscr{E}$ then $A \cap B \in \mathscr{E} \subseteq \delta(\mathscr{E})$ since $\mathscr{E}$ is assumed to be $\cap$-stable. By definition of $\delta(\mathscr{E})$, we have therefore that $\delta(\mathscr{E}) \subseteq \mathscr{D}_B$, concluding the proof of Step 1.

Step 2: We show that $A, B \in \delta(\mathscr{E}) \Rightarrow A \cap B \in \delta(\mathscr{E})$ (i.e. (4.55)). We define

$$\mathscr{D}_A = \{B \subseteq \Omega \, : \, A \cap B \in \delta(\mathscr{E})\}. \tag{4.59}$$

This definition coincides with (4.56) (but we allow $A \in \delta(\mathscr{E})$). As in Step 1, one can see that $\mathscr{D}_A$ is a Dynkin-system. Now, by Step 1, we note that

$$B \in \mathscr{D}_A \text{ for every } B \in \mathscr{E}. \tag{4.60}$$

It follows that $\mathscr{E} \subseteq \mathscr{D}_A$, and thus $\delta(\mathscr{E}) \subseteq \mathscr{D}_A$. Written out, this is exactly the claim (4.55). □

We are now in the position to prove the announced converse of Lemma 4.12.

Theorem 4.19 (Lebesgue-Stieltjes measures). *Let $F : \mathbb{R} \to \mathbb{R}$ satisfying the properties (i) – (iv) of Lemma 4.12. Then there exists a probability measure $\mathbf{P}$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $F = F_{\mathbf{P}}$, and $\mathbf{P}$ is uniquely determined by F.*

Proof. We define the *pseudoinverse* $F^{-1} : (0,1) \to \mathbb{R}$ as follows:

$$F^{-1}(\omega) = \sup\{y \in \mathbb{R} \,:\, F(y) < \omega\}. \tag{4.61}$$

Note that

$$\{\omega \in (0,1) \,:\, F^{-1}(\omega) \leq x\} = \{\omega \in (0,1) \,:\, \omega \leq F(x)\}, \qquad x \in \mathbb{R}. \tag{4.62}$$

Indeed, if $\omega \leq F(x)$, then $x \notin \{y \in \mathbb{R} \,:\, F(y) < \omega\}$, which implies $x \geq F^{-1}(\omega)$.

On the other hand, if $\omega \in (0,1)$ with $F(x) < \omega$, since F is right continuous, there is $\varepsilon > 0$ with $F(x+\varepsilon) < \omega$. Therefore $F^{-1}(\omega) \geq x + \varepsilon > x$. This means that $F(x) < \omega$ implies that $F^{-1}(\omega) > x$. Consequently, $x \geq F^{-1}(\omega)$ implies $\omega \leq F(x)$.

We then equip the space $((0,1), \mathscr{B}((0,1)))$ with the uniform distribution $\mathbf{Q} = \mathscr{U}((0,1))$,[2] and set $\mathbf{P}[A] = \mathbf{Q}[\{\omega \in (0,1) \,:\, F^{-1}(\omega) \in A\}]$ for $A \in \mathscr{B}(\mathbb{R})$. It is left as an exercise to show that $\{\omega \in (0,1) \,:\, F^{-1}(\omega) \in A\} \in \mathscr{B}(\mathbb{R})$ and that $\mathbf{P}$ is indeed a probability measure. We now claim that $\mathbf{P}$ fulfills $F_{\mathbf{P}} = F$. Indeed:

$$F_{\mathbf{P}}(x) = \mathbf{Q}\big[\{\omega \in (0,1) \,:\, F^{-1}(\omega) \leq x\}\big] \overset{(4.62)}{=} \mathbf{Q}\big[(0, F(x)] \cap (0,1)\big] = F(x). \tag{4.63}$$

This proves the existence part. For the uniqueness, note that the set

$$\mathscr{E} = \{(-\infty, b] \,:\, b \in \mathbb{R}\} \subseteq \mathscr{B}(\mathbb{R}) \tag{4.64}$$

is $\cap$-stable and fulfills $\mathscr{B}(\mathbb{R}) = \sigma(\mathscr{E})$. Suppose that $\mathbf{P}$ and $\widetilde{\mathbf{P}}$ are two probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with $F_{\widetilde{\mathbf{P}}} = F = F_{\mathbf{P}}$, then we find that for any $A = (-\infty, b] \in \mathscr{E}$:

$$\mathbf{P}[A] = F(b) = \widetilde{\mathbf{P}}[A], \tag{4.65}$$

and by Corollary (4.18), we see that $\mathbf{P} = \widetilde{\mathbf{P}}$. □

End of Lecture 9

We close this section with a number of comments and a reflection on general probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Remark 4.20. (i) The proof above is constructive: For any cumulative distribution function $F : \mathbb{R} \to \mathbb{R}$ (i.e. a function satisfying the properties (i) – (iv) of Lemma 4.12), the formula (4.61) gives us a way to simulate the distribution associated with F, if one is able to simulate a $\mathscr{U}((0,1))$-distribution.

(ii) In line with the first remark, note the following:

- In the previous Chapter 3, we gave a complete characterization and construction of discrete probability measures in terms of the probability mass function p (see Proposition 3.4).

[2] Here, $\mathscr{U}((0,1))$ stands for the uniform distribution on the open interval $(0,1)$, viewed as a probability measure on $\big((0,1), \mathscr{B}((0,1))\big)$. Since continuous distributions give zero mass to points, this is essentially the same distribution as the uniform distribution $\mathscr{U}([0,1])$ on $[0,1]$, viewed as a probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

- Note that *any* probability measure **P** on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ can be constructed using the previous theorem, given the existence of $\mathscr{U}((0,1))$. This means in particular that in Proposition 4.8, the *only* part of the construction we do not explicitly give is the existence of $\mathscr{U}((0,1))$, which is equivalent to the existence of the *Lebesgue-measure* λ_1 on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

(iii) Note that the cumulative distribution function of a continuous distribution is continuous, but the converse is not true. Instead, it is a standard fact from analysis that the set of functions $F : \mathbb{R} \to [0,1]$ with $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ for a $f \geq 0$ with $\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1$[3] is exactly the class

$$\begin{aligned} \mathscr{F}_{\text{abs}} = \{F : \mathbb{R} \to [0,1] \text{ absolutely continuous, right-continuous, increasing,} \\ \text{and fulfills } \lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1\}. \end{aligned} \tag{4.66}$$

Here, $g : \mathbb{R} \to \mathbb{R}$ is called absolutely continuous if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for each sequence of pairwise disjoint intervals $((x_j, y_j))_{j=1}^{N}$, with $x_j < y_j$, we have that $\sum_{j=1}^{N}(y_j - x_j) < \delta$ implies $\sum_{j=1}^{N} |g(y_j) - g(x_j)| < \varepsilon$. Note that not every continuous function is absolutely continuous. Continuous but not absolutely continuous cumulative distribution functions give rise to *singular continuous distributions.*

[3] In fact, the function f need not be piecewise continuous, so our definition of (absolutely) continuous distributions is yet (very slightly) more restrictive!

5. Random variables

(Reference: [GS01, Section 2.1], or [Geo12, Sections 1.3])

Random variables are an extremely important tool in probability theory, allowing for the manipulation of different sources of randomness. We will motivate this with a simple example and then move on to a more general framework.

5.1. Definition of random variables

Let us consider rolling two dice. We are interested in the sum of the outcomes of the two dice. If the dice are fair, the probability space modelling this problem is given by (1.31), namely

$$\begin{aligned} &\Omega = \{1,2,3,4,5,6\}^2 = \{(1,1),(1,2),\dots,(6,6)\},\\ &\mathscr{F} = \mathscr{P}(\Omega),\\ &\mathbf{P} = \mathscr{U}(\Omega), \qquad \text{i.e. } \mathbf{P}[A] = \frac{|A|}{|\Omega|} = \frac{|A|}{36}, \qquad \text{for all } A \in \mathscr{P}(\Omega). \end{aligned} \tag{5.1}$$

The sum of the outcomes is given by the number

$$S(\omega) = \omega_1 + \omega_2, \tag{5.2}$$

and it can attain values in $\{2,3,\dots,12\}$. We can calculate the probability that the sum of the outcomes is 3 as follows:

$$\mathbf{P}[\{\omega \in \Omega \,:\, S(\omega) = 3\}] = \mathbf{P}[\{(1,2),(2,1)\}] = \frac{2}{36} = \frac{1}{18}. \tag{5.3}$$

More generally, if we ask for the probability that the sum of the outcomes is $k \in \{2,\dots,12\}$:

$$\mathbf{P}[\{\omega \in \Omega \,:\, S(\omega) = k\}] = \mathbf{P}\big[S^{-1}(\{k\})\big]. \tag{5.4}$$

We observe that for the function $S : \Omega \to \mathbb{R}$ and every $A \in \mathscr{B}(\mathbb{R})$, the set

$$\{S \in A\} = \{\omega \in \Omega \,:\, S(\omega) \in A\} = S^{-1}(A) \in \mathscr{P}(\Omega) \tag{5.5}$$

is an event. The function S is called a random variable. Let us give a general definition.

Definition 5.1. (i) Let $(\Omega, \mathscr{F})$, $(S, \mathscr{S})$ be two measurable spaces (i.e. $\mathscr{F}$ is a σ-algebra on Ω and $\mathscr{S}$ is a σ-algebra on S). A function

$$f : \Omega \to S \tag{5.6}$$

is called *$\mathscr{F}$-$\mathscr{S}$-measurable* if

$$f^{-1}(A) \in \mathscr{F} \qquad \text{for all } A \in \mathscr{S}. \tag{5.7}$$

In this case we also write $f : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$.

(ii) Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $(S, \mathscr{S})$ a measurable space. A $\mathscr{F}$-$\mathscr{S}$-measurable function $X : \Omega \to S$ is called a *S-valued random variable*. We say that $X : \Omega \to \mathbb{R}$ is a real random variable if it is $\mathscr{F}$-$\mathscr{B}(\mathbb{R})$-measurable.

Remark 5.2. (i) It can be argued that a function $X : \Omega \to \mathbb{R}$ is a real random variable if and only if

$$\{X \leq a\} = \{\omega \in \Omega\,;\, X(\omega) \leq a\} \in \mathscr{F} \qquad \text{for all } a \in \mathbb{R}. \tag{5.8}$$

(ii) If Ω is countable and $\mathscr{F} = \mathscr{P}(\Omega)$, then every function $X : \Omega \to \mathbb{R}$ is a random variable.

(iii) We can also view X as a map from Ω to $\Omega_X = \{X(\omega) : \omega \in \Omega\}$. This in particular useful, if Ω_X is itself countable. In this case, we say that X is a *discrete random variable.*

(iv) Note that the definition of a random variable does not depend on the probability measure, but only on the σ-algebra $\mathscr{F}$ used.

5.2. Law and cumulative distribution of a real random variable

Definition 5.3. Let $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a real random variable. The *law of X under* $\mathbf{P}$ is the probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ defined by

$$\mathbf{P}_X[B] = \mathbf{P}\big[X^{-1}(B)\big], \qquad B \in \mathscr{B}(\mathbb{R}). \tag{5.9}$$

More generally, we can define (mutatis mutandis, via $\mathbf{P}_X[B] = \mathbf{P}\big[X^{-1}(B)\big], B \in \mathscr{S}$) the law on $(S, \mathscr{S})$ under $\mathbf{P}$ of an S-valued random variable $X : (\Omega, \mathscr{F}) \to (S, \mathscr{S})$.

Remark 5.4. (i) It is easy to see that $\mathbf{P}_X$ is indeed a probability measure, so $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbf{P}_X)$ is a probability space.

(ii) In the case where Ω_X is countable, we can also consider $\mathbf{P}_X$ as a probability measure on $(\Omega_X, \mathscr{P}(\Omega_X))$. In this case, (5.9) becomes

$$\mathbf{P}_X[B] = \mathbf{P}\big[X^{-1}(B)\big], \qquad B \in \mathscr{P}(\Omega_X). \tag{5.10}$$

(iii) We write as abbreviation $X \sim \mathbf{P}_X$. For instance, $X \sim Bin(n, p)$ means that $\mathbf{P}_X = Bin(n, p)$.

(iv) If the law $\mathbf{P}_X$ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is a continuous distribution, i.e. it is defined in terms of a probability density f_X as in (4.13), we say that X is a *continuous random variable.*

(v) We write F_X as a shorthand notation for $F_{\mathbf{P}_X}$ (the cumulative distribution function of the law of X under $\mathbf{P}$).

End of Lecture 10

Definition 5.5. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $X, Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ two random variables. We say that X and Y are *equal in distribution* or *identically distributed* if

$$\mathbf{P}_X = \mathbf{P}_Y \qquad (\Leftrightarrow F_X(x) = F_Y(x), \text{ for all } x \in \mathbb{R}). \tag{5.11}$$

This is denoted as $X \overset{d}{=} Y$.

Example 5.6. (i) Consider again throwing two dice. We use the probability space (5.1). The result of the first and second die are given by the random variables

$$\begin{aligned} X &: \{1, ..., 6\}^2 \to \{1, ..., 6\}, \qquad X(\omega_1, \omega_2) = \omega_1, \\ Y &: \{1, ..., 6\}^2 \to \{1, ..., 6\}, \qquad Y(\omega_1, \omega_2) = \omega_2. \end{aligned} \tag{5.12}$$

We see that $X \sim \mathcal{U}(\{1, ..., 6\})$ and $Y \sim \mathcal{U}(\{1, ..., 6\})$, so $X \overset{d}{=} Y$. Note that of course $X \neq Y$, since for instance $X(1,2) = 1 \neq 2 = Y(1,2)$.

(ii) Let X be any continuous random variable and f_X the probability density of its law. Assume that f_X is an even function, i.e.

$$f_X(x) = f_X(-x), \qquad \text{for all } x \in \mathbb{R}. \tag{5.13}$$

Then $X \overset{d}{=} -X$. Indeed, we have

$$\begin{aligned} F_{-X}(x) = \mathbf{P}[-X \le x] &= \mathbf{P}[X \ge -x] = 1 - \mathbf{P}[X < -x] \\ &= 1 - \int_{-\infty}^{-x} f_X(t)\mathrm{d}t \\ &= 1 + \int_{\infty}^{x} f_X(-t)\mathrm{d}t \\ &= 1 - \int_{x}^{\infty} f_X(t)\mathrm{d}t \\ &= \mathbf{P}[X \le x] = F_X(x). \end{aligned} \tag{5.14}$$

Here we repeatedly used the results of Lemma 4.9. A more concrete example: If $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma > 0$, we have that $-X \sim \mathcal{N}(0, \sigma^2)$ as well.

The example above already gives a hint that random variables are a useful tool for algebraic manipulations. We will see more of this in the next section.

5.3. Transformation of random variables

Example 5.7. We measure the temperature of a liquid in ° C and want to transform it into ° F.

$$\begin{aligned} {}^\circ\text{C} &: \text{random variable } X, \\ {}^\circ\text{F} &: \text{random variable } Y. \end{aligned}$$

We can use the known formula

$$Y = \frac{9}{5} \cdot X + 32, \qquad \text{more generally } Y = a \cdot X + b, \tag{5.15}$$

for $a \neq 0$, $b \in \mathbb{R}$. We assume that X is a continuous random variable with probability density (distribution) function f_X (F_X), for instance $X \sim \mathcal{N}(\mu, \sigma^2)$. What is the distribution of Y, i.e. how do f_Y or F_Y look like? For simplicity, we also assume $a > 0$.

$$\begin{aligned} & F_Y(y) = \mathbf{P}_Y\big[(-\infty, y]\big] = \mathbf{P}\big[Y \leq y\big] = \mathbf{P}\big[aX + b \leq y\big] = \mathbf{P}\left[X \leq \tfrac{y-b}{a}\right] = F_X\left(\tfrac{y-b}{a}\right) \\ \Rightarrow \quad & f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X\left(\tfrac{y-b}{a}\right) = \frac{1}{a} \cdot f_X\left(\tfrac{y-b}{a}\right). \end{aligned} \tag{5.16}$$

If $Y = a \cdot X + b$ with general $a \neq 0$, we have

$$f_Y(y) = \frac{1}{|a|} \cdot f_X\left(\tfrac{y-b}{a}\right). \tag{5.17}$$

In the special case where $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$\begin{aligned} & f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ \Rightarrow \quad & f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma|a|} e^{-\frac{1}{2\sigma^2 a^2}(y-b-a\mu)^2} \\ \Rightarrow \quad & Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2). \end{aligned} \tag{5.18}$$

Formula (5.17) is the linear transformation rule for continuous random variables. We now show a general transformation rule for continuous random variables.

Theorem 5.8. *Let X be a continuous random variable and f_X the probability density function of its law. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing or strictly decreasing and differentiable.*[1] *Then*

$$Y = g(X) \tag{5.19}$$

is also a continuous random variable and has density

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right|, & y = g(x) \textit{ with } f_X(x) > 0, \\ 0, & \textit{else.} \end{cases} \tag{5.20}$$

[1] To be very precise, we also need to make sure that $Y = g(X)$ is still a random variable (i.e. that it is $\mathscr{F} - \mathscr{B}(\mathbb{R})$-measurable. This follows from the fact that g is differentiable and thus $\mathscr{B}(\mathbb{R}) - \mathscr{B}(\mathbb{R})$-measurable (in fact continuity is sufficient) and the simple fact that the composition of measurable functions is measurable.

Proof. Let g be strictly increasing. Then, for $[a,b) \subseteq \{g(x)\,;\, f_X(x) > 0\}$, we have

$$\begin{aligned}\mathbf{P}_Y\big[[a,b)\big] &= \mathbf{P}\big[g(X) \in [a,b)\big] = \mathbf{P}\big[X \in [g^{-1}(a), g^{-1}(b))\big] \\ &= \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)\mathrm{d}x = \int_a^b \underbrace{f_X\left(g^{-1}(y)\right) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)}_{=f_Y(y)} \mathrm{d}y.\end{aligned} \tag{5.21}$$

The proof for g strictly decreasing is similar. □

Example 5.9. Let $X \sim \mathcal{U}([0,1])$ and consider $Y = \exp(X) = e^X$. The function exp satisfies the requirements of the previous theorem, and its inverse is log. Moreover, $f_X(x) \neq 0$ if and only if $x \in [0,1]$. We have

$$f_Y(y) = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}y}\log(y) = \frac{1}{y}, & y \in [1,e], \\ 0, & y \notin [1,e]. \end{cases} \tag{5.22}$$

End of Lecture 11

In the next example, we use the same method as in Theorem 5.8 to introduce the χ^2 distribution (with one degree of freedom).

Example 5.10. Let $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. We want to calculate $f_Y(y)$. Unfortunately the function $g : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ is not strictly increasing / decreasing, but we can still use the same idea as in the proof of the transformation rule. Indeed, we see that $g^{-1}(y) = \sqrt{y}$ and $\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = \frac{1}{2\sqrt{y}}$ for $y > 0$. Then, for $0 \leq a < b < \infty$, we find

$$\begin{aligned}\mathbf{P}\big[Y \in [a,b)\big] &= \mathbf{P}\big[X \in [\sqrt{a}, \sqrt{b})\big] + \mathbf{P}\big[X \in (-\sqrt{b}, -\sqrt{a}]\big] \\ &= 2\mathbf{P}\big[X \in [\sqrt{a}, \sqrt{b})\big] \\ &= 2\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{2\sqrt{y}} \mathrm{d}y.\end{aligned} \tag{5.23}$$

We used the symmetry of $X \sim \mathcal{N}(0,1)$ and the fact that $\mathbf{P}_X$ does not give mass to points. It follows that

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} \mathbb{1}_{[0,\infty)}(y). \tag{5.24}$$

We say that a random variable Y with a law given by the density is χ^2*-distributed (with one degree of freedom).*

6. Expectation, variance and higher moments of random variables

(Reference: [GS01, Sections 3.3, 4.3], or [Geo12, Sections 4,1, 4.3])

6.1. Expectation

We now introduce the notion of the *expectation* or *expected value* of a real random variable. The idea is to somehow quantify the "typical" or "average" value of a random variable X. Let us motivate the definition with an example.

Example 6.1. We consider a game where we throw a die once, and get the following rewards:

- \$ 1 if the die shows 1 or 2,
- \$ 2 if the die shows 3 or 4,
- \$ 4 if the die shows 5, and
- \$ 8 if the die shows 6.

What would be a "fair price" for playing this game? We would like to stakes to be such that we do not lose money on average by playing. To describe the (random) return in one round, we can define the random variable $X : \{1, 2, 3, 4, 5, 6\} \to \{1, 2, 4, 8\}$ on the Laplace probability space $(\{1, 2, 3, 4, 5, 6\}, \mathscr{P}(\{1, 2, 3, 4, 5, 6\}, \mathscr{U}(\{1, 2, 3, 4, 5, 6\}))$, by

$$X(\omega) = \mathbb{1}_{\{1,2\}}(\omega) + 2 \cdot \mathbb{1}_{\{3,4\}}(\omega) + 4 \cdot \mathbb{1}_{\{5\}}(\omega) + 8 \cdot \mathbb{1}_{\{6\}}(\omega). \tag{6.1}$$

Assume that we play $n \in \mathbb{N}$ times, and n_1, n_2, ..., n_6 denotes the number of 1, 2, ..., 6 that show up in n rounds. Our return (in \$) after n steps is

$$1 \cdot n_1 + 1 \cdot n_2 + 2 \cdot n_3 + 2 \cdot n_4 + 4 \cdot n_5 + 8 \cdot n_6. \tag{6.2}$$

So, the average return in one round is

$$1 \cdot \frac{n_1}{n} + 1 \cdot \frac{n_2}{n} + 2 \cdot \frac{n_3}{n} + 2 \cdot \frac{n_4}{n} + 4 \cdot \frac{n_5}{n} + 8 \cdot \frac{n_6}{n}. \tag{6.3}$$

The idea is now that for large n, the relative fractions $\frac{n_i}{n}$ should be close to $\frac{1}{6}$. This gives us the value

$$\begin{aligned} \mathbf{E}[X] &= 1 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} \\ &= 1 \cdot \mathbf{P}[X = 1] + 2 \cdot \mathbf{P}[X = 2] + 4 \cdot \mathbf{P}[X = 4] + 8 \cdot \mathbf{P}[X = 6] \\ &= 18 \cdot \frac{1}{6} = 3. \end{aligned} \tag{6.4}$$

This somehow suggests that the “fair” price to play the game is \$ 3.

This example motivates the definition of the expectation.

Definition 6.2. (i) Let X be a discrete real random variable with values in $\Omega_X (\subseteq \mathbb{R})$ and let p_X be the probability mass function of its law $\mathbf{P}_X$. We define the *expectation of* X as

$$\mathbf{E}\big[X\big] = \sum_{\omega \in \Omega_X} \omega \cdot p_X(\omega), \tag{6.5}$$

if $\sum_{\omega\in\Omega_X} |\omega| p_X(\omega) < \infty$.

(ii) Let X be a continuous real random variable and let f_X be the probability density function of its law $\mathbf{P}_X$. We define the *expectation of* X as

$$\mathbf{E}\big[X\big] = \int_{-\infty}^{\infty} x \cdot f_X(x) \mathrm{d}x, \tag{6.6}$$

if $\int_{-\infty}^{\infty} |x| f_X(x) \mathrm{d}x < \infty$.

For real random variables X that are neither discrete nor continuous, we typically cannot easily define the expectation by a formula as above (since $\mathbf{P}_X$ may neither admit a probability mass function nor a probability density function). To define the expectation of any real random variable $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and $n \in \mathbb{N}$, we define the $\frac{1}{n}$-discretization

$$X_{(n)} : \Omega \to \mathbb{R}, \qquad \omega \mapsto X_{(n)}(\omega) = \frac{\lfloor nX(\omega) \rfloor}{n} = \sum_{k \in \mathbb{Z}} \frac{k}{n} \mathbb{1}_{\left\{\frac{k}{n} \le X(\omega) < \frac{k+1}{n}\right\}}. \tag{6.7}$$

$X_{(n)}$ is a discrete random variable with $\Omega_{X_{(n)}} \subseteq \frac{1}{n}\mathbb{Z}$. We have the following:

Lemma 6.3. *Let* $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ *be a real random variable.*

(i) For any $n \geq 1$, *we have* $X_{(n)} \leq X \leq X_{(n)} + \frac{1}{n}$.

(ii) If $X_{(n)}$ *has an expectation for some* $n \geq 1$, *it has an expectation for every* $n \in \mathbb{N}$, *and* $\big(\mathbf{E}\big[X_{(n)}\big]\big)_{n\in\mathbb{N}}$ *is a Cauchy sequence.*

Proof. Item (i) follows immediately from the definition. We now turn to (ii). For $m, n \geq 1$, we have both

$$X_{(m)} < X_{(n)} + \frac{1}{n}, \qquad \text{and} \qquad X_{(n)} < X_{(m)} + \frac{1}{m}, \tag{6.8}$$

which implies that

$$|X_{(n)}| < |X_{(m)}| + \max\left\{\frac{1}{m}, \frac{1}{n}\right\}. \tag{6.9}$$

It now follows from the definition that if $|X_{(m)}|$ has finite expectation, also $|X_{(n)}|$ has a finite expectation. Moreover, we see that for $n, m \in \mathbb{N}$,

$$\mathbf{E}\big[X_{(m)}\big] \leq \mathbf{E}\big[X_{(n)}\big] + \frac{1}{n}. \tag{6.10}$$

The last inequality holds if we exchange the roles of $m, n \in \mathbb{N}$, so that

$$\left|\mathbf{E}\left[X_{(m)}\right] - \mathbf{E}\left[X_{(n)}\right]\right| \leq \max\left\{\frac{1}{m}, \frac{1}{n}\right\}, \tag{6.11}$$

from which it is clear that $\left(\mathbf{E}\left[X_{(n)}\right]\right)_{n\in\mathbb{N}}$ is a Cauchy sequence. □

Definition 6.4. Let $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a real random variable. If $X_{(n)}$ has an expectation for some (and hence all) n, we define the *expectation of X* as

$$\mathbf{E}\left[X\right] = \lim_{n\to\infty} \mathbf{E}\left[X_{(n)}\right]. \tag{6.12}$$

Remark 6.5. (i) The expectation only depends on the law $\mathbf{P}_X$ of X. In other words: If $X \overset{d}{=} Y$ and the expectation of X exists, then the expectation of Y exists as well and $\mathbf{E}[X] = \mathbf{E}[Y]$.

(ii) One can show that Definition 6.4 contains Definition 6.2, (ii) as a special case; see, e.g., [Geo12, Theorem 4.12].

(iii) We will see in the exercises that for $X \geq 0$ either discrete or continuous, we have

$$\mathbf{E}\left[X\right] = \int_0^\infty (1 - F_X(x))\mathrm{d}x, \tag{6.13}$$

with F_X the cumulative distribution function of the law of X. Defining the *positive* and *negative part* $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$ (such that $X = X^+ - X^-$), we can also obtain the general formula for the expectation

$$\mathbf{E}\left[X\right] = \int_0^\infty (1 - F_{X^+}(x))\mathrm{d}x - \int_0^\infty (1 - F_{X^-}(x))\mathrm{d}x, \tag{6.14}$$

if both of the integrals are finite.

It is sometimes expedient to also define the expectation as $+\infty$ or $-\infty$: for instance if $\Omega_X \subseteq [0, \infty)$ in (6.5), or if $f_X(x) = 0$ whenever $x < 0$ in (6.6), the value of the sum (resp. integral) always exists in $[0, \infty]$. Let us give a couple of examples.

Example 6.6. (i) Let $X = c \in \mathbb{R}$. Then

$$\mathbf{E}\left[X\right] = c, \tag{6.15}$$

since X is a discrete random variable and $\Omega_X = \{c\}$, $\mathbf{P}[X = c] = 1$.

(ii) Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $A \in \mathscr{F}$. Then $\mathbb{1}_A$, defined by

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases} \tag{6.16}$$

is a random variable and

$$\mathbf{E}\left[\mathbb{1}_A\right] = \mathbf{P}[A]. \tag{6.17}$$

Indeed, $\mathbb{1}_A^{-1}(B) \in \{\emptyset, A, A^c, \Omega\}$ for every $B \in \mathscr{B}(\mathbb{R})$, and $\Omega_{\mathbb{1}_A} = \{0,1\}$. Therefore, we have

$$\mathbf{E}\big[\mathbb{1}_A\big] = 0 \cdot \mathbf{P}[\mathbb{1}_A = 0] + 1 \cdot \underbrace{\mathbf{P}[\mathbb{1}_A = 1]}_{=\mathbf{P}[A]}. \tag{6.18}$$

(iii) $X \sim \mathit{Pois}(\lambda)$, where $\lambda > 0$. The corresponding probability mass function is $p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ (for $k \in \mathbb{N}_0$), and

$$\mathbf{E}\big[X\big] = \sum_{k=0}^{\infty} k p_X(k) = \lambda \sum_{k=0}^{\infty} k \cdot \frac{\lambda^{k-1}}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda. \tag{6.19}$$

(iv) $X \sim \mathscr{U}([a,b])$ for $a < b$. The corresponding probability density function is $f_X(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$. We have

$$\mathbf{E}\big[X\big] = \int_a^b x \cdot \frac{1}{b-a} \mathrm{d}x = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}. \tag{6.20}$$

(v) $X \sim \mathscr{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Here the probability density function is $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

$$\begin{aligned} \mathbf{E}\big[X\big] &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y \\ &\quad + \mu \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}}}_{=(*)} \mathrm{d}y = 0 + \mu = \mu, \end{aligned} \tag{6.21}$$

where we used that the expression $(*)$ is again a probability density function of $\mathscr{N}(0, \sigma^2)$, and therefore its integral is one.

(vi) $X \sim \mathit{Geo}(p)$, $p \in (0,1)$ has expectation

$$\mathbf{E}\big[X\big] = \frac{1}{p}. \tag{6.22}$$

(vii) $X \sim \mathscr{E}(\lambda)$, $\lambda > 0$ has expectation

$$\mathbf{E}\big[X\big] = \frac{1}{\lambda}. \tag{6.23}$$

(viii) $X \sim \mathit{Bin}(n,p)$, $n \in \mathbb{N}$, $p \in (0,1)$ has expectation

$$\mathbf{E}\big[X\big] = np. \tag{6.24}$$

(ix) Consider the probability distribution on $\mathbb{N}$ characterized by the probability mass function $p(k) = \frac{6}{\pi^2}\frac{1}{k^2}$.[1] Let $X \sim \mathbf{P}$. Then the expectation of X does not exist. Indeed:

$$\sum_{k=1}^{\infty} k \cdot p_X(k) = \sum_{k=1}^{\infty} \frac{6}{\pi^2}\frac{1}{k} = \infty. \tag{6.25}$$

The claims in (vi) and (vii) are Exercises. Claim (viii) will be shown very easily later, after introducing the notion of independent random variables.

Theorem 6.7. *Let $g : \mathbb{R} \to \mathbb{R}$.*

(i) If X is a discrete real random variable with probability mass function $(p_X(\omega))_{\omega\in\Omega_X}$, then

$$\mathbf{E}\big[g(X)\big] = \sum_{\omega\in\Omega_X} g(\omega)p_X(\omega), \qquad \textit{if} \sum_{\omega\in\Omega_X} |g(\omega)|p_X(\omega) < \infty. \tag{6.26}$$

(ii) If X is a continuous real random variable with probability density function f_X and g is $\mathscr{B}(\mathbb{R})$-$\mathscr{B}(\mathbb{R})$-measurable, then

$$\mathbf{E}\big[g(X)\big] = \int_{-\infty}^{\infty} g(x)f_X(x)\mathrm{d}x, \qquad \textit{if} \int_{-\infty}^{\infty} |g(x)|f_X(x)\mathrm{d}x < \infty. \tag{6.27}$$

Proof. For (i), let $Y = g(X)$, then $\Omega_Y = \{y_1, y_2, ...\}$. Let $A_i = g^{-1}(\{y_i\})$ for $i \in \mathbb{N}$. Clearly, $\Omega_X = \bigcup_{i=1}^{\infty} A_i$, and the A_j are pairwise disjoint. We have

$$\begin{aligned}\mathbf{E}\big[g(X)\big] = \mathbf{E}\big[Y\big] &= \sum_{i=1}^{\infty} y_i \cdot \mathbf{P}_Y[\{y_i\}] = \sum_{i=1}^{\infty} y_i \sum_{\omega_j\in A_i} p_X(\omega_j) \\ &= \sum_{i=1}^{\infty} \sum_{\omega_j\in A_i} g(\omega_j)\cdot p_X(\omega_j) \\ &= \sum_{\omega\in\Omega_X} g(\omega)p_X(\omega).\end{aligned} \tag{6.28}$$

For (ii), the proof of the general case is more complicated and relies on Measure Theory. For the special case where g is strictly increasing / strictly decreasing and differentiable, we can however use Theorem 5.8 and see that for $Y = g(X)$

$$\mathbf{E}\big[g(X)\big] = \int_{-\infty}^{\infty} y f_Y(y)\mathrm{d}y = \int_{-\infty}^{\infty} y f_X\left(g^{-1}(y)\right)\cdot\left|\tfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|\mathrm{d}y = \int_{-\infty}^{\infty} g(x)f_X(x)\mathrm{d}x. \tag{6.29}$$

□

End of Lecture 12

[1] The prefactor is chosen since $\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}$. This can be shown using Fourier series.

Corollary 6.8. *Let X be a real random variable with finite expectation $\mathbf{E}\big[X\big]$. Then, for $a, b \in \mathbb{R}$,*

$$\mathbf{E}\big[aX + b\big] = a\mathbf{E}\big[X\big] + b. \tag{6.30}$$

Proof. We only prove this statement for X continuous or discrete. Without loss of generality, assume that X is continuous (the discrete case proceeds analogously). Consider $g(x) = ax + b$. Then

$$\mathbf{E}\big[g(X)\big] = \int_{-\infty}^{\infty} (ax + b) f_X(x)\mathrm{d}x = a \underbrace{\int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x}_{=\mathbf{E}[X]} + b \underbrace{\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x}_{=1}. \tag{6.31}$$

A similar calculation shows that the integral $\int_{-\infty}^{\infty} |g(x)| f_X(x)\mathrm{d}x$ is finite. □

Theorem 6.9. *Let X, Y be two real random variables, both with finite expectation. Then*

$$\mathbf{E}\big[X + Y\big] = \mathbf{E}\big[X\big] + \mathbf{E}\big[Y\big]. \tag{6.32}$$

We will show this later after introducing the joint distribution of random variables in the next section.

6.2. Variance

Definition 6.10. Let X be a real random variable.

(i) For $k \in \mathbb{N}$, the *k-th moment of X* is defined by

$$\mu_k = \mathbf{E}\big[X^k\big], \qquad \text{if } \mathbf{E}\big[|X|^k\big] < \infty. \tag{6.33}$$

Note that if $\mathbf{E}\big[|X|^k\big] < \infty$ for some $k \in \mathbb{N}$, then $\mathbf{E}\big[|X|^\ell\big] < \infty$ for all $1 \le \ell \le k$, since $|X|^\ell \le 1 + |X|^k$.

(ii) Assume that X has a finite second moment, $\mathbf{E}\big[X^2\big] < \infty$. We define the *variance of X* as

$$\mathrm{Var}[X] = \mathbf{E}\big[(X - \mathbf{E}[X])^2\big]. \tag{6.34}$$

We also define the *standard deviation of X* by

$$\sigma[X] = \sqrt{\mathrm{Var}[X]}. \tag{6.35}$$

The variance is a measure how much the distribution of X typically spreads around its expectation. If it is large, the distribution is well spread-out. If it is small, the the distribution is concentrated around the expectation. Both the notion of kth moment and variance only depend on the law $\mathbf{P}_X$ of X, similar as for the expectation.

Proposition 6.11. *Let X be a real random variable with $\mathbf{E}\big[X^2\big] < \infty$ and $a, b \in \mathbb{R}$. Then*

(i)

$$\mathrm{Var}\big[X\big] = \mathbf{E}\big[X^2\big] - \big(\mathbf{E}\big[X\big]\big)^2 = \mu_2 - \mu_1^2. \tag{6.36}$$

(ii)

$$\mathrm{Var}\big[aX+b\big] = a^2\mathrm{Var}\big[X\big]. \tag{6.37}$$

Proof. We first prove (i):

$$\begin{aligned}\mathrm{Var}\big[X\big] &= \mathbf{E}\left[(X-\mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2\right] \\ &= \mathbf{E}\left[X^2\right] - 2\mathbf{E}[X]^2 + \mathbf{E}[X]^2 = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2.\end{aligned} \tag{6.38}$$

For (ii), we calculate

$$\begin{aligned}\mathrm{Var}[aX+b] &= \mathbf{E}\left[(aX+b-a\mathbf{E}[X]-b)^2\right] \\ &= \mathbf{E}\left[a^2(X-\mathbf{E}[X])^2\right] = a^2\mathrm{Var}[X].\end{aligned} \tag{6.39}$$

□

Let us give some examples.

Example 6.12. (i) Let $X \sim Ber(p)$, $p \in (0,1)$. We have

$$\begin{aligned}\mathbf{E}\big[X\big] &= 0\cdot(1-p) + 1\cdot p = p, \\ \mathbf{E}\big[X^2\big] &= 0^2\cdot(1-p) + 1^2\cdot p = p, \\ \mathrm{Var}[X] &= p - p^2 = p(1-p).\end{aligned} \tag{6.40}$$

(ii) Let $X \sim \mathscr{U}(\{1,...,6\})$. We have

$$\begin{aligned}\mathbf{E}\big[X\big] &= \sum_{k=1}^{6} k\cdot\mathbf{P}[X=k] = 3.5, \\ \mathbf{E}\big[X^2\big] &= \sum_{k=1}^{6} k^2\cdot\mathbf{P}[X=k] = \frac{91}{6}, \\ \Rightarrow \qquad \mathrm{Var}[X] &= \frac{91}{6} - \frac{49}{4} = \frac{70}{24} = \frac{35}{12} \approx 2.92.\end{aligned} \tag{6.41}$$

(iii) Let $X \sim \mathscr{N}(0,1)$. We already saw that $\mathbf{E}\big[X\big] = 0$ (see (6.21)). We now evaluate

$$\begin{aligned}\mathrm{Var}[X] &= \mathbf{E}\left[(X - \underbrace{\mathbf{E}[X]}_{=0})^2\right] = \int_{-\infty}^{\infty} x^2\varphi(x)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2}\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\left[-xe^{-\frac{1}{2}x^2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\mathrm{d}x \\ &= 0 + 1 = 1.\end{aligned} \tag{6.42}$$

Now let $Y = \sigma X + \mu$ for $\mu \in \mathbb{R}$, $\sigma > 0$. Then

$$\begin{aligned} Y \sim& \mathcal{N}(\mu, \sigma^2) \qquad \text{(by (5.18))} \\ \mathrm{Var}[Y] = \sigma^2 &\mathrm{Var}[X] = \sigma^2 \qquad \text{(by (6.37))}. \end{aligned} \tag{6.43}$$

We have established that

$$Z \sim \mathcal{N}(\mu, \sigma^2) \qquad \Rightarrow \qquad \mathbf{E}\big[Z\big] = \mu,\ \mathrm{Var}[Z] = \sigma^2. \tag{6.44}$$

In other words: The standard deviation of $\mathcal{N}(\mu, \sigma^2)$ is exactly σ.

(iv) $X \sim Pois(\lambda)$. Recall that $\mathbf{E}[X] = \lambda$. Then

$$\begin{aligned} \mathbf{E}\big[X(X-1)\big] &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}e^{-\lambda} = \lambda^2 \\ \Rightarrow \qquad \mathbf{E}\big[X^2\big] &= \mathbf{E}\big[X(X-1)\big] + \mathbf{E}[X] = \lambda^2 + \lambda, \\ \Rightarrow \qquad \mathrm{Var}[X] &= \mathbf{E}\big[X^2\big] - \mathbf{E}\big[X\big]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned} \tag{6.45}$$

(v) $X \sim Bin(n,p)$, $n \in \mathbb{N}$, $p \in (0,1)$ has variance

$$\mathrm{Var}[X] = np(1-p). \tag{6.46}$$

End of Lecture 13

We now argue why the variance is indeed a useful quantification how the distribution is spread out. We study the expression $\mathbf{P}\left[|X - \mathbf{E}[X]| \geq \varepsilon\right]$ for $\varepsilon > 0$. The following inequality is called the *Markov inequality.*

Theorem 6.13. *Let X be a non-negative random variable with finite expectation. Then, for $\varepsilon > 0$,*

$$\mathbf{P}\big[X \geq \varepsilon\big] \leq \frac{\mathbf{E}\big[X\big]}{\varepsilon}. \tag{6.47}$$

Proof. We first prove the case where X is discrete. Let $\Omega_X = \{\omega_1, \omega_2, ...\}$. Then

$$\begin{aligned} \mathbf{P}\big[X \geq \varepsilon\big] &= \sum_{\substack{i=1 \\ \omega_i \geq \varepsilon}}^{\infty} \mathbf{P}\big[X = \omega_i\big] \\ &\leq \sum_{\substack{i=1 \\ \omega_i \geq \varepsilon}}^{\infty} \frac{\omega_i}{\varepsilon}\mathbf{P}\big[X = \omega_i\big] \\ &\leq \frac{1}{\varepsilon}\sum_{i=1}^{\infty} \omega_i \mathbf{P}\big[X = \omega_i\big] = \frac{1}{\varepsilon}\mathbf{E}\big[X\big]. \end{aligned} \tag{6.48}$$

Now assume that X is arbitrary and let $n \in \mathbb{N}$. By using the discrete case, we have for any $\varepsilon' > 0$ that

$$\mathbf{P}\big[X_{(n)} \geq \varepsilon'\big] \leq \frac{\mathbf{E}\big[X_{(n)}\big]}{\varepsilon'}. \tag{6.49}$$

However, we know that $\{X \geq \varepsilon\} \subseteq \{X_{(n)} \geq \varepsilon - \frac{1}{n}\}$ by Lemma 6.3, (i). For fixed $\varepsilon > 0$ and large enough $n \in \mathbb{N}$, we therefore obtain by (6.49) that

$$\mathbf{P}\big[X \geq \varepsilon\big] \leq \mathbf{P}\big[X_{(n)} \geq \varepsilon - \tfrac{1}{n}\big] \leq \frac{\mathbf{E}\big[X_{(n)}\big]}{\varepsilon - \frac{1}{n}}. \tag{6.50}$$

The result in the general case now follows by letting $n \to \infty$ and by using the definition (6.4) for the expectation. □

From the Markov inequality, we have the following result, called the *Chebyshev inequality.*

Theorem 6.14. *Let* $\mathbf{E}\big[X^2\big] < \infty$. *For any* $a \in \mathbb{R}$ *and* $\varepsilon > 0$, *one has*

$$\mathbf{P}\big[|X - a| \geq \varepsilon\big] \leq \frac{\mathbf{E}\big[(X-a)^2\big]}{\varepsilon^2}. \tag{6.51}$$

In particular, for $a = \mathbf{E}[X]$ *one has*

$$\mathbf{P}\big[|X - \mathbf{E}[X]| \geq \varepsilon\big] \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}. \tag{6.52}$$

Proof. We apply the Markov inequality (6.47) to the non-negative random variable $(X-a)^2$. It follows that

$$\begin{aligned} \mathbf{P}\big[|X - a| \geq \varepsilon\big] &= \mathbf{P}\big[(X-a)^2 \geq \varepsilon^2\big] \\ &\overset{(6.47)}{\leq} \frac{\mathbf{E}\big[(X-a)^2\big]}{\varepsilon^2}. \end{aligned} \tag{6.53}$$

□

Chebyshev's inequality gives a bound how likely it is that the result of a random number deviates by a certain amount from its expectation. In particular, we have the following "$k\sigma$-rules":

Corollary 6.15. *Let* $\mathbf{E}\big[X^2\big] < \infty$.

(i) If $\sigma = \sigma[X] = \sqrt{\mathrm{Var}(X)} > 0$, *we have*

$$\mathbf{P}\left[|X - \mathbf{E}[X]| \geq k\sigma\right] \leq \frac{1}{k^2}, \tag{6.54}$$

for $k > 0$. *In particular:*

$$\begin{aligned} \mathbf{P}\left[|X - \mathbf{E}[X]| \geq 2\sigma\right] &\leq \frac{1}{4}, \\ \mathbf{P}\left[|X - \mathbf{E}[X]| \geq 3\sigma\right] &\leq \frac{1}{9}. \end{aligned} \tag{6.55}$$

(ii) If $\mathrm{Var}\big[X\big] = 0$*, then*

$$\mathbf{P}\big[X = \mathbf{E}[X]\big] = 1. \tag{6.56}$$

Remark 6.16. The Chebyshev inequality is very general, but only gives very rough bounds. Consider for instance $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$, $\sigma > 0$. Then

$$\begin{aligned}\mathbf{P}\big[|X-\mu| < k\sigma\big] &= \mathbf{P}\bigg[\bigg|\underbrace{\frac{X-\mu}{\sigma}}_{\sim\mathcal{N}(0,1)}\bigg| < k\bigg] \\ &= \mathbf{P}\bigg[-k < \frac{X-\mu}{\sigma} \le k\bigg] = \Phi(k) - \Phi(-k) \\ &= 2\Phi(k) - 1.\end{aligned} \tag{6.57}$$

From this it follows that

$$\mathbf{P}\big[|X-\mu| \ge k\sigma\big] = 1 - (2\Phi(k) - 1) = 2 - 2\Phi(k). \tag{6.58}$$

For instance, we have

$$\begin{aligned}\mathbf{P}\big[|X-\mu| \ge \sigma\big] &\approx 2 - 2\cdot 0.84 = 0.32,\\ \mathbf{P}\big[|X-\mu| \ge 2\sigma\big] &\approx 2 - 2\cdot 0.98 = 0.04,\\ \mathbf{P}\big[|X-\mu| \ge 3\sigma\big] &\approx 2 - 2\cdot 0.9987 = 0.0026.\end{aligned} \tag{6.59}$$

The last equality in (6.57) follows from the symmetry of the density φ.

$$\begin{aligned}\Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t &= 1 - \int_{x}^{\infty} \varphi(t)\mathrm{d}t \\ &= 1 - \int_{-\infty}^{-x} \varphi(t)\mathrm{d}t \\ &= 1 - \Phi(-x).\end{aligned} \tag{6.60}$$

7. Joint distributions of random variables, covariance, extremes

(Reference: [GS01, Sections 3.2, 3.6, 3.8, 4.2, 4.5, 4.8], or [Geo12, Sections 3.3, 3.4, 4.3])

7.1. Joint distributions of random variables

In several situations, it is necessary to study multiple random variables at once and the way they are related. For instance, consider

$$\begin{aligned} X &= \text{ length of a randomly chosen fish of species A,} \\ Y &= \text{ age of a randomly chosen fish of species A.} \end{aligned} \tag{7.1}$$

Of course we expect that there should be a certain dependence between the values of X and Y. In this context, we want to study the joint distribution of X and Y, that is the random vector (!) $(X, Y) \in \mathbb{R}^2$.

Let us start with a simple example. Consider tossing a coin three times.

$$\begin{aligned} &\Omega = \{0,1\}^3 = \{(\omega_1, \omega_2, \omega_3)\,;\, \omega_i \in \{0,1\}\}, \\ &X(\omega) = \omega_1, \qquad \Omega_X = \{0,1\}, \\ &Y(\omega) = \sum_{i=1}^{3} \omega_i, \qquad \Omega_Y = \{0,1,2,3\}. \end{aligned} \tag{7.2}$$

The random vector (X, Y) then takes values in $\Omega_{(X,Y)} = \Omega_X \times \Omega_Y \subseteq \mathbb{R}^2$. Since $\Omega_X \times \Omega_Y$ is a finite set, we can consider $Z = (X, Y)$ as a random variable with values in $\Omega_X \times \Omega_Y$ and study its distribution on

$$\begin{aligned} &\big(\Omega_X \times \Omega_Y, \mathscr{P}(\Omega_X \times \Omega_Y)\big), \text{ where } \Omega_X \times \Omega_Y = \big\{(x,y)\,;\, x \in \Omega_X, y \in \Omega_Y\big\}. \\ &\mathbf{P}_{(X,Y)}[A] = \mathbf{P}\big[(X,Y)^{-1}(A)\big], \qquad A \in \mathscr{P}(\Omega_X \times \Omega_Y). \end{aligned} \tag{7.3}$$

We equip $\big(\Omega, \mathscr{P}(\Omega)\big)$ with the uniform distribution $\mathbf{P} = \mathscr{U}(\Omega)$. We say that $\mathbf{P}_{(X,Y)}$ is the *joint distribution* of X and Y. As an example:

$$\begin{aligned} &A = \big\{(0,1),(1,2)\big\}, \qquad (X,Y)^{-1}(A) = \big\{(0,1,0),(0,0,1),(1,1,0),(1,0,1)\big\} \\ \Rightarrow \qquad &\mathbf{P}_{(X,Y)}[A] = \frac{|(X,Y)^{-1}(A)|}{|\Omega|} = \frac{4}{8} = \frac{1}{2}. \end{aligned} \tag{7.4}$$

We say that

$$p_{X,Y}(x_i, y_j) = \mathbf{P}_{(X,Y)}\big[\{(x_i, y_j)\}\big], \qquad (x_i, y_j) \in \Omega_X \times \Omega_Y \tag{7.5}$$

is the *joint probability mass function of X and Y*. Let us compute all of the values in the example above:

$x_i \setminus y_j$	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

Can we obtain the probability mass function of X or Y from $p_{(X,Y)}$? Let more generally be $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be two discrete random variables and let $\Omega_X = \{x_1, x_2, ...\}$ and $\Omega_Y = \{y_1, y_2, ...\}$.

$$\begin{aligned} p_X(x_i) = \mathbf{P}[\{X = x_i\}] &= \sum_{j=1}^{\infty} \mathbf{P}[X = x_i, Y = y_j] \\ &= \sum_{j=1}^{\infty} p_{(X,Y)}(x_i, y_j). \end{aligned} \tag{7.6}$$

Similarly, we have

$$p_Y(y_j) = \sum_{i=1}^{\infty} p_{(X,Y)}(x_i, y_j). \tag{7.7}$$

The probability distributions $\mathbf{P}_X$ and $\mathbf{P}_Y$, which are given in terms of p_X and p_Y are called the *marginal distributions* of (X, Y). Turning back to the example above, can fill in the table

$x_i \setminus y_j$	0	1	2	3	$p_X(x_i)$
0	1/8	2/8	1/8	0	1/2
1	0	1/8	2/8	1/8	1/2
$p_Y(y_j)$	1/8	3/8	3/8	1/8	1

We want to make this discussion more general by

- considering more than two random variables,
- dropping the assumption that the random variables are discrete.

To this end, we want to define an equivalent notion of measurability as in Definition 5.1 for a vector $(X_1, ..., X_n)$ consisting of n random variables $X_1, ..., X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ defined on the same probability space $(\Omega, \mathscr{F}, \mathbf{P})$. As previously, we want to speak about the law of the random vector $(X_1, ..., X_n)$, $\mathbf{P}_{(X_1,...,X_n)}[B]$ for certain subsets of $\mathbb{R}^n$, similar as in Definition 5.3. To do this, we need to introduce a σ-algebra on $\mathbb{R}^n$.

Definition 7.1. Consider the system of subsets of $\mathbb{R}^n$ given by

$$\mathcal{E} = \big\{[a_1, b_1) \times [a_2, b_2) \times \dots \times [a_n, b_n)\,;\, a_i < b_i, a_i, b_i \in \mathbb{R}\big\}. \tag{7.8}$$

The *Borel σ-algebra* $\mathcal{B}(\mathbb{R}^n)$ on $\mathbb{R}^n$ is defined by

$$\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{E}). \tag{7.9}$$

Again, every set of our imagination is contained in $\mathcal{B}(\mathbb{R}^n)$, such as every countable set of points, lines, hyperplanes, cubes, cylinders, balls, ...

The following lemma ensures that we can consider a vector of random variable as a random variable with values in $\mathbb{R}^n$.

Lemma 7.2. *Let $(\Omega, \mathcal{F})$ be a measurable space and $X_1, \dots, X_n : \Omega \to \mathbb{R}$ be maps. Then*

$$X_1, \dots, X_n \text{ are } \mathcal{F}\text{-}\mathcal{B}(\mathbb{R}) \text{ measurable} \;\Leftrightarrow\; (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n \text{ is } \mathcal{F}\text{-}\mathcal{B}(\mathbb{R}^n) \text{ measurable.} \tag{7.10}$$

The latter means that

$$(X_1, \dots, X_n)^{-1}(A) \in \mathcal{F} \qquad \text{for all } A \in \mathcal{B}(\mathbb{R}^n). \tag{7.11}$$

Proof. ($\Leftarrow$) Consider for any $1 \le i \le n$ the *projection maps*

$$\pi_i : \mathbb{R}^n \to \mathbb{R}, \qquad (x_1, \dots, x_n) \mapsto x_i. \tag{7.12}$$

Note that $\pi_i^{-1}((-\infty, a]) = \mathbb{R} \times \dots \times (-\infty, a] \times \dots \times \mathbb{R} \in \mathcal{B}(\mathbb{R}^n)$ (with the replacement at the i-th entry). By Remark 5.2, π_i is $\mathcal{B}(\mathbb{R})$-$\mathcal{B}(\mathbb{R}^n)$-measurable. Therefore, $X_i = \pi_i \circ (X_1, \dots, X_n)$ is $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$-measurable as a composition of measurable maps.

($\Rightarrow$) We only sketch the proof. Assume that the X_i are $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$-measurable and let $a_1, \dots, a_n \in \mathbb{R}$. We see that

$$(X_1, \dots, X_n)^{-1}((-\infty, a_1] \times \dots \times (-\infty, a_n]) = \bigcup_{i=1}^{n} X_i^{-1}((-\infty, a_i]) \in \mathcal{F}. \tag{7.13}$$

Since the family $\mathcal{E}' = \{(-\infty, a_1] \times \dots \times (-\infty, a_n] \;:\; a_1, \dots, a_n \in \mathbb{R}\}$ generates $\mathcal{B}(\mathbb{R}^n)$, one can show that $(X_1, \dots, X_n)$ is $\mathcal{F}$-$\mathcal{B}(\mathbb{R}^n)$-measurable, analogously to Remark 5.2. □

Definition 7.3. Let $X_1, \dots, X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be n real random variables and $\mathbf{P}$ a probability measure on $(\Omega, \mathcal{F})$.

(i) The *joint law of $X_1, \dots, X_n$ under* $\mathbf{P}$ is the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

$$\mathbf{P}_{(X_1,\dots,X_n)}[B] = \mathbf{P}\big[(X_1, \dots, X_n)^{-1}(B)\big], \qquad B \in \mathcal{B}(\mathbb{R}^n). \tag{7.14}$$

(ii) For $x_1, ..., x_n \in \mathbb{R}$, we set

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \mathbf{P}_{(X_1,...,X_n)}\big[(-\infty, x_1] \times ... \times (-\infty, x_n]\big] \\ = \mathbf{P}\big[X_1 \leq x_1, ..., X_n \leq x_n\big]. \tag{7.15}$$

The function $F_{X_1,...,X_n}$ is called the *joint cumulative distribution function* of $X_1, ..., X_n$ / the *cumulative distribution function of the law of* $(X_1, ..., X_n)$.[1]

Note that by Lemma 7.2, the event $(X_1, ..., X_n)^{-1}(B)$ under the probability in (7.14) is in $\mathcal{F}$.

End of Lecture 14

We are now ready to define multivariate continuous distributions.

Definition 7.4. A Riemann-integrable function $f : \mathbb{R}^n \to [0, \infty)$ is called a *probability density function* if

$$\int_{\mathbb{R}^n} f(x_1, ..., x_n)\mathrm{d}^n x = 1. \tag{7.16}$$

We define $\mathbf{P}$ to be the (unique) probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ that fulfills

$$\mathbf{P}\big[[a_1, b_1) \times [a_2, b_2) \times ... \times [a_n, b_n)\big] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} ... \int_{a_n}^{b_n} f(x_1, ...x_n)\mathrm{d}x_n ... \mathrm{d}x_2\, \mathrm{d}x_1. \tag{7.17}$$

Such a probability measure is called a *(multivariate) continuous distribution.*

Example 7.5. The function

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right) \tag{7.18}$$

is a probability density function on $\mathbb{R}^2$, the density of a *multivariate standard normal distribution* $\mathcal{N}(0, I_{2\times 2})$. We will study such multivariate normal distributions later.

Remark 7.6. (i) One can show that the measure $\mathbf{P}$ in the above definition exists (using measure theory), and for any $A \in \mathcal{B}(\mathbb{R}^n)$, we would like to write

$$\mathbf{P}[A] = \int_A f(x_1, ..., x_n)\mathrm{d}^n x, \tag{7.19}$$

The definition in (7.19) can be understood as a multidimensional Riemann-integral if A is Jordan measurable. Typically we are only interested in sets A over which this integral can be performed as an iterated Riemann integral.

(ii) The uniqueness follows by Corollary 4.18 to Dynkin's π-λ-theorem. Indeed $\mathbf{P}$ is defined by (7.17) on the $\cap$-stable generator $\mathcal{E} \cup \{\varnothing\}$ with $\mathcal{E}$ as in (7.8).

[1] This wording reflects that there are two ways to view how $X_1, ..., X_n$ "interact": We may view them as n different $\mathbb{R}$-valued functions, or we look at the single, $\mathbb{R}^n$-valued function $(X_1, ..., X_n)$. In the former interpretation, we may think about $\mathbf{P}_{(X_1,...,X_n)}$ as the joint law of $X_1, ..., X_n$, in the latter we can say that $\mathbf{P}_{(X_1,...,X_n)}$ is *the* law of the vector-valued random variable $(X_1, ..., X_n)$.

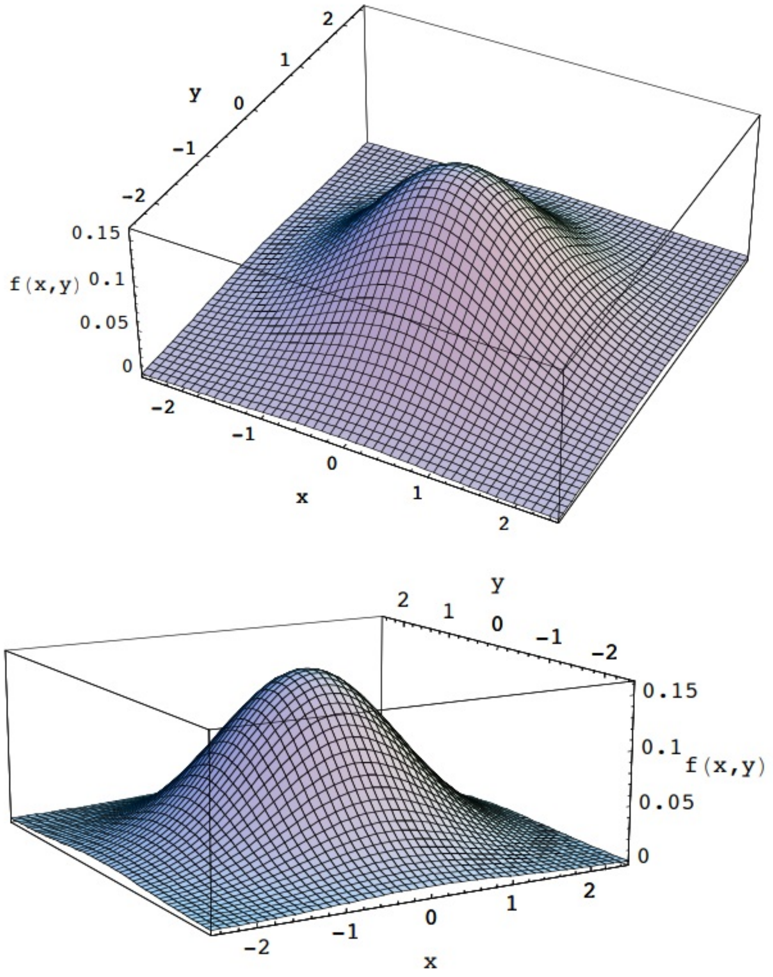

Figure 7.1.: Plot of the density f in (7.18)

(iii) Equality in distributions of random vectors $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$ is defined as in Definition 5.5. Similarly as in the one-dimensional case, the law $\mathbf{P}_{(X_1,...,X_n)}$ is uniquely determined by $F_{X_1,...,X_n}$.

Let us now now restrict our attention to the case of either discrete or continuous real random vectors.

Definition 7.7. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X_1, ..., X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ random variables.

(i) If all random variables are discrete with values in the countable sets $\Omega_{X_1}, ..., \Omega_{X_n} \subseteq \mathbb{R}$, then $(X_1, ..., X_n)$ only takes values in the countable set $\Omega_{X_1} \times ... \times \Omega_{X_n}$. We say that $(X_1, ..., X_n)$ is a *discrete (real) random vector*. The law $\mathbf{P}_{(X_1,...,X_n)}$ on $(\Omega_{X_1} \times ... \times \Omega_{X_n}, \mathscr{P}(\Omega_{X_1} \times ... \times \Omega_{X_n}))$ is characterized by

$$\begin{aligned} p_{X_1,...,X_n}(x_1, ..., x_n) &= \mathbf{P}_{(X_1,...,X_n)}[\{(x_1, ..., x_n)\}] \\ &= \mathbf{P}[X_1 = x_1, ..., X_n = x_n], \qquad x_i \in \Omega_{X_i}, 1 \le i \le n. \end{aligned} \tag{7.20}$$

Here, $(p_{X_1,...,X_n}(x_1, ..., x_n))_{(x_1,...,x_n)\in\Omega_{X_1}\times...\times\Omega_{X_n}}$ is the *joint probability mass function* of $X_1, ..., X_n$. For $I \subseteq \{1, ..., n\}$ one has

$$p_{(X_i;i\in I)}(x_i; i \in I) = \mathbf{P}[X_i = x_i; i \in I] = \sum_{\substack{x_j\in\Omega_{X_j} \\ j\notin I}} p_{X_1,...,X_n}(x_1, ..., x_n) \tag{7.21}$$

for the marginal joint probability mass function of $(X_i)_{i\in I}$.

(ii) If the joint law $\mathbf{P}_{(X_1,...,X_n)}$ of $X_1, ..., X_n$ is a continuous distribution on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, i.e. it is defined in terms of a multivariate probability density $f_{X_1,...,X_n}$ as in (7.19), then $(X_1, ...X_n)$ is a *continuous (real) random vector*, or $X_1, ..., X_n$ are *jointly continuous*. The function $f_{X_1,...X_n}$ is called the *joint probability density function* of $X_1, ..., X_n$. For $I \subseteq \{1, ..., n\}$ one has

$$f_{(X_i;i\in I)}(x_i; i \in I) = \iint ... \int_{\mathbb{R}^{n-|I|}} f_{X_1,...,X_n}(x_1, ..., x_n) \prod_{j\notin I}(\mathrm{d}x_j) \tag{7.22}$$

for the marginal joint probability density function of $(X_i)_{i\in I}$.

Remark 7.8. Note that it contrary to the case where all $X_1, ..., X_n$ are discrete, it is *not* the case that $(X_1, ..., X_n)$ is a continuous random vector, if all $X_1, ..., X_n$ are all continuous. As a counterexample, take $X = Y \sim \mathscr{N}(0, 1)$. Then the vector $(X, Y) = (X, X)$ is not a continuous random vector, since

$$\mathbf{P}\big[(X, Y) \in \Delta\big] = \mathbf{P}_{(X,Y)}\big[\Delta\big] = 1, \qquad \text{where } \Delta = \{(t, t) \,;\, t \in \mathbb{R}\}. \tag{7.23}$$

This cannot be true if $\mathbf{P}_{(X,Y)}$ was given in terms of a multivariate probability density f, since $\int_\Delta f(x, y)\mathrm{d}x\mathrm{d}y = 0$.

We want to practice a bit how to work with joint cumulative distribution functions / probability density functions / probability mass functions.

Proposition 7.9. *Let $(X_1, ..., X_n)$ be a continuous random vector. We have*

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{X_1,...,X_n}(t_1, ..., t_n)\mathrm{d}t_n...\mathrm{d}t_1. \tag{7.24}$$

Proof. First, we see that

$$\begin{aligned}&\mathbf{P}_{(X_1,...,X_n)}[A_1 \times A_2 \times ... \times \underbrace{\{a\}}_{\text{position } j} \times ... \times A_n] = 0,\\ &\qquad\text{for all } A_1, ..., A_{j-1}, A_{j+1}, ..., A_n \in \mathscr{B}(\mathbb{R}), a \in \mathbb{R}.\end{aligned} \tag{7.25}$$

This follows similarly as (4.14) in the one-dimensional case. Therefore:

$$\mathbf{P}_{(X_1,...,X_n)}\big[(a_1, x_1] \times (a_2, x_2] \times ... \times (a_n, x_n]\big] = \int_{a_1}^{x_1}\int_{a_2}^{x_2} ... \int_{a_n}^{x_n} f(t_1, ..., t_n)\mathrm{d}t_n ... \mathrm{d}t_2\, \mathrm{d}t_1. \tag{7.26}$$

We see that $(a_1, x_1] \times ... \times (-k, x_n] \subseteq (a_1, x_1] \times ... \times (-k, x_n]$ and

$$\bigcup_{k=1}^{\infty}(a_1, x_1] \times ... \times (-k-1, x_n] = (a_1, x_1] \times ... \times (-\infty, x_n]. \tag{7.27}$$

By Proposition 1.15, (vi), it follows that

$$\begin{aligned}&\mathbf{P}_{(X_1,...,X_n)}\big[(a_1, x_1] \times (a_2, x_2] \times ... \times (-\infty, x_n]\big]\\ &\qquad= \lim_{k\to\infty}\int_{a_1}^{x_1}\int_{a_2}^{x_2} ... \int_{-k}^{x_n} f(t_1, ..., t_n)\mathrm{d}t_n...\mathrm{d}t_2\mathrm{d}t_1\\ &\qquad= \int_{a_1}^{x_1}\int_{a_2}^{x_2} ... \int_{-\infty}^{x_n} f(t_1, ..., t_n)\mathrm{d}t_n...\mathrm{d}t_2\mathrm{d}t_1.\end{aligned} \tag{7.28}$$

Repeating this procedure for the other integration variables yields the claim. □

Example 7.10. (i) Let X, Y be distributed with the joint density

$$f_{X,Y}(x, y) = \begin{cases} 2x + 2y - 4xy, & x, y \in [0, 1],\\ 0, & x, y \notin [0, 1].\end{cases} \tag{7.29}$$

We have that

$$\begin{aligned}f_X(x) = \int_0^1 (2x + 2y - 4xy)\mathrm{d}y = 2x + 1 - 2x = 1, \qquad & x \in [0, 1],\\ f_Y(y) = \int_0^1 (2x + 2y - 4xy)\mathrm{d}x = 1, \qquad & y \in [0, 1].\end{aligned} \tag{7.30}$$

Therefore $X \sim \mathcal{U}([0,1])$, $Y \sim \mathcal{U}([0,1])$. Let us calculate $\mathbf{P}[X^2 \leq Y]$:

$$\begin{aligned}\mathbf{P}[X^2 \leq Y] &= \int_0^1 \int_{x^2}^1 (2x + 2y - 4xy)\mathrm{d}y\mathrm{d}x \\ &= \frac{19}{30}.\end{aligned} \tag{7.31}$$

(ii) We now consider X, Y with the joint density

$$f_{X,Y}(x,y) = \mathbb{1}_{[0,1]^2}(x,y). \tag{7.32}$$

This is the *uniform distribution on* $[0,1]^2$, denoted by $\mathcal{U}([0,1]^2)$. We see that

$$\begin{aligned}f_X(x) &= \int_0^1 1 \cdot \mathrm{d}y = 1, \qquad x \in [0,1], \\ f_Y(y) &= \int_0^1 1 \cdot \mathrm{d}x = 1, \qquad y \in [0,1].\end{aligned} \tag{7.33}$$

So again, $X \sim \mathcal{U}([0,1])$ and $Y \sim \mathcal{U}([0,1])$. We calculate again $\mathbf{P}[X^2 \leq Y]$:

$$\begin{aligned}\mathbf{P}[X^2 \leq Y] &= \int_0^1 \int_{x^2}^1 1 \cdot \mathrm{d}y\mathrm{d}x = \int_0^1 (1 - x^2)\mathrm{d}x \\ &= \frac{2}{3}.\end{aligned} \tag{7.34}$$

This shows that even if X and Y have the same marginal densities, their joint distribution can be different from case to case.

Theorem 7.11. *Let $g : \mathbb{R}^n \to \mathbb{R}$.*

(i) If $X_1, ..., X_n$ are discrete real random variables with joint probability mass function given by $(p_{X_1,...,X_n}(x_1,...,x_n))_{(x_1,...,x_n) \in \Omega_{X_1} \times ... \Omega_{X_N}}$, then

$$\begin{aligned}\mathbf{E}\big[g(X_1,...,X_n)\big] &= \sum_{(x_1,...,x_n)\in\Omega_{X_1}\times...\times\Omega_{X_n}} g(x_1,...,x_n)p_{X_1,...,X_n}(x_1,...,x_n), \\ \textit{if} &\sum_{(x_1,...,x_n)\in\Omega_{X_1}\times...\times\Omega_{X_n}} |g(x_1,...,x_n)|p_{X_1,...,X_n}(x_1,...,x_n) < \infty.\end{aligned} \tag{7.35}$$

(ii) If $(X_1,...,X_n)$ is a continuous random vector and $X_1,...,X_n$ have the joint probability density function $f_{X_1,...,X_n}$ (and g is measurable), then

$$\begin{aligned}\mathbf{E}\big[g(X_1,...,X_n)\big] &= \iint ... \int_{\mathbb{R}^n} g(x_1,...,x_n)f_{X_1,...,X_n}(x_1,...,x_n)\mathrm{d}^n x, \\ \textit{if} \iint ... \int_{\mathbb{R}^n} &|g(x_1,...,x_n)|f_{X_1,...,X_n}(x_1,...,x_n)\mathrm{d}^n x < \infty.\end{aligned} \tag{7.36}$$

Proof. This is analogous to Theorem 6.7. □

End of Lecture 15

We can now prove the additivity of the expectation claimed in the previous section:

Proof of Theorem 6.9. We first assume that X and Y are both discrete. Then, by applying (7.35) with the function $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x, y) = x + y$:

$$\begin{aligned}\mathbf{E}\big[X+Y\big] &= \sum_{x\in\Omega_X, y\in\Omega_Y} (x+y)p_{X,Y}(x,y) \\ &= \sum_{x\in\Omega_X} x \underbrace{\sum_{y\in\Omega_Y} p_{X,Y}(x,y)}_{=p_X(x)} + \sum_{y\in\Omega_Y} y \underbrace{\sum_{x\in\Omega_X} p_{X,Y}(x,y)}_{=p_Y(y)} \\ &= \mathbf{E}\big[X\big] + \mathbf{E}\big[Y\big].\end{aligned} \tag{7.37}$$

The general case follows by approximation: Indeed, we have that $|(X+Y)_{(n)} - (X+Y)| \le \frac{1}{n}$ as well as $|X_{(n)} - X| \le \frac{1}{n}$ and $|Y_{(n)} - Y| \le \frac{1}{n}$ for any $n \in \mathbb{N}$. It follows that

$$|(X+Y)_{(n)} - X_{(n)} - Y_{(n)}| \le \frac{3}{n}. \tag{7.38}$$

It therefore follows that $\lim_{n\to\infty} |\mathbf{E}[(X+Y)_{(n)}] - \mathbf{E}[X_{(n)}] - \mathbf{E}[Y_{(n)}]| = 0$, and by the general definition of the expectation, the claim follows. □

7.2. Independence of random variables

Consider two random variables $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. If $A, B \in \mathscr{B}(\mathbb{R})$, we can consider the two events

$$X^{-1}(A) \qquad \text{and} \qquad Y^{-1}(B). \tag{7.39}$$

We already have a notion of independence of these events: If we cannot obtain any information about $X^{-1}(A)$ from knowing whether $Y^{-1}(B)$ has occured, then the two events are independent. If this is the case for *any* choice of the sets A and B, we will not be able to obtain any information from X about Y or vice versa, since $\{X^{-1}(A)\,;\, A \in \mathscr{B}(\mathbb{R})\}$ contains all possible events that we can observe by knowing X[2], and $\{Y^{-1}(B)\,;\, B \in \mathscr{B}(\mathbb{R})\}$ contains all possible events we can observe by knowing Y. We come to the general definition of stochastic independence of random variables.

Definition 7.12. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space.

[2] We later see that this can be understood as the *generated σ-algebra of X*, written $\sigma(X)$. For instance, if $X : \Omega = \{1,2,3,4,5,6\} \to \mathbb{R}$, $X(\omega) = \mathbb{1}_{\{1,3,5\}}(\omega)$ (i.e. $X(\omega) = 1$ if ω is odd and $X(\omega) = 0$ if ω is even). Then $\sigma(X) = \{\varnothing, \{1,3,5\}, \{2,4,6\}, \Omega\}$. This means that the “information” contained in knowing the value of $X(\omega)$ is exactly whether or not ω is even, namely whether $\omega \in \{1,3,5\}$ or $\omega \in \{2,4,6\}$.

(i) The random variables $X_1, ..., X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are *independent*, if
(7.40)
$$\mathbf{P}\big[X_1 \in A_1, X_2 \in A_2, ..., X_n \in A_n\big] = \prod_{i=1}^{n} \mathbf{P}\big[X_i \in A_i\big], \qquad \text{for } A_1, ..., A_n \in \mathscr{B}(\mathbb{R}).$$

(ii) The random variables $(X_i\,;\, i \in \mathscr{I})$, where $X_i : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and $\mathscr{I}$ is an arbitrary set, are *independent*, if for all finite sets $\{i_1, ..., i_n\} \subseteq \mathscr{I}$ (with $i_1, ..., i_n$ pairwise distinct):
(7.41)
$$\mathbf{P}\big[X_{i_1} \in A_1, X_{i_2} \in A_2, ..., X_{i_n} \in A_n\big] = \prod_{j=1}^{n} \mathbf{P}\big[X_{i_j} \in A_j\big], \qquad \text{for } A_1, ..., A_n \in \mathscr{B}(\mathbb{R}).$$

Of course (i) is a special case of (ii) with $\mathscr{I} = \{1, ..., n\}$. Let us discuss some special cases:

Proposition 7.13. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X_1, ..., X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ random variables.*

(i) $X_1, ..., X_n$ are independent if and only if

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \qquad \text{for all } x_1, ..., x_n \in \mathbb{R}. \tag{7.42}$$

(ii) Assume that $X_1, ..., X_n$ are discrete random variables. Then $X_1, ..., X_n$ are independent if and only if

$$\mathbf{P}\big[X_1 = x_1, ..., X_n = x_n\big] = \prod_{i=1}^{n} \mathbf{P}\big[X_i = x_i\big], \qquad \text{for all } x_i \in \Omega_{X_i}, 1 \le i \le n. \tag{7.43}$$

(iii) Assume that $X_1, ..., X_n$ are continuous random variables. Then $X_1, ..., X_n$ are independent if and only if they are jointly continuous and

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \qquad \text{for all } x_1, ..., x_n \in \mathbb{R}. \tag{7.44}$$

Proof. For (i), we assume that $X_1, ..., X_n$ are independent, then

$$\mathbf{P}\big[X_1 \le x_1, ..., X_n \le x_n\big] = \prod_{i=1}^{n} \mathbf{P}\big[X_i \le x_i\big]. \tag{7.45}$$

The other direction follows from Corollary 4.18 to Dynkin's π-λ-theorem: Indeed, suppose that (7.42) holds, then the two probability measures on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, $\mathbf{P}_{(X_1,...,X_n)}$ and the *product measure* $\bigotimes_{i=1}^{n} \mathbf{P}_{X_i}$, defined[3] by

$$\bigotimes_{i=1}^{n} \mathbf{P}_{X_i}[A_1 \times ... \times A_n] = \prod_{i=1}^{n} \mathbf{P}_{X_i}[A_i], \qquad A_i \in \mathscr{B}(\mathbb{R}) \tag{7.46}$$

[3] In fact, one needs an extension theorem from measure theory to define the latter on all $\mathscr{B}(\mathbb{R}^n)$.

coincide on the $\cap$-stable generator $\mathscr{E}' = \{(-\infty, a_1] \times ... \times (-\infty, a_n] \,:\, a_1, ..., a_n \in \mathbb{R}\}$, and since $\mathscr{B}(\mathbb{R}^n) = \sigma(\mathscr{E}')$, they are equal.

For (ii), we consider $A_1 \in \mathscr{P}(\Omega_{X_1})$,..., $A_n \in \mathscr{P}(\Omega_{X_n})$. Then

$$\begin{aligned} \mathbf{P}\big[X_1 \in A_1, ..., X_n \in A_n\big] &= \sum_{x_1 \in A_1, ..., x_n \in A_n} \mathbf{P}\big[X_1 = x_1, ..., X_n = x_n\big] \\ &= \sum_{x_1 \in A_1, ..., x_n \in A_n} \prod_{i=1}^{n} \mathbf{P}\big[X_i = x_i\big] = \prod_{i=1}^{n} \mathbf{P}\big[X_i \in A_i\big]. \end{aligned} \tag{7.47}$$

For (iii), assume that $X_1, ..., X_n$ are independent, then by part (i) we have (7.42). The claim then follows by taking the partial derivatives with respect to $x_1, ..., x_n$. For the other direction, let us assume that (7.44) holds. Take $x_1, ..., x_n \in \mathbb{R}$, then by Proposition 7.9,

$$\begin{aligned} F_{X_1,...,X_n}(x_1, ..., x_n) &= \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{X_1,...,X_n}(t_1, ..., t_n)\mathrm{d}t_n...\mathrm{d}t_1 \\ &= \prod_{i=1}^{n} \int_{-\infty}^{x_i} f_{X_i}(t_i)\mathrm{d}t_i = \prod_{i=1}^{n} F_{X_i}(x_i). \end{aligned} \tag{7.48}$$

It follows from part (i) that $X_1, ..., X_n$ are independent. □

Example 7.14. Consider again Example 7.10. The random variables X and Y in part (i) are not independent. On the other hand, the random variables X and Y in part (ii), since $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Intuitively, if the random variables X and Y are independent, then $Z = f(X)$ and $W = g(Y)$ should also be independent for any functions $f, g : \mathbb{R} \to \mathbb{R}$. This is indeed the case, and we have the *propagation of independence*:

Proposition 7.15. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X_1, ..., X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ independent real random variables. Furthermore, assume that $h_i : (\mathbb{R}, \mathscr{B}(\mathbb{R})) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are measurable functions. Then the functions*

$$h_1(X_1), ..., h_n(X_n) \text{ are independent.} \tag{7.49}$$

Proof. For all $B_i \in \mathscr{B}(\mathbb{R})$, we have

$$\begin{aligned} \mathbf{P}\big[h_1(X_1) \in B_1, ..., h_n(X_n) \in B_n\big] &= \mathbf{P}\big[X_1 \in h_1^{-1}(B_1), ..., X_n \in h_n^{-1}(B_n)\big] \\ &= \prod_{i=1}^{n} \mathbf{P}\big[X_i \in h_i^{-1}(B_i)\big] = \prod_{i=1}^{n} \mathbf{P}\big[h_i(X_i) \in B_i\big]. \end{aligned} \tag{7.50}$$

□

Example 7.16. If X and Y are independent, then $Z = X^2 - \sin(X)$ and $W = \cos(Y^2)\mathbb{1}_{\{Y \geq 0\}}$ are independent as well.

The proof of Proposition 7.15 works in a similar way for vector-valued random variables. With this, we can show that any functions of two disjoint subsets of $\{X_1, ..., X_n\}$ are again independent, if $X_1, ..., X_n$ are independent.

Example 7.17. If X_1, X_2, X_3, X_4, X_5 are independent, then $Z = \sin(X_1^2 + X_3^2)$ and $Y = X_5 \exp(X_2)$ are independent.

Theorem 7.18. *Let X and Y be independent real random variables with $\mathbf{E}\left[X^2\right] < \infty$ and $\mathbf{E}\left[Y^2\right] < \infty$.[4] Then*

$$\mathbf{E}\left[XY\right] = \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]. \tag{7.51}$$

Proof. We first assume that X and Y are both discrete. Then, by applying (7.35) with the (measurable) function $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x, y) = xy$:

$$\begin{aligned} \mathbf{E}\left[XY\right] &= \sum_{x \in \Omega_X, y \in \Omega_Y} xy \underbrace{p_{X,Y}(x, y)}_{= p_X(x) p_Y(y)} \\ &= \sum_{x \in \Omega_X} x p_X(x) \sum_{y \in \Omega_Y} y p_Y(y) \\ &= \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]. \end{aligned} \tag{7.52}$$

Now suppose that X, Y are arbitrary random variables. By Proposition 7.15, also $X_{(n)}$ and $Y_{(n)}$ are independent, and we have

$$|(XY)_{(n)} - X_{(n)}Y_{(n)}| \leq \frac{1}{n} + \frac{1}{n}\left(|X| + |Y| + \frac{1}{n}\right) \tag{7.53}$$

using the triangle inequality. By the discrete case we see that $\mathbf{E}[X_{(n)}Y_{(n)}] = \mathbf{E}[X_{(n)}]\mathbf{E}[Y_{(n)}]$. Using this in combination with (7.53) yields the claim. □

Example 7.19. Let X and Y be the side-lengths of a random rectangle. We assume that $X, Y \sim \mathcal{U}([0, 1])$ and X is independent of Y. The expected value of the area of the rectangle is

$$\mathbf{E}\left[XY\right] = \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \tag{7.54}$$

Clearly, independence is crucial for (7.51) to hold. Indeed, if we look at $\widetilde{Y} = X$, then

$$\mathbf{E}\left[X\widetilde{Y}\right] = \mathbf{E}\left[X^2\right] = \int_0^1 x^2 \mathrm{d}x = \frac{1}{3}. \tag{7.55}$$

The previous example is a special case of a general concept, that will be very important in the rest of the course.

[4] The assumption that $\mathbf{E}\left[X^2\right] < \infty$ and $\mathbf{E}\left[Y^2\right] < \infty$ is to guarantee the existence of $\mathbf{E}[|XY|]$, as we will later see using the Cauchy-Schwarz inequality. In fact due to the independence of X and Y, it is redundant in this case.

Definition 7.20. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables $X_1, X_2, \ldots : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is called *independent and identically distributed*, or *i.i.d.*, if $(X_i)_{i\in\mathbb{N}}$ are independent and for every $i, j \in \mathbb{N}$, $\mathbf{P}_{X_i} = \mathbf{P}_{X_j}$. Being i.i.d. is defined simularly for finitely many random variables $X_1, \ldots, X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Can we always find and i.i.d. sequence of random variables with a given distribution?

Theorem 7.21. *Let* $\mathbf{Q}$ *be a probability measure on* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. *Then there exists a probability space* $(\Omega, \mathscr{F}, \mathbf{P})$ *and a sequence* $(X_n)_{n\in\mathbb{N}}$ *of random variables* $X_1, X_2, \ldots : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ *that are i.i.d. with* $\mathbf{P}_{X_n} = \mathbf{Q}$ *for all* $n \in \mathbb{N}$.

Proof. See, e.g., [Geo12, Theorem 3.26]. □

In the next sections, we will study some important ways to combine independent random variables.

End of Lecture 16

7.3. Extremes

Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X_1, \ldots, X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be i.i.d. real random variables. We are interested in the distribution of the maximum

(7.56)
$$X^{[n]}(\omega) = \max\{X_1, \ldots, X_n\}(\omega) = \max\{X_1(\omega), \ldots, X_n(\omega)\}$$

and the minimum

(7.57)
$$X_{[n]}(\omega) = \min\{X_1, \ldots, X_n\}(\omega) = \min\{X_1(\omega), \ldots, X_n(\omega)\}.$$

Note that $X^{[n]}$ and $X_{[n]}$ are indeed random variables, by Remark 5.2, (i).

Proposition 7.22. *The distribution functions of* $X^{[n]}$ *and* $X_{[n]}$ *are given by*

(7.58)
$$\begin{aligned} F_{X^{[n]}}(x) &= \big(F_{X_1}(x)\big)^n, \\ F_{X_{[n]}}(x) &= 1 - \big(1 - F_{X_1}(x)\big)^n, \qquad x \in \mathbb{R}, \end{aligned}$$

respectively.

Proof. We have for $x \in \mathbb{R}$,

(7.59)
$$\begin{aligned} F_{X^{[n]}}(x) &= \mathbf{P}\big[\max\{X_1, \ldots, X_n\} \le x\big] = \mathbf{P}\left[\bigcap_{i=1}^{n} \{X_i \le x\}\right] \\ &= \prod_{i=1}^{n} \mathbf{P}\big[X_i \le x\big] \qquad \text{(by independence)} \\ &= \big(F_{X_1}(x)\big)^n. \end{aligned}$$

Similarly, we see that for $x \in \mathbb{R}$,

$$\begin{aligned} 1 - F_{X_{[n]}}(x) &= \mathbf{P}\big[\min\{X_1, ..., X_n\} > x\big] = \mathbf{P}\left[\bigcap_{i=1}^{n}\{X_i > x\}\right] \\ &= \prod_{i=1}^{n}\mathbf{P}\big[X_i > x\big] \qquad \text{(by independence)} \\ &= \big(1 - F_{X_1}(x)\big)^n. \end{aligned} \tag{7.60}$$

□

As an example, consider $X_1, ..., X_n \sim \mathscr{E}(\lambda)$ i.i.d. random variables. These may be interpreted as waiting times for independent, exponentially distributed events. What is the law of the maximum and minimum? We have

$$\begin{aligned} &F_{X^{[n]}}(x) = \big(F_{X_1}(x)\big)^n = \big(1 - e^{-\lambda x}\big)^n \mathbb{1}_{[0,\infty)}(x), \\ &F_{X_{[n]}}(x) = 1 - \big(1 - F_{X_1}(x)\big)^n = 1 - \big(1 - (1 - e^{-\lambda x})\mathbb{1}_{[0,\infty)}(x)\big)^n = \big(1 - e^{-\lambda n x}\big)\mathbb{1}_{[0,\infty)}(x). \end{aligned} \tag{7.61}$$

Here we used the expression for the cumulative distribution function of an exponentially distributed random variable (4.33). In particular, we see that $X_{[n]} \sim \mathscr{E}(\lambda n)$.

7.4. Sums of independent random variables

Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Assume that $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are independent real random variables with known distributions $\mathbf{P}_X$ and $\mathbf{P}_Y$. What is the distribution of $X + Y$? We will assume that X and Y are both discrete or both continuous. Let us start with the former (easier) case.

Proposition 7.23. *Assume that X, Y are independent discrete real random variables with probability mass functions $(p_X(k))_{k\in\Omega_X}$ and $(p_Y(\ell))_{\ell\in\Omega_Y}$. Then the sum $Z = X + Y$ is again discrete and its probability mass function is given by*

$$p_Z(k) = \sum_{\ell\in\Omega_Y} p_X(k-\ell)p_Y(\ell), \tag{7.62}$$

for $k \in \Omega_X + \Omega_Y = \{a + \ell \,:\, a \in \Omega_X, \ell \in \Omega_Y\}$.

Proof. Clearly, Z is discrete, since it takes only values in $\Omega_X + \Omega_Y$, which is countable. Now take $k \in \Omega_X + \Omega_Y$. Then (since $\bigcup_{\ell\in\Omega_Y}\{Y = \ell\} = \Omega$):

$$\begin{aligned} \mathbf{P}\big[Z = k\big] &= \mathbf{P}\left[\bigcup_{\ell\in\Omega_Y}\{X = k - \ell, Y = \ell\}\right] \\ &= \sum_{\ell\in\Omega_Y} \underbrace{\mathbf{P}\big[X = k - \ell\big]}_{=p_X(k-\ell)} \cdot \underbrace{\mathbf{P}\big[Y = \ell\big]}_{=p_Y(\ell)}, \end{aligned} \tag{7.63}$$

where we used the independence assumption in the second step. □

Example 7.24. (i) Let $X \sim Bin(n,p)$ and $Y \sim Bin(m,p)$ for $n,m \in \mathbb{N}$ and $p \in (0,1)$ be independent. Then $X+Y \sim Bin(n+m,p)$. Indeed, the random variable $X+Y$ can attain values in $\{0,...,n+m\}$, so let $0 \le k \le n+m$. Then

$$\begin{aligned} p_Z(k) &= \sum_{\ell=0}^{m} p_X(\ell)p_Y(k-\ell) \\ &= \sum_{\ell=0}^{k} \binom{n}{k-\ell} p^{k-\ell}(1-p)^{n-(k-\ell)} \binom{m}{\ell} p^{\ell}(1-p)^{m-\ell} \\ &= \sum_{\ell=0}^{k} \binom{n}{k-\ell}\binom{m}{\ell} p^{k}(1-p)^{n+m-k} \\ &= \binom{n+m}{k} p^{k}(1-p)^{n+m-k}. \end{aligned} \tag{7.64}$$

In the last line we used the *Vandermonde identity*

$$\binom{n+m}{k} = \sum_{\ell=0}^{k} \binom{n}{k-\ell}\binom{m}{\ell}, \qquad n,m,k \in \mathbb{N}_0. \tag{7.65}$$

There are various ways to prove this identity. A particularly simple proof goes as follows: Consider the polynomial functions $x \mapsto (1+x)^r$, for $r \in \mathbb{N}$. We see that

$$\begin{aligned} & (1+x)^{n+m} = (1+x)^n \cdot (1+x)^m \\ \Rightarrow \quad & \sum_{k=0}^{n+m} \binom{m+n}{k} x^k = \left(\sum_{i=0}^{n} \binom{n}{i} x^i\right)\left(\sum_{j=0}^{m} \binom{m}{j} x^j\right) \\ & \qquad = \sum_{k=0}^{m+n} \left(\sum_{\ell=0}^{k} \binom{n}{k-\ell}\binom{m}{\ell}\right) x^k. \end{aligned} \tag{7.66}$$

By extracting the coefficient k on both sides, we obtain (7.65).

(ii) Let $X \sim Pois(\lambda)$ and $Y \sim Pois(\mu)$, $\lambda, \mu > 0$ be independent. Then $X+Y \sim Pois(\lambda+\mu)$. We leave this as an Exercise.

In particular, part (i) shows that for i.i.d. $X_1,...,X_n \sim Ber(p)$, the random variable $S_n = \sum_{i=1}^{n} X_i$ fulfills $S_n \sim Bin(n,p)$. In particular, we see

$$\mathbf{E}\big[S_n\big] = n\mathbf{E}\big[X_1\big] = np, \tag{7.67}$$

where we used (6.40), and Theorem 6.9.

We now move to the sum of continuous, independent random variables.

Proposition 7.25. *Assume that X, Y are independent continuous real random variables with probability density functions f_X and f_Y. Then the sum $Z = X + Y$ is again continuous and its probability density function is given by*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \mathrm{d}y. \tag{7.68}$$

Proof. Let $z \in \mathbb{R}$. We have

$$\begin{aligned} F_Z(z) &= \mathbf{P}\big[X + Y \leq z\big] \\ &= \mathbf{P}\big[(X,Y) \in \{(x,y) \in \mathbb{R}^2 \,;\, x + y \leq z\}\big] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathbb{1}_{\{(x,y)\in\mathbb{R}^2\,;\,x+y\leq z\}} \mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f_{X,Y}(x,y)\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} f_X(x)\mathrm{d}x\right) f_Y(y)\mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_X(z-y) f_Y(y) \mathrm{d}y. \end{aligned} \tag{7.69}$$

We differentiate with respect to z on both sides to obtain

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \mathrm{d}y. \tag{7.70}$$

□

Remark 7.26. If X and Y are independent real random variables defined on $(\Omega, \mathscr{F}, \mathbf{P})$ with laws $\mathbf{P}_1 = \mathbf{P}_X$ and $\mathbf{P}_2 = \mathbf{P}_Y$, we say that the law $\mathbf{Q} = \mathbf{P}_{X+Y}$ of the sum of X and Y is the *convolution* of the laws of X and Y, and we write

$$\mathbf{Q} = \mathbf{P}_1 * \mathbf{P}_2 = \mathbf{P}_X * \mathbf{P}_Y. \tag{7.71}$$

For instance, one can show that if $X, Y \sim \mathscr{E}(\lambda)$ are independent, then $X + Y \sim \Gamma(2, \lambda)$ (the *Gamma*-distribution with parameters $\alpha = 2$ and λ, characterized by the density $f(x) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}\mathbb{1}_{[0,\infty)}(x)$ for $\lambda, \alpha > 0$, with $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$). Thus one has

$$\mathscr{E}(\lambda) * \mathscr{E}(\lambda) = \Gamma(2, \lambda). \tag{7.72}$$

For other distributions, one has (see Exercises and Example 7.24)

$$\begin{aligned} Bin(n,p) * Bin(m,p) &= Bin(n+m,p), \qquad n, m \in \mathbb{N}, p \in (0,1), \\ Pois(\lambda) * Pois(\mu) &= Pois(\lambda + \mu), \qquad \lambda, \mu > 0, \\ \mathscr{N}(\mu_1, \sigma_1^2) * \mathscr{N}(\mu_2, \sigma_2^2) &= \mathscr{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \qquad \mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0, \\ \Gamma(\alpha_1, \lambda) * \Gamma(\alpha_2, \lambda) &= \Gamma(\alpha_1 + \alpha_2, \lambda), \qquad \alpha_1, \alpha_2, \lambda > 0. \end{aligned} \tag{7.73}$$

End of Lecture 17

7.5. Covariance and correlation

As we have already seen, the expectation is additive. We want to find a similar expression for the variance of the sum of two random variables. This motivates the following definition.

Definition 7.27. Let X and Y be two real random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbf{P})$, fulfilling $\mathbf{E}\left[X^2\right] < \infty$ and $\mathbf{E}\left[Y^2\right] < \infty$.[5]

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

$$\operatorname{Cov}\left[X, Y\right] = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right]. \tag{7.74}$$

(ii) The *correlation coefficient* of X and Y is defined as

$$\rho(X, Y) = \begin{cases} \frac{\operatorname{Cov}\left[X,Y\right]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}}, & \operatorname{Var}[X] \neq 0, \operatorname{Var}[Y] \neq 0, \\ 0, & \text{else.} \end{cases} \tag{7.75}$$

We collect some properties of the covariance and correlation coefficient.

Proposition 7.28. *Let X and Y be two real random variables defined on the same probability space, fulfilling $\mathbf{E}\left[X^2\right] < \infty$ and $\mathbf{E}\left[Y^2\right] < \infty$.*

(i) The covariance can be written as

$$\operatorname{Cov}\left[X, Y\right] = \mathbf{E}\left[XY\right] - \mathbf{E}[X]\mathbf{E}[Y] = \operatorname{Cov}\left[Y, X\right]. \tag{7.76}$$

(ii)

$$\operatorname{Var}\left[X\right] = \operatorname{Cov}\left[X, X\right]. \tag{7.77}$$

(iii) If X and Y are independent, then $\operatorname{Cov}\left[X, Y\right] = \rho(X, Y) = 0$.

(iv) If $X = \pm Y$, then

$$\operatorname{Cov}\left[X, Y\right] = \pm\operatorname{Var}\left[X\right] \text{ and } \rho(X, Y) = \pm 1 \text{ (if } \operatorname{Var}[X] \neq 0 \text{ and } \operatorname{Var}[Y] \neq 0\text{)}. \tag{7.78}$$

Proof. For (i), we use

$$\begin{aligned} \operatorname{Cov}\left[X, Y\right] &= \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right] \\ &= \mathbf{E}\left[XY\right] - \mathbf{E}\left[X\mathbf{E}[Y]\right] - \mathbf{E}\left[Y\mathbf{E}[X]\right] + \mathbf{E}[X]\mathbf{E}[Y] \\ &= \mathbf{E}\left[XY\right] - \mathbf{E}[X]\mathbf{E}[Y] = \operatorname{Cov}\left[Y, X\right]. \end{aligned} \tag{7.79}$$

The claim (ii) follows directly from the definition of the variance (6.34).

For (iii), we see that

$$\operatorname{Cov}\left[X, Y\right] = \mathbf{E}\left[XY\right] - \mathbf{E}[X]\mathbf{E}[Y] \overset{(7.51)}{=} \mathbf{E}[X]\mathbf{E}[Y] - \mathbf{E}[X]\mathbf{E}[Y] = 0. \tag{7.80}$$

[5] We will show in the following Chapter that these guarantee the existence of $\mathbf{E}[|XY|]$.

(iv) The first part is obvious. Now

$$\rho(X, \pm X) = \frac{\pm \mathrm{Var}[X]}{\left(\sqrt{\mathrm{Var}[X]}\right)^2} \overset{\mathrm{Var}[X]>0}{=} \pm 1. \tag{7.81}$$

□

We now prove the bilinearity of the covariance.

Proposition 7.29. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ random variables on the same probability space, whith finite second moment. For $a, b, c_1, \ldots, c_n, d_1, \ldots, d_m \in \mathbb{R}$ one has*

$$\mathrm{Cov}\left[a + \sum_{i=1}^{n} c_i X_i, b + \sum_{j=1}^{m} d_j Y_j\right] = \sum_{i=1}^{n} \sum_{j=1}^{m} c_i d_j \mathrm{Cov}[X_i, Y_j]. \tag{7.82}$$

Proof. We use (7.76) to see that for every random variables Z_1, Z_2, Z_3 with finite second moment:

$$\begin{aligned} \mathrm{Cov}[Z_1 + Z_2, Z_3] &= \mathbf{E}\big[(Z_1 + Z_2) Z_3\big] - \mathbf{E}\big[Z_1 + Z_2\big]\mathbf{E}\big[Z_3\big] \\ &= \mathbf{E}\big[Z_1 Z_3\big] + \mathbf{E}\big[Z_2 Z_3\big] - \mathbf{E}[Z_1]\mathbf{E}[Z_3] - \mathbf{E}[Z_2]\mathbf{E}[Z_3] \\ &= \mathrm{Cov}\big[Z_1, Z_3\big] + \mathrm{Cov}\big[Z_2, Z_3\big]. \end{aligned} \tag{7.83}$$

Moreover, for $\lambda \in \mathbb{R}$:

$$\begin{aligned} \mathrm{Cov}\big[\lambda Z_1, Z_2\big] &= \mathbf{E}\big[\lambda Z_1 Z_2\big] - \mathbf{E}[\lambda Z_1]\mathbf{E}[Z_2] \\ &= \lambda \mathrm{Cov}\big[Z_1, Z_2\big]. \end{aligned} \tag{7.84}$$

Finally, we have

$$\mathrm{Cov}\big[\lambda, Z_1\big] = \mathbf{E}\big[\lambda Z_1\big] - \mathbf{E}[\lambda]\mathbf{E}[Z_1] = 0. \tag{7.85}$$

Equation (7.82) follows from the previous three displays and the symmetry of Cov. □

A special case of this is the *Bienaymé formula* for the variance of the sum of random variables.

Corollary 7.30. *Let $X_1, \ldots, X_n$ be real random variables defined on the same probability space, with finite second moment.*

(i)

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i,j=1}^{n} \mathrm{Cov}[X_i, X_j] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{\substack{i=1 \\ i \neq j}}^{n} \mathrm{Cov}[X_i, X_j]. \tag{7.86}$$

(ii) If $X_1, \ldots, X_n$ are uncorrelated (i.e. $\mathrm{Cov}[X_i, X_j] = 0$ *for every $1 \le i \neq j \le n$), then*

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i]. \tag{7.87}$$

Let us combine part (ii) with Example 7.24, (i): Since for i.i.d. $X_1, ..., X_n \sim Ber(p)$, the random variable $S_n = \sum_{i=1}^n X_i$ fulfills $S_n \sim Bin(n, p)$, we have

$$\mathrm{Var}\big[S_n\big] = n\mathrm{Var}\big[X_1\big] = np(1-p), \tag{7.88}$$

where we used (6.40).

Theorem 7.31. *Let X, Y be real random variables defined on some probability space with $\mathbf{E}\big[X^2\big], \mathbf{E}\big[Y^2\big] < \infty$.*

(i) $|\rho(X,Y)| \leq 1$.

(ii) $\rho(X,Y) = \pm 1$ *if and only if there are* $a, b \in \mathbb{R}$, $a \neq 0$ *with*

$$\mathbf{P}\big[Y = aX + b\big] = 1, \tag{7.89}$$

and we have

$$\begin{aligned} &a > 0, && \textit{if}\, \rho(X,Y) = 1, \\ &a < 0, && \textit{if}\, \rho(X,Y) = -1. \end{aligned} \tag{7.90}$$

Proof. We first show (i). The variance of any random variables is nonnegative, so

$$\begin{aligned} 0 &\leq \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}} + \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) \\ &= \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}\right) + \mathrm{Var}\left(\frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) + 2\mathrm{Cov}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}, \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) \\ &= \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} + \frac{\mathrm{Var}(Y)}{\mathrm{Var}(Y)} + \frac{2\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = 2(1 + \rho(X,Y)), \end{aligned} \tag{7.91}$$

where we used Corollary 7.30, (i). We see that $\rho(X,Y) \geq -1$. By replacing X by $-X$, we also have $\rho(X,Y) \leq 1$.

For (ii), we first assume that $\rho(X,Y) = -1$. Then, by (7.91),

$$\mathbf{P}\left[\frac{X}{\sqrt{\mathrm{Var}(X)}} + \frac{Y}{\sqrt{\mathrm{Var}(Y)}} = c\right] = 1, \tag{7.92}$$

where we used Corollary 6.15. In other words, we have

$$\mathbf{P}[Y = aX + b] = 1, \qquad a = -\frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{\mathrm{Var}(X)}} < 0. \tag{7.93}$$

If $\rho(X,Y) = 1$, we can do the same calculation but with $\mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}} - \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right) = ... = 0$. □

Definition 7.32. Let $X_1, ..., X_n$ be random variables with finite second moment, defined on the same probability space. The matrix

$$\Sigma = \Sigma(X) = (\text{Cov}(X_i, X_j))_{i,j=1}^n \tag{7.94}$$

is called the *covariance matrix* of the random vector $X = (X_1, ..., X_n)$. For $u \in \mathbb{R}^n$, one has

$$\text{Var}\,(u \cdot X) = u \cdot \Sigma u. \tag{7.95}$$

Here for $u, v \in \mathbb{R}^n$, $u \cdot v = \sum_{i=1}^d u_i v_i$ denotes the standard scalar product in $\mathbb{R}^n$.

Note that the matrix Σ is symmetric and non-negative definite.

8. Jensen and Hölder inequalities

(Reference: [Geo12, Section 5.6])

In this short section, we state two useful inequalities that are relevant for applications.

8.1. Jensen's inequality

In this subsection, we will show a useful inequality for the expectation. We start with a reminder on convexity.

Definition 8.1. Consider $\varphi : \mathbb{R} \to \mathbb{R}$. We say that φ is *convex* if for every $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$, one has

$$\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y). \tag{8.1}$$

We say that φ is *concave* if $-\varphi$ is convex.

In particular, if φ is twice differentiable, it is convex if and only if $\varphi''(x) \geq 0$ for every $x \in \mathbb{R}$.

Example 8.2. The functions $\varphi_j : \mathbb{R} \to \mathbb{R}$ with $1 \leq j \leq 3$ with $\varphi_1(x) = x^2$, $\varphi_2(x) = e^{rx}$ for $r > 0$, $\varphi_3(x) = |x|$ are convex.

We now state *Jensen's inequality.*

Lemma 8.3. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and X a real random variable with $\mathbf{E}[|X|] < \infty$ and $\mathbf{E}[|\varphi(X)|] < \infty$. Then*

$$\mathbf{E}\big[\varphi(X)\big] \geq \varphi\big(\mathbf{E}[X]\big). \tag{8.2}$$

End of Lecture 18

Proof. We set

$$\mathscr{E}_\varphi = \{(a,b) \in \mathbb{R}^2 \,:\, \varphi(x) \geq ax + b\}. \tag{8.3}$$

By elementary properties of convex functions, we have that

$$\varphi(x) = \sup_{(a,b)\in\mathscr{E}_\varphi} (ax+b). \tag{8.4}$$

Since $\varphi(X) \geq aX + b$ for every $(a,b) \in \mathscr{E}_\varphi$, we obtain

$$\mathbf{E}[\varphi(X)] \geq \sup_{(a,b)\in\mathscr{E}_\varphi} \mathbf{E}[aX+b] = \sup_{(a,b)\in\mathscr{E}_\varphi} (a\mathbf{E}[X]+b) = \varphi(\mathbf{E}[X]), \tag{8.5}$$

which gives us the claim. □

A special case is the function $\varphi(x) = x^2$, where one has the (known) inequality

$$\mathbf{E}\big[X^2\big] \geq \big(\mathbf{E}[X]\big)^2. \tag{8.6}$$

8.2. Hölder's inequality

Suppose that $X, Y : (\Omega, \mathscr{F}, \mathbf{P}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are real random variables. By considering the example that $X = Y$ with law on $\mathbb{N}$ given by

$$\mathbf{P}[X = n] = \frac{1}{c_3} \cdot \frac{1}{n^3}, \qquad c_3 = \sum_{n=1}^{\infty} \frac{1}{n^3} \simeq 1.202057..., \tag{8.7}$$

we see that even though $\mathbf{E}[X] = \mathbf{E}[Y]$ exist, the expectation of $\mathbf{E}[X \cdot Y] = \mathbf{E}[X^2]$ is infinite. Are there situations where we can conclude that $X \cdot Y$ has a finite expectation, based on information on the moments of X and Y? It turns out that this is the case, and we have the *Hölder inequality*:

Theorem 8.4. *Let $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Let X, Y be two real random variables, defined on the same probability space and assume that the expectations of $|X|^p$ and $|Y|^q$ exist. Then $\mathbf{E}[|XY|] < \infty$ and*

$$\mathbf{E}\big[|X \cdot Y|\big] \le \big(\mathbf{E}\big[|X|^p\big]\big)^{\frac{1}{p}} \cdot \big(\mathbf{E}\big[|Y|^q\big]\big)^{\frac{1}{q}}. \tag{8.8}$$

Proof. We first show that for all $a, b \ge 0$, one has

$$ab \le \frac{a^p}{p} + \frac{b^q}{q} \qquad \textit{(Young's inequality).} \tag{8.9}$$

Indeed, fix $b > 0$ (the inequality is trivial if $a = 0$ or $b = 0$) and set $f(x) = \frac{x^p}{p} + \frac{b^q}{q} - xb$, then f is twice differentiable, and

$$f'(x) = x^{p-1} - b, \qquad f''(x) = (p-1)x^{p-2}. \tag{8.10}$$

In particular, f is attains its (unique) minimum at $x_0 = b^{\frac{1}{p-1}}$. Since $q = \frac{p}{p-1}$, $x_0^p = b^q$, so

$$f(x_0) = \left(\frac{1}{p} + \frac{1}{q}\right) b^q - b^{\frac{1}{p-1}} b = 0, \tag{8.11}$$

showing (8.9). Suppose now that the expectations of $|X|^p$ and $|Y|^q$ are not zero (otherwise the claim becomes trivial). We can then apply (8.9) to

$$a = \frac{|X(\omega)|}{\big(\mathbf{E}\big[|X|^p\big]\big)^{\frac{1}{p}}}, \qquad b = \frac{|Y(\omega)|}{\big(\mathbf{E}\big[|Y|^q\big]\big)^{\frac{1}{q}}}. \tag{8.12}$$

This yields

$$\frac{|X(\omega)Y(\omega)|}{\big(\mathbf{E}\big[|X|^p\big]\big)^{\frac{1}{p}} \cdot \big(\mathbf{E}\big[|Y|^q\big]\big)^{\frac{1}{q}}} \le \frac{|X(\omega)|^p}{p\mathbf{E}\big[|X|^p\big]} + \frac{|Y(\omega)|^q}{q\mathbf{E}\big[|Y|^q\big]}. \tag{8.13}$$

In particular, the expectation of $|XY|$ exists, and the inequality (8.8) follows from taking the expectation and using that $\frac{1}{p} + \frac{1}{q} = 1$. □

If $p = q = 2$, the special case of the Hölder inequality is called the *Cauchy-Schwarz inequality*, and states that if $\mathbf{E}[X^2] < \infty$ and $\mathbf{E}[Y^2] < \infty$, then

$$\mathbf{E}\big[|X \cdot Y|\big] \leq \sqrt{\mathbf{E}\big[X^2\big]} \cdot \sqrt{\mathbf{E}\big[Y^2\big]}. \tag{8.14}$$

We have used this implicitly in the proof of Theorems 7.18 and in the definition of the covariance, see Definition 7.27. In fact, we also have the bound

$$\mathrm{Cov}\big[X, Y\big] \leq \sqrt{\mathrm{Var}[X]} \cdot \sqrt{\mathrm{Var}[Y]}, \tag{8.15}$$

which follows from (8.14) and the fact that $|\mathbf{E}[Z]| \leq \mathbf{E}[|Z|]$ by replacing X and Y by $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$, respectively.

9. Conditional distributions and conditional expectation

(Reference: [GS01, Sections 3.7, 4.6])

In this chapter, we define conditional distributions and conditional expectations of random variables. Recall that we already discussed the *joint* distribution of random variables X and Y. A natural question is then:

What is the distribution of X, if the value of Y is known?

For instance, we may think of the following:

- X is the value of the first outcome when rolling a fair die three times, Y is the sum of outcomes, what is $\mathbf{P}[X = x|Y = y]$?
- X is the length of a randomly chosen fish of species A, Y is its age. What is $\mathbf{P}[X \in B|Y = y]$, for $B \in \mathscr{B}(\mathbb{R})$?

9.1. Discrete conditional distributions

Suppose that X, Y are two discrete real random variables with joint probability mass function $p_{X,Y}(x, y)$ for $x \in \Omega_X$, $y \in \Omega_Y$. Clearly, one has

$$\mathbf{P}[X = x|Y = y] = \frac{\mathbf{P}[X = x, Y = y]}{\mathbf{P}[Y = y]} = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \tag{9.1}$$

if $p_Y(y) > 0$. This leads to the following fact.

Proposition 9.1. *Let $(\Omega, \mathscr{F}, \boldsymbol{P})$ be a probability space and $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ two discrete real random variables with probability mass functions $(p_X(x))_{x\in\Omega_X}$ and $(p_Y(y))_{y\in\Omega_Y}$ and joint probability mass function $p_{X,Y}(x, y))_{(x,y)\in\Omega_X\times\Omega_Y}$. We define*

$$p_{X|Y=y}(x) = \begin{cases} \frac{p_{X,Y}(x,y)}{p_Y(y)}, & \textit{if } p_Y(y) > 0, \\ 0, & \textit{else.} \end{cases} \tag{9.2}$$

Then $p_{X|Y=y}(x) \geq 0$, $\sum_{x\in\Omega_X} p_{X|Y=y}(x) = 1$ for any $y \in \Omega_Y$ with $p_Y(y) > 0$, and

$$p_X(x) = \sum_{y\,;\,p_Y(y)>0} p_{X|Y=y}(x)p_Y(y). \tag{9.3}$$

Proof. This all follows from the fact that $\mathbf{P}[\,\cdot\,|\{Y = y\}]$ is a probability measure as long as $\mathbf{P}[Y = y] = p_Y(y) > 0$ (see Proposition 2.4), and the law of X under $\mathbf{P}[\,\cdot\,|\{Y = y\}]$ is discrete with probability mass function $\mathbf{P}[X = x|Y = y] = \frac{p_{X,Y}(x,y)}{p_Y(y)}$. Equation (9.3) follows from the law of total probability. □

We say that $\mathbf{P}[\,\cdot\,|Y = y]_X$ is the *conditional distribution / law of X given $Y = y$*, and its probability mass function $p_{X|Y=y}$ is the *conditional probability mass function of (the law of) X given $Y = y$*. We also record that if X and Y are independent, then

$$p_{X|Y=y}(x) = p_X(x) \qquad \text{for all } x \in \Omega_X, p_Y(y) > 0. \tag{9.4}$$

Example 9.2. (i) Recall the example of tossing a coin three times from (7.2) in Chapter 7, i.e.

$$\begin{aligned} \Omega &= \{0,1\}^3 = \{(\omega_1,\omega_2,\omega_3)\,;\,\omega_i \in \{0,1\}\}, \\ X(\omega) &= \omega_1, \qquad \Omega_X = \{0,1\}, \\ Y(\omega) &= \sum_{i=1}^{3} \omega_i, \qquad \Omega_Y = \{0,1,2,3\}. \end{aligned} \tag{9.5}$$

We found for the joint probability mass function:

$x_i \setminus y_j$	0	1	2	3	$p_X(x_i)$
0	1/8	2/8	1/8	0	1/2
1	0	1/8	2/8	1/8	1/2
$p_Y(y_j)$	1/8	3/8	3/8	1/8	1

With this, we have for instance:

$$\begin{aligned} p_{X|Y=1}(0) &= \frac{p_{X,Y}(0,1)}{p_Y(1)} = \frac{2/8}{2/8+1/8} = \frac{2}{3}, \\ p_{X|Y=1}(1) &= \frac{p_{X,Y}(1,1)}{p_Y(1)} = \frac{1/8}{2/8+1/8} = \frac{1}{3}. \end{aligned}$$

(ii) Suppose that X and Y are independent Poisson random variables with parameters $\lambda > 0$ and $\mu > 0$, respectively. We want to calculate the conditional distribution of X given $X + Y = n$.

$$\begin{aligned} \mathbf{P}[X = k|X + Y = n] &= \frac{\mathbf{P}[X = k]\mathbf{P}[Y = n-k]}{\mathbf{P}[X + Y = n]} \\ &= \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-\mu}\mu^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-(\lambda+\mu)}(\lambda+\mu)^n} \\ &= \binom{n}{k} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^k \cdot \left(\frac{\mu}{\lambda+\mu}\right)^{n-k}. \end{aligned}$$

Here we used that $X + Y \sim Pois(\lambda + \mu)$, see Example 7.24, (ii). In other words, $X \sim Bin(n, \frac{\lambda}{\lambda+\mu})$ under $\mathbf{P}[\,\cdot\,|X + Y = n]$. This property is known as *splitting of Poisson random variables.*

9.2. Continuous conditional distributions

We wish to extend the discussion of the previous section to the case where X, Y are jointly continuous real random variables with joint density $f_{X,Y}$. Unfortunately, $\mathbf{P}[Y = y] = 0$, and we cannot really "condition" on the event $\{Y = y\}$. However, the following heuristics still gives a valuable definition of a conditional density: Suppose $a < b$ and $f_Y(y) > 0$, then

$$\begin{aligned}\mathbf{P}[X \in [a,b)|Y = y] &\approx \mathbf{P}[X \in [a,b)|Y \in [y, y+\Delta y)], \qquad \Delta y \text{ small} \\ &= \frac{\mathbf{P}[X \in [a,b), Y \in [y, y+\Delta y)]}{\mathbf{P}[Y \in [y, y+\Delta y)]} \\ &= \frac{\int_a^b \int_y^{y+\Delta y} f_{X,Y}(x,z)\mathrm{d}z\mathrm{d}x}{\int_{y+\Delta y} f_Y(z)\mathrm{d}z} \\ &\approx \frac{\int_a^b f_{X,Y}(x,y)\Delta y \mathrm{d}x}{f_Y(y)\Delta y} = \frac{\int_a^b f_{X,Y}(x,y)\mathrm{d}x}{f_Y(y)}.\end{aligned}$$

This suggests that we may define the conditional density of X given $Y = y$ as a quotient of the joint density $f_{X,Y}$ and the marginal density f_Y.

Proposition 9.3. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ two jointly continous real random variables joint probability density function $f_{X,Y}$. We define*

$$f_{X|Y=y}(x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)}, & \textit{if } f_Y(y) > 0, \\ 0, & \textit{else.} \end{cases} \tag{9.6}$$

Then $f_{X|Y=y}(x) \geq 0$, $\int_{-\infty}^{\infty} f_{X|Y=y}(x) = 1$ for any $y \in \mathbb{R}$ with $f_Y(y) > 0$, and

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y=y}(x) f_Y(y)\mathrm{d}y. \tag{9.7}$$

The quantity $f_{X|Y=y}$ is called the *conditional probability density funcion of (the law of) X given $Y = y$.*

End of Lecture 19

9.3. Conditional expectation

We have seen in the previous section how to define the conditional probability mass function $p_{X|Y=y}$ or the conditional probability density function $f_{X|Y=y}$. We can also calculate expectations with these quantities.

Definition 9.4. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ two real random variables.

(i) Suppose both X and Y are discrete with values in Ω_X and Ω_Y and joint probability mass function $p_{X,Y}$. For any $y \in \Omega_Y$ with $p_Y(y) > 0$, we define the *conditional expectation of*

X given $Y = y$ as

$$\mathbf{E}[X|Y = y] = \sum_{x \in \Omega_X} x \cdot p_{X|Y=y}(x), \tag{9.8}$$

if $\sum_{x \in \Omega_X} |x| \cdot p_{X|Y=y}(x) < \infty$.

(ii) Suppose that X, Y are jointly continuous with joint probability density function $f_{X,Y}$. For $y \in \mathbb{R}$ with $f_Y(y) > 0$, we define the *conditional expectation of X given $Y = y$* as

$$\mathbf{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x), \tag{9.9}$$

if $\int_{-\infty}^{\infty} |x| \cdot f_{X|Y=y}(x) < \infty$.

Remark 9.5. In the discrete case, the definition of the conditional expectation of X given $Y = y$ is exactly the expectation of X under the conditional probability measure $\mathbf{P}[\,\cdot\,|Y = y]$, since the the distribution of X under this probability measure $(\mathbf{P}[\,\cdot\,|Y = y])_X$ has probability mass function $p_{X|Y=y}(x)$.

Example 9.6. Suppose that the joint probability density function of X and Y is given by

$$f_{X,Y}(x, y) = \frac{e^{-\frac{x}{y}} e^{-y}}{y} \mathbb{1}_{(0,\infty)^2}(x, y). \tag{9.10}$$

We want to calculate $\mathbf{E}[X|Y = y]$ and start by computing the conditional density. For $x, y > 0$:

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{f_{X,Y}(x, y)}{f_Y(y)} \\ &= \frac{(1/y)e^{-\frac{x}{y}} e^{-y}}{\int_0^{\infty} (1/y)e^{-\frac{x}{y}} e^{-y} \mathrm{d}x} = \frac{1}{y} e^{-\frac{x}{y}}. \end{aligned} \tag{9.11}$$

So the conditional distribution of X, given $Y = y$ is the exponential distribution $\mathscr{E}(\frac{1}{y})$, and thus

$$\mathbf{E}[X|Y = y] = y. \tag{9.12}$$

Note that in the discrete case, the set of $\omega \in \Omega$ for which $p_y(Y(\omega)) = 0$ has probability zero. We then set

$$\mathbf{E}[X|Y](\omega) = \sum_{y \in \Omega_Y, p_Y(y) > 0} \mathbf{E}[X|Y = y] \mathbb{1}_{\{y = Y(\omega)\}}. \tag{9.13}$$

With this definition, $\mathbf{E}[X|Y] : \Omega \to \mathbb{R}$ is also a random variable. A similar argument can be performed in the continuous case (requiring some measure theory).

Theorem 9.7. *Let $X, Y, Z : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be real random variables. Assume that $\mathbf{E}[|X|] < \infty$ and $\mathbf{E}[|Y|] < \infty$. Then* *(i)* $\mathbf{E}\big[\mathbf{E}[X|Y]\big] = \mathbf{E}[X]$*,*

(ii) X, Y *stochastically independent, then* $\mathbf{E}[X|Y] = \mathbf{E}[X]$*,* [1]

(iii) $\mathbf{E}[Xh(Y)|Y] = h(Y)\mathbf{E}[X|Y]$ *for any (measurable) function* h*, in particular*

- $\mathbf{E}[\mathit{const.}|Y] = \mathit{const.}$,
- $\mathbf{E}[h(Y)|Y] = h(Y)$,

(iv) $\mathbf{E}[\alpha X + \beta Y|Z] = \alpha\mathbf{E}[X|Z] + \beta\mathbf{E}[Y|Z]$ *(linearity) for* $\alpha, \beta \in \mathbb{R}$*.*

Proof. We only consider the case where X and Y are both discrete.

(i) We calculate

$$\begin{aligned}\mathbf{E}\big[\mathbf{E}[X|Y]\big] &= \sum_{y\in\Omega_Y, p_Y(y)>0} \mathbf{E}[X|Y=y]p_Y(y)\\ &= \sum_{y\in\Omega_Y, p_Y(y)>0}\sum_{x\in\Omega_X} x\,\underbrace{p_{X|Y=y}(x)p_Y(y)}_{=p_{X,Y}(x,y)} = \mathbf{E}[X].\end{aligned} \tag{9.14}$$

(ii) Note that we have

$$\mathbf{E}[X|Y=y] = \sum_{x\in\Omega_X} x\,\underbrace{p_{X|Y=y}(x)}_{=p_X(x)} = \mathbf{E}[X]. \tag{9.15}$$

(iii) Suppose that $\omega \in \Omega$ and $y = Y(\omega)$. Then we have (for $p_Y(y) > 0$):

$$\begin{aligned}\mathbf{E}\big[X\cdot h(Y)|Y\big](\omega) &= \mathbf{E}[X\cdot h(Y)|Y=y] = h(y)\mathbf{E}[X|Y=y]\\ &= h(Y(\omega))\mathbf{E}[X|Y](\omega).\end{aligned} \tag{9.16}$$

(iv) Follows from a calculation.

□

End of Lecture 20

As an application, we discuss *random sums.*

Example 9.8. Suppose that in an insurance company, a random (integer) number N of claims is made during a year. We assume that the size of the claims form an i.i.d. sequence $(X_n)_{n\in\mathbb{N}}$ of (non-negative) real random variables which we assume to be independent from N. The total amount of claims to the company is

$$S = \sum_{j=1}^{N} X_j. \tag{9.17}$$

[1] Apart from a set N with $\mathbf{P}[N] = 0$.

Note that this is really a random variable of the form

$$\omega \mapsto S(\omega) = \sum_{j=1}^{N(\omega)} X_j(\omega) = \sum_{j=1}^{\infty} X_j(\omega)\mathbb{1}_{\{N(\omega)\geq j\}}. \tag{9.18}$$

We are interested in $\mathbf{E}[S]$ and $\text{Var}[S]$ and assume also that $\mathbf{E}[X_1^2] < \infty$ and $\mathbf{E}[N^2] < \infty$.

$$\begin{aligned} \mathbf{E}[S|N = n] = \mathbf{E}\left[\sum_{j=1}^{N} X_j \middle| N = n\right] &= \sum_{j=1}^{n} \mathbf{E}[X_j] = n\mathbf{E}[X_1], \\ \Rightarrow \qquad \mathbf{E}[S|N] &= N\mathbf{E}[X_1] \\ \Rightarrow \qquad \mathbf{E}[S] &= \mathbf{E}\big[\mathbf{E}[S|N]\big] = \mathbf{E}[N]\mathbf{E}[X_1]. \end{aligned} \tag{9.19}$$

Moreover, we have

$$\begin{aligned} \mathbf{E}[S^2|N = n] = \mathbf{E}\left[\left(\sum_{j=1}^{N} X_j\right)^2 \middle| N = n\right] &= \mathbf{E}\left[\left(\sum_{j=1}^{n} X_j\right)^2\right] \\ &= \text{Var}\left[\sum_{j=1}^{n} X_j\right] + \left(\mathbf{E}\left[\sum_{j=1}^{n} X_j\right]\right)^2 \\ &= n\text{Var}[X_1] + n^2(\mathbf{E}[X_1])^2. \end{aligned} \tag{9.20}$$

From here, it follows that

$$\begin{aligned} \mathbf{E}[S^2|N] &= N\text{Var}[X_1] + N^2(\mathbf{E}[X_1])^2 \\ \Rightarrow \qquad \mathbf{E}[S^2] &= \mathbf{E}[N]\text{Var}[X_1] + \mathbf{E}[N^2](\mathbf{E}[X_1])^2 \\ \Rightarrow \qquad \text{Var}[S] &= \mathbf{E}[S^2] - (\mathbf{E}[S])^2 = \mathbf{E}[N]\text{Var}[X_1] + \text{Var}[N](\mathbf{E}[X_1])^2. \end{aligned} \tag{9.21}$$

Compare this to the variance of a *fixed* sum $\widetilde{S} = \sum_{j=1}^{n} X_j$, which is $\text{Var}[\widetilde{S}] = n\text{Var}[X_1]$.

The conditional expectation $\mathbf{E}[Y|X]$ gives the "average" over many independent realizations of Y, when the value X is given. We can make this more rigorous, and discuss as a final part of this section the *best prediction / best linear prediction* of random variables.

Lemma 9.9. *Suppose X, Y are real random variables, defined on the same probability space and $\mathbf{E}[X^2] < \infty$, $\mathbf{E}[Y^2] < \infty$.*

(i) The conditional expectation $\mathbf{E}[Y|X] =: h_*(X)$ minimizes the expression $\mathbf{E}\left[(Y - h(X))^2\right]$ among all $h : \mathbb{R} \to \mathbb{R}$ (measurable) such that $\mathbf{E}[h(X)^2] < \infty$.

(ii) The values $a, b \in \mathbb{R}$ that minimize $\mathbf{E}\left[(Y - (a + bX)^2\right]$ are given by

$$b = \frac{\text{Cov}[X,Y]}{\text{Var}[X]}, \qquad a = \mathbf{E}[Y] - \frac{\text{Cov}[X,Y]}{\text{Var}[X]}\mathbf{E}[X]. \tag{9.22}$$

Proof. (i) is given as an exercise. For (ii), consider

$$\begin{aligned}\mathbf{E}[(Y-a-bX)^2] &= \mathrm{Var}[Y-a-bX] + (\mathbf{E}[Y]-a-b\mathbf{E}[X])^2\\ &= \mathrm{Var}[Y-bX] + (\mathbf{E}[Y]-a-b\mathbf{E}[X])^2.\end{aligned}$$

Both terms are non-negative, so we see immediately (by minimizing the second term) that

$$a = \mathbf{E}[Y] - b\mathbf{E}[X]. \tag{9.23}$$

On the other hand:

$$\begin{aligned}\mathrm{Var}[Y-bX] &= \mathrm{Var}[Y] + b^2\mathrm{Var}[X] - 2b\mathrm{Cov}[X,Y]\\ \frac{\partial}{\partial b}\mathrm{Var}[Y-bX] &= 2b\mathrm{Var}[X] - 2\mathrm{Cov}[X,Y] = 0 \qquad \Leftrightarrow \qquad b = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}.\end{aligned}$$

This concludes the proof. □

We also remark that the *mean square error* for these optimal values of a and b is given by

$$\mathbf{E}[(Y-a-bX)^2] = \mathrm{Var}[Y] - \frac{\mathrm{Cov}[X,Y]^2}{\mathrm{Var}[X]} = \mathrm{Var}[Y](1-\rho(X,Y)^2). \tag{9.24}$$

Note that $\rho(X,Y)$ close to ± 1 means that we typically only make a small error when approximating Y as a linear function of X.

Lemma 9.9 thus tells us: if we want to *predict* the value of Y, *knowing* the value of X:

- The best way in general to do it is to choose $\mathbf{E}[Y|X]$,
- The best way to do it with a linear function is $a + bX$ with a, b given in (9.22).

Example 9.10. Let $Y = X^2 + X + Z$ with $X, Z \sim \mathcal{N}(0,1)$ i.i.d., suppose we want to find a good prediction of Y knowing X.

- The best prediction is given by the conditional expectation:

  $$\mathbf{E}[Y|X] = X^2 + X + \mathbf{E}[Z|X] = X^2 + X,$$

  and the mean square error is

  $$\mathbf{E}[(Y-X^2-X)^2] = \mathbf{E}[Z^2] = 1.$$

- The best *linear* prediction of Y is given by $a + bX$ with

  $$\begin{aligned}b &= \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} = \mathrm{Cov}[X, X^2+X] = \mathrm{Cov}[X,X^2] + \mathrm{Var}[X] = 1,\\ a &= \mathbf{E}[Y] - \mathbf{E}[X] = \mathbf{E}[X^2] = 1,\end{aligned}$$

  so $1 + X$ is the best linear prediction. Its mean square error is

  $$\begin{aligned}\mathbf{E}\left[(Y-(1+X))^2\right] &= \mathbf{E}[(X^2+Z-1)^2]\\ &= \mathbf{E}[X^4] + 2\mathbf{E}[X^2Z] + \mathbf{E}[Z^2] - 2\mathbf{E}[X^2] - 2\mathbf{E}[Z] + 1 = 3.\end{aligned}$$

10. Generating functions

(Reference: [GS01, Sections 5.1], or [Geo12, Section 4.4])

In this chapter, we will define three functions which characterize probability distributions $\mathbf{P}_X$ of a real random variable X and are also helpful in the calculation of moments.

Definition 10.1. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ a real random variable. We define the *moment generating function* ψ_X of X by

(10.1)
$$\psi_X(t) = \mathbf{E}\big[e^{tX}\big] = \begin{cases} \sum_{k \in \Omega_X} e^{tk}\mathbf{P}[X = k], & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) \mathrm{d}x, & \text{if } X \text{ is continuous and its law has density } f_X, \end{cases}$$

for $t \in \mathbb{R}$, if this expectation exists.

Note that $\psi_X(t)$ always exists for $t = 0$, but may not exist for general $t \neq 0$. Here are some properties of moment generating functions:

Lemma 10.2. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X, Y : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ independent real random variables, and $\lambda, \mu \in \mathbb{R}$.*

(i) Assume that $\psi_X(t)$, $\psi_Y(t)$ and $\psi_{X+Y}(t)$ exist for some $t \in \mathbb{R}$, then

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t). \tag{10.2}$$

(ii) Assume that $\psi_X(\lambda t)$ exists, then also $\psi_{\lambda X+\mu}(t)$ exists and

$$\psi_{\lambda X+\mu}(t) = e^{\mu t}\psi_X(\lambda t). \tag{10.3}$$

(iii) Suppose that $\psi_X(t)$ exists for all $t \in (-\varepsilon, \varepsilon)$, $\varepsilon > 0$. Then

$$\psi_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\psi_X(t)\bigg|_{t=0} = \mathbf{E}[X^n]. \tag{10.4}$$

Proof. (i) We have

$$\psi_{X+Y}(t) = \mathbf{E}[e^{t(X+Y)}] = \mathbf{E}[e^{tX}e^{tY}] \overset{(7.52)}{=} \mathbf{E}[e^{tX}]\mathbf{E}[e^{tY}] = \psi_X(t)\psi_Y(t), \tag{10.5}$$

where we used that since X and Y are independent, so are e^{tX} and e^{tY} (see Proposition 7.15).

(ii) We see that

$$\psi_{\lambda X+\mu}(t) = \mathbf{E}[e^{\lambda tX}e^{\mu t}] = e^{\mu t}\psi_X(\lambda t). \tag{10.6}$$

(iii) Let X be discrete, and we assume that we can interchange differentiation and summation[1]:

$$\begin{aligned}\psi_X^{(n)}(0) &= \sum_{k\in\Omega_X} \frac{\mathrm{d}^n}{\mathrm{d}t^n} e^{tk}\mathbf{P}[X=k]\Big|_{t=0} \\ &= \sum_{k\in\Omega_X} k^n e^{tk}\mathbf{P}[X=k]\Big|_{t=0} = \sum_{k\in\Omega_X} k^n\mathbf{P}[X=k] = \mathbf{E}[X^n].\end{aligned} \tag{10.7}$$

If X is continous, one has (interchanging integration and differentiation)

$$\begin{aligned}\psi_X^{(n)}(0) &= \int_{-\infty}^{\infty} \frac{\mathrm{d}^n}{\mathrm{d}t^n} e^{tx} f_X(x)\mathrm{d}x\Big|_{t=0} \\ &= \int_{-\infty}^{\infty} x^n e^{tx} f_X(x)\mathrm{d}x\Big|_{t=0} = \int_{-\infty}^{\infty} x^n f_X(x)\mathrm{d}x = \mathbf{E}[X^n].\end{aligned} \tag{10.8}$$

□

The third property justifies the name moment generating function. Note that the assumption in part (i) the above Lemma can typically be justified by Hölder's inequality: Indeed, if $\psi_X(2t)$ and $\psi_Y(2t)$ exist for $t > 0$, then so does $\psi_{X+Y}(t)$ (see Theorem 8.4 with $p = q = 2$), similarly for (ii). Also note that the moment generating function exists for all $t \in \mathbb{R}$ if X is bounded.

Example 10.3. Consider $S \sim Bin(n, p)$ for $n \in \mathbb{N}$ and $p \in [0, 1]$, then $\psi_S(t)$ exists for every $t \in \mathbb{R}$ and

$$\psi_S(t) = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = (pe^t + 1 - p)^n. \tag{10.9}$$

By differentiation, we have

$$\psi_S'(t) = n(pe^t + 1 - p)^{n-1}pe^t \qquad \overset{(10.4)}{\Rightarrow} \qquad \mathbf{E}[S] = np. \tag{10.10}$$

With this we reproduce (7.67).

We now quote without proof another fundamental property of the moment generating function, which is that it characterizes the law of X.

Theorem 10.4. *Suppose that X and Y are two real random variables with moment generating functions ψ_X and ψ_Y, which both exist in $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Then*

$$\psi_X(t) = \psi_Y(t) \text{ for all } t \in (-\varepsilon, \varepsilon) \qquad \Leftrightarrow \qquad \mathbf{P}_X = \mathbf{P}_Y. \tag{10.11}$$

Proof. See, e.g., [i, Theorem 2.1] for a proof.[2] □

[1] This is immediate if Ω_X is finite, and can be justified generally.

[2] In fact this reference reduces the statement to the claim (10.14) for characteristic functions, using the identity theorem from complex analysis.

This gives another characterization of equality in law, besides the characterization with cumulative distribution functions (see (5.11)), or the equality of the probability mass functions / probability density functions. To give an example, we consider the case of Binomial distributions.

Example 10.5. Let $X \sim Bin(n,p)$ and $Y \sim Bin(m,p)$ be independent with $n,m \in \mathbb{N}$ and $p \in [0,1]$. The moment generating functions of X and Y exist for all $t \in \mathbb{R}$. Then we see, using (10.9) and Lemma 10.2, (i),

$$\psi_{X+Y}(t) = (pe^t + (1-p))^n \cdot (pe^t + (1-p))^m = (pe^t + (1-p))^{n+m}, \tag{10.12}$$

and this is the moment generating function for a $Bin(n+m,p)$, so $X+Y \sim Bin(n+m,p)$ (which we already saw in Example 7.24, (i)).

For completeness, we also give the definitions and properties for two other relevant functions.

Definition 10.6. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X : (\Omega,\mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ a real random variable. We define the *characteristic function* $\varphi_X : \mathbb{R} \to \mathbb{C}$ of X by

$$\varphi_X(t) = \mathbf{E}\big[e^{\mathrm{i}tX}\big] = \begin{cases} \sum_{k\in\Omega_X} e^{\mathrm{i}tk}\mathbf{P}[X=k], & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{\mathrm{i}tx} f_X(x)\mathrm{d}x, & \text{if } X \text{ is continuous and its law has density } f_X, \end{cases} \tag{10.13}$$

for $t \in \mathbb{R}$. Here, i is the imaginary unit (with $\mathrm{i}^2 = -1$).

End of Lecture 21

The characteristic function has a similar uniqueness property as we saw for the moment generating function (Theorem 10.4), and it has the advantage that it always exists for real random variables. We can say

$$\varphi_X(t) = \varphi_Y(t) \text{ for all } t \in \mathbb{R} \qquad \Leftrightarrow \qquad \mathbf{P}_X = \mathbf{P}_Y. \tag{10.14}$$

The (harder, but more important) direction ($\Rightarrow$) follows from the inversion formula in [GS01, Section 5.9].

Definition 10.7. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X : (\Omega,\mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ a discrete real random variable with $\Omega_X \subseteq \mathbb{N}_0$ (i.e. X only attains values in $\mathbb{N}_0$). We define the *probability generating function* G_X of X by

$$G_X(t) = \mathbf{E}\big[t^X\big] = \sum_{k=0}^{\infty} t^k \mathbf{P}[X=k], \tag{10.15}$$

for all $t \geq 0$ for which it exists.

Clearly, $G_X(t)$ exists for $t \in [0,1]$ and is differentiable at least on $[0,1)$. We have similar properties as for the moment generating functions, which we state here without proof (see [Geo12, Theorem 4.33] for a proof).

Lemma 10.8. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $X : (\Omega,\mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ a discrete real random variable with $\Omega_X \subseteq \mathbb{N}_0$. Then* *(i) For every $k \in \mathbb{N}$, one has*

$$\mathbf{P}[X = k] = \frac{G_X^{(k)}(0)}{k!}. \tag{10.16}$$

In particular, $\mathbf{P}_X$ is uniquely determined by G_X.[3]*.*

(ii) $\mathbf{E}[X]$ *exists if and only if $G'_X(1) = \lim_{t\uparrow 1} G'_X(t)$ exists and then*

$$\mathbf{E}[X] = G'_X(1). \tag{10.17}$$

[3] So we have, just as in (10.11) and (10.14), that for two $\mathbb{N}_0$-valued random variables X and Y, that $G_X(t) = G_Y(t)$ for every $t \in [0,1] \Leftrightarrow \mathbf{P}_X = \mathbf{P}_Y$.

11. Convergence in probability, almost sure convergence and the law of large numbers

(Reference: [GS01, Sections 7.1–7.5], or [Geo12, Section 5.1])

11.1. Convergence in probability and the weak law of large numbers

Definition 11.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(Z_n)_{n\in\mathbb{N}}$ a sequence of random variables $Z_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $Z : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(i) We say that $(Z_n)_{n\in\mathbb{N}}$ *converges in probability* to Z, if

$$\lim_{n\to\infty} \mathbf{P}\big[|Z_n - Z| > \varepsilon\big] = 0, \qquad \text{for all } \varepsilon > 0. \tag{11.1}$$

We write this as

$$Z_n \xrightarrow[n\to\infty]{\mathbf{P}} Z. \tag{11.2}$$

(ii) Suppose that for $p \geq 1$, we have that $\mathbf{E}[|Z_n|^p] < \infty$ for $n \in \mathbb{N}$ and $\mathbf{E}[|Z|^p] < \infty$. We say that $(Z_n)_{n\in\mathbb{N}}$ *converges in p^{th} mean* to Z if

$$\mathbf{E}\big[|Z_n - Z|^p\big] \xrightarrow[n\to\infty]{} 0. \tag{11.3}$$

We write this as

$$Z_n \xrightarrow[n\to\infty]{(p)} Z. \tag{11.4}$$

Convergence in p^{th} mean implies convergence in probability.

Proposition 11.2. *Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of real random variables on $(\Omega, \mathcal{F}, \mathbf{P})$, and Z a random variable on the same space. Let $p \geq 1$. Then*

$$Z_n \xrightarrow[n\to\infty]{(p)} Z \qquad \Rightarrow \qquad Z_n \xrightarrow[n\to\infty]{\mathbf{P}} Z. \tag{11.5}$$

Proof. The function $x \mapsto x^p$ is strictly increasing on $[0, \infty)$, so by Markov's inequality (Theorem 6.13), we have for $\varepsilon > 0$,

$$\mathbf{P}\big[|Z_n - Z| > \varepsilon\big] = \mathbf{P}\big[|Z_n - Z|^p > \varepsilon^p\big] \le \frac{\mathbf{E}\big[|Z_n - Z|^p\big]}{\varepsilon^p}. \tag{11.6}$$

Since the right-hand side converges to 0 by assumption as n tends to infinity, the claim follows. □

Theorem 11.3 (Weak law of large numbers)**.** *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $(X_n)_{n\in\mathbb{N}}$ a sequence of i.i.d. random variables $X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with $\mathbf{E}\big[X_1^2\big] < \infty$ and $\mu = \mathbf{E}[X_1]$. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{(2)} \mu, \tag{11.7}$$

so in particular

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbf{P}} \mu. \tag{11.8}$$

Proof. Let $\sigma^2 = \mathrm{Var}[X_1]$. We have

$$\mathbf{E}\big[|\overline{X}_n - \mu|^2\big] = \frac{\sum_{i=1}^n \mathbf{E}[(X_i - \mu)^2]}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0, \tag{11.9}$$

proving (11.7). The claim (11.8) follows by Proposition 11.2. □

Remark 11.4.
(i) The assumption that $\mathbf{E}\big[X_1^2\big] < \infty$ can be relaxed to $\mathbf{E}[|X_1|] < \infty$ (see Theorem 5.7 in [Geo12]). However, if $\mathbf{E}[|X_1|] = \infty$, the weak law of large numbers does not necessarily hold anymore.

(ii) One can see from the proof that one can relax the assumption that $(X_n)_{n\in\mathbb{N}}$ are i.i.d. by the weaker assumption that $(X_n)_{n\in\mathbb{N}}$ are identically distributed and pairwise uncorrelated, namely $\mathrm{Cov}[X_n, X_m] = 0$ for all $n \neq m$.

11.2. Almost sure convergence and the strong law of large numbers

We have seen that for $(X_n)_{n\in\mathbb{N}}$ i.i.d. with $\mathbf{E}[X_1^2] < \infty$, the probability to see a deviation from the average by at least $\varepsilon > 0$ converges to zero. In fact, something stronger is true: We will see that not only does this probability go to zero, but in fact with probability 1, $\overline{X}_n$ converges to $\mathbf{E}[X_1]$. We formalize this kind of convergence.

Definition 11.5. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $(Z_n)_{n\in\mathbb{N}}$ a sequence of random variables $Z_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. We say that $(Z_n)_{n\in\mathbb{N}}$ *converges* $\mathbf{P}$*-almost surely* (abbreviated as $\mathbf{P}$*-a.s.*) to some random variable $Z : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, if

$$\mathbf{P}\left[\left\{\omega \in \Omega\,;\, \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\right\}\right] = 1 \tag{11.10}$$

We write this as

$$Z_n \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} Z. \tag{11.11}$$

One can show that the set $\{\omega \in \Omega\,;\, \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\}$ is in $\mathscr{F}$. Almost sure convergence is a stronger notion than convergence in probability. Indeed, one has:

Proposition 11.6. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $(Z_n)_{n\in\mathbb{N}}$ a sequence of random variables $Z_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$.*

$$Z_n \xrightarrow[n\to\infty]{\mathbf{P}\text{-}a.s.} Z \qquad \Rightarrow \qquad Z_n \xrightarrow[n\to\infty]{\mathbf{P}} Z. \tag{11.12}$$

End of Lecture 22

Proof. Note that

(11.13)

$$\begin{aligned}\mathbf{P}\big[|Z_n - Z| > \varepsilon\big] &\leq \mathbf{P}\left[\bigcup_{k=n}^{\infty}\{|Z_k - Z| > \varepsilon\}\right]\\ &\xrightarrow[n\to\infty]{} \mathbf{P}\left[\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty}\{|Z_k - Z| > \varepsilon\}\right] \leq \mathbf{P}\left[\left\{\omega \in \Omega\,:\, \lim_{n\to\infty} Z_n(\omega) \neq Z(\omega)\right\}\right].\end{aligned}$$

We have used Proposition 1.15, (vii) in the second step. Now if $Z_n \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} Z$, then the right-hand side of the above inequality is zero, and thus $\lim_{n\to\infty} \mathbf{P}[|Z_n - Z| > \varepsilon] = 0$ follows. □

We now state a deep result relating to almost sure convergence.

Lemma 11.7 (Borel-Cantelli). *(i) (First Borel-Cantelli lemma) Let $A_n \in \mathscr{F}$ for $n \in \mathbb{N}$. Then*

$$\sum_{n=1}^{\infty}\mathbf{P}[A_n] < \infty \qquad \Rightarrow \qquad \mathbf{P}\left[\bigcap_{n=1}^{\infty}\bigcup_{k\geq n} A_k\right] = 0. \tag{11.14}$$

(ii) (Second Borel-Cantelli lemma) Let $A_n \in \mathscr{F}$ for $n \in \mathbb{N}$ be independent. Then

$$\sum_{n=1}^{\infty}\mathbf{P}[A_n] = \infty \qquad \Rightarrow \qquad \mathbf{P}\left[\bigcap_{n=1}^{\infty}\bigcup_{k\geq n} A_k\right] = 1. \tag{11.15}$$

Proof. For (i) note that

$$\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k \subseteq \bigcup_{k \geq n} A_k, \qquad \text{for all } n \in \mathbb{N}. \tag{11.16}$$

It follows that

$$\mathbf{P}\left[\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right] \leq \sum_{k=n}^{\infty} \mathbf{P}[A_k] \to 0 \qquad \text{as } n \to \infty. \tag{11.17}$$

For (ii), note first that

$$\left(\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right)^c = \bigcup_{n=1}^{\infty} \bigcap_{k \geq n} A_k^c \Rightarrow \mathbf{P}\left[\left(\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right)^c\right] = \lim_{n\to\infty} \mathbf{P}\left[\bigcap_{k \geq n} A_k^c\right], \tag{11.18}$$

since the sequence $(\bigcap_{k \geq n} A_k^c)_{n \geq 1}$ is increasing. We then calculate for $r > n$,

$$\mathbf{P}\left[\bigcap_{k=n}^{r} A_k^c\right] = \prod_{k=1}^{r} (1 - \mathbf{P}[A_k]) \leq \prod_{k=n}^{r} \exp(-\mathbf{P}[A_k]) = \exp\left(-\sum_{k=n}^{r} \mathbf{P}[A_k]\right), \tag{11.19}$$

where we used independence in the first step, and the bound $e^x \geq 1 + x$ for every $x \in \mathbb{R}$ in the second. Thus, letting $r \to \infty$ and using that $\sum_{k=n}^{\infty} \mathbf{P}[A_k] = \infty$ for every $n \in \mathbb{N}$, we see that $\mathbf{P}\left[\left(\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right)^c\right] = 0$. □

Theorem 11.8 (Strong law of large numbers). *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space and $(X_n)_{n\in\mathbb{N}}$ a sequence of i.i.d. random variables $X_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with $\mathbf{E}[X_1^2] < \infty$ and $\mu = \mathbf{E}[X_1]$. Then*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} \mu. \tag{11.20}$$

Proof. We follow the proof from [Geo12, Theorem 5.16]. Define $Y_n = X_n - \mu$ so that $(Y_n)_{n\in\mathbb{N}}$ are i.i.d. and $\mathbf{E}[Y_1] = 0$. We define $Z_n = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. We first show that $Z_{n^2} \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} 0$. Indeed, let $\varepsilon > 0$ and $\sigma^2 = \mathrm{Var}[Y_1] = \mathrm{Var}[X_1]$. Then

$$\mathbf{P}\left[Z_n^2 > \varepsilon\right] \leq \frac{\sigma^2}{n^2 \varepsilon^2}, \tag{11.21}$$

as in the proof of the weak law of large numbers. By the first Borel-Cantelli lemma, we see that

$$\mathbf{P}\left[\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} \{Z_{k^2} > \varepsilon\}\right] = 0. \tag{11.22}$$

In particular, we see that

$$\mathbf{P}\left[\limsup_{n\to\infty} Z_{n^2} > \varepsilon\right] = 0. \tag{11.23}$$

We now let $\varepsilon = 1/m$ for $m \in \mathbb{N}$ and note that

$$\begin{aligned}\mathbf{P}\left[\{\lim_{n\to\infty} Z_{n^2} = 0\}^c\right] &= \mathbf{P}\left[\limsup_{n\to\infty} Z_{n^2} > 0\right] \\ &\overset{\text{Prop. 1.15, (vi)}}{=} \lim_{m\to\infty} \mathbf{P}\left[\limsup_{n\to\infty} Z_{n^2} > \frac{1}{m}\right] = 0,\end{aligned} \tag{11.24}$$

so the claim follows. We now need to establish the $\mathbf{P}$-a.s., convergence of $(Z_n)_{n\in\mathbb{N}}$, i.e. extend the result from the convergence along the fixed subsequence $(Z_{n^2})_{n\in\mathbb{N}}$. To this end, let $\ell \in \mathbb{N}$ and choose the unique $n = n(\ell)$ such that $n^2 \le \ell < (n+1)^2$. Set $S_k = kZ_k = \sum_{i=1}^k Y_i$. By Chebyshev's inequality, we have

$$\mathbf{P}\left[|S_\ell - S_{n^2}| > \varepsilon n^2\right] \le \frac{1}{n^4\varepsilon^2}\mathrm{Var}\left[\sum_{n^2<i\le\ell} Y_i\right] \le \frac{\sigma^2(\ell - n^2)}{\varepsilon^2 n^4}, \tag{11.25}$$

and moreover

$$\begin{aligned}\sum_{\ell\ge1}\left[|S_\ell - S_{n(\ell)^2}| > \varepsilon n(\ell)^2\right] &\le \frac{\sigma^2}{\varepsilon^2}\sum_{n=1}^{\infty}\sum_{\ell=n^2}^{(n+1)^2-1}\frac{\ell - n^2}{n^4} \\ &= \frac{\sigma^2}{\varepsilon^2}\sum_{n=1}^{\infty}\sum_{k=1}^{2n}\frac{k}{n^4} = \frac{\sigma^2}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{2n(2n+1)}{2n^4} < \infty,\end{aligned} \tag{11.26}$$

so again by the first Borel-Cantelli lemma, we see that

$$\mathbf{P}\left[\left|\frac{S_\ell}{n(\ell)^2} - Z_{n(\ell)^2}\right| \xrightarrow[\ell\to\infty]{} 0\right] = 1. \tag{11.27}$$

Now since the intersection of two events with probability 1 has again probability 1 (by considering the complements and using the union bound), we find that

$$\mathbf{P}\left[\frac{S_\ell}{n(\ell)^2} \xrightarrow[\ell\to\infty]{} 0\right] = 1. \tag{11.28}$$

Finally, since $|Z_\ell| = |S_\ell|/\ell \le |S_\ell|/n(\ell)^2$, we find that $\mathbf{P}[Z_\ell \to 0] = 1$. Recalling now that $Z_\ell = \frac{1}{\ell}\sum_{i=1}^{\ell}(X_i - \mu) = \overline{X}_\ell - \mu$, the statement of the theorem follows. □

To close this section, we give the following result known as the *continuous mapping theorem.*

Proposition 11.9. *Consider a family $(Z_n)_{n\in\mathbb{N}}$ of random variables $Z_n : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ and a random variable $Z : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Furthermore, let $g : (\mathbb{R}, \mathscr{B}(\mathbb{R})) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ be a measurable function such that*

$$\mathbf{P}[Z \in D_g] = 0, \qquad D_g = \{x \in \mathbb{R}\,;\, g \text{ is not continuous in } x\}. \tag{11.29}$$

(i) If $Z_n \xrightarrow[n\to\infty]{\mathbf{P}} Z$*, then also* $g(Z_n) \xrightarrow[n\to\infty]{\mathbf{P}} g(Z)$*,*

(ii) If $Z_n \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} Z$*, then also* $g(Z_n) \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} g(Z)$*.*

For instance, consider $(X_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence with $X_1 \sim \mathscr{E}(\lambda)$. By the strong law of large numbers,

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} \mathbf{E}[X_1] = \frac{1}{\lambda}. \tag{11.30}$$

Therefore, by the continuous mapping theorem, Proposition 11.9, (ii), we have

$$\frac{1}{\overline{X}_n} \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} \lambda. \tag{11.31}$$

Proof. Claim (ii) is straightforward: Take $\widetilde{\Omega} \in \mathscr{F}$ with $\mathbf{P}[\widetilde{\Omega}] = 1$ such that for every $\omega \in \widetilde{\Omega}$, $Z_n(\omega) \to Z(\omega)$. Then $\{Z \in D_g\}^c \cap \widetilde{\Omega}$ still has probability 1, and $g(Z_n(\omega)) \to g(Z(\omega))$ as $n \to \infty$.

We now turn to (i). Fix $\varepsilon > 0$ and define the set $B_\delta = \{x \in \mathbb{R} : x \notin D_g$ and there exists $y \in \mathbb{R}$ with $|x - y| < \delta, |g(x) - g(y)| > \varepsilon\}$. We see that $\lim_{\delta\downarrow 0} B_\delta = \varnothing$. Now $\mathbf{P}[|g(Z_n) - g(Z)| > \varepsilon] \leq \mathbf{P}[|Z_n - Z| \geq \delta] + \mathbf{P}[Z \in B_\delta] + \mathbf{P}[Z \in D_g]$. Letting $n \to \infty$ and afterwards $\delta \downarrow 0$ gives the claim. □

End of Lecture 23

11.3. Application: Monte-Carlo integration

The law of large numbers essentially states that the average $\overline{X}_n(\omega)$ of i.i.d. random variables $X_1, ..., X_n$ is close to μ with high probability of large n. An important application of the law of large numbers are *Monte-Carlo methods* to calculate integrals or sums.

Example 11.10. Consider a piecewise continuous function $h : [0,1] \to \mathbb{R}$ we want to calculate the integral

$$I = \int_0^1 h(x)\mathrm{d}x, \tag{11.32}$$

which cannot be calculated elementary. We also assume that $\int_0^1 h^2(x)\mathrm{d}x < \infty$[1]. The *Monte-Carlo integration* gives us a way to find an estimate of I.

Let $X_1, X_2, ... \sim \mathscr{U}([0,1])$ be i.i.d. random variables. Then also $h(X_1), h(X_2), ...$ are i.i.d. by Proposition 7.15 and the fact (not proved here) that the function h is measurable, since it is piecewise continuous. Note that

$$\mathbf{E}\big[h(X_1)\big] \overset{(6.27)}{=} \int_0^1 h(x)\mathrm{d}x = I. \tag{11.33}$$

[1] For instance since we know some bound. Note in particular that I is therefore well-defined and finite.

By the strong law of large numbers, Theorem 11.8, we have

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow[n\to\infty]{\mathbf{P}\text{-a.s.}} I. \tag{11.34}$$

This means that:

$$\mathbf{P}\left[\left\{\omega \in \Omega \,:\, \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = I\right\}\right] = 1. \tag{11.35}$$

In other words, by simulating i.i.d. uniform random variable, we have a method to estimate the integral I.

Remark 11.11. (i) In the same way, we can calculate approximately other integrals or series by considering i.i.d. random variables $X_1, X_2, \ldots$ with a different distribution. For instance, if $X_1, X_2, \ldots \sim \mathcal{N}(0,1)$ i.i.d., we can approximate

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} h(x)\mathrm{d}x \approx \frac{\sqrt{2\pi}}{n}\sum_{i=1}^{n} h(X_i), \tag{11.36}$$

if $h : \mathbb{R} \to \mathbb{R}$ is piecewise continuous and $\int_{-\infty}^{\infty} h^2(x)e^{-\frac{x^2}{2}}\mathrm{d}x < \infty$.

(ii) The speed of convergence of Monte-Carlo is quite poor (we can find the speed by the central limit theorem later). Therefore, Monte-Carlo methods are typically only used if h is very irregular, or to perform high-dimensional integration. For regular, one-dimensional functions h numerical methods are typically much better than Monte-Carlo methods.

12. Weak convergence and the central limit theorem

(Reference: [GS01, Section 5.10], or [Geo12, Section 5.2, 5.3])

Consider rolling a die n times. We are interested in the sum of all numbers that came during the n rolls. By now we know that such a process can be described in different ways, and one of them is to consider $X_1, X_2, ..., X_n$ i.i.d. random variables with $X_1 \sim \mathcal{U}(\{1, ..., 6\})$. From the previous section we learned that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbf{P}} \mathbf{E}[X_1]. \tag{12.1}$$

This means that the average of the numbers shown should be close to $\mathbf{E}[X_1] = 3.5$ for large n, consequently the sum of the numbers should be close to $3.5 \cdot n$. What is not clear however is *how* close it really is. For instance, we could ask

(12.2) *How likely is it that the sum of outcomes when rolling a die* 100 *times exceeds* 400*?*

In other words, we are interested in *fluctuations* around $\mathbf{E}[X_1]$. This will be the main statement of the central limit theorem. For this, we need another notion of convergence.

12.1. Convergence in distribution

Definition 12.1. Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of real random variables with cumulative distribution functions F_{Z_n} and Z a real random variable with cumulative distribution function F_Z. We say that Z_n *converges in law / in distribution* if

$$F_{Z_n}(z) = \mathbf{P}[Z_n \le z] \xrightarrow[n\to\infty]{} F_Z(z), \tag{12.3}$$

for all points of continuity $z \in \mathbb{R}$ of F_Z. We write

$$Z_n \xrightarrow[n\to\infty]{d} Z. \tag{12.4}$$

Remark 12.2. (i) Since F_Z is monotone and bounded, one can show that there are at most countably many points of discontinuity.

(ii) Note that the above notion of convergence does not depend on the random variables Z_n and Z, but only on their laws $\mathbf{P}_{Z_n}$ and $\mathbf{P}_Z$. If $Z_n \xrightarrow[n\to\infty]{d} Z$ also say that $\mathbf{P}_{Z_n}$ converges *weakly* to $\mathbf{P}_Z$, also written as

$$\mathbf{P}_{Z_n} \xrightarrow[n\to\infty]{w} \mathbf{P}_Z. \tag{12.5}$$

(iii) To understand why we exclude discontinuity points of F_Z, consider the deterministic functions $Z_n = \frac{1}{n}$ and $Z = 0$. Then

$$F_{Z_n}(z) = \begin{cases} 1, & z \geq \frac{1}{n}, \\ 0, & z < \frac{1}{n}, \end{cases} \qquad F_Z(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases} \tag{12.6}$$

We see that $F_{Z_n}(0) = 0$ does not converge to $F_Z(0) = 1$, but of course we would like to say that $Z_n \xrightarrow[n\to\infty]{d} Z$, so we exclude the convergence at this point.

12.2. The central limit theorem

Let us start with a motivation: As explained at the beginning of this chapter, we want to be able to calculate probabilities of fluctuations for sums of i.i.d. random variables. Naturally we should ask, how large these fluctuations typically are.

Example 12.3. Let $X_1, \dots, X_n$ be i.i.d. real random variables with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \tfrac{\sigma^2}{n}) \qquad \Rightarrow \qquad \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1). \tag{12.7}$$

In other words, we have that

$$F_{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}}(x) = \Phi(x), \qquad x \in \mathbb{R}, \tag{12.8}$$

where Φ is the cumulative distribution function of a $\mathcal{N}(0,1)$-distributed random variable.

From this, we see that the scaling $\sqrt{n}$ is the "correct" scaling to measure the fluctuations of $\overline{X}_n$, at least if $X_1, \dots, X_n$ are i.i.d. with $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. That this is true in general is the statement of the celebrated *central limit theorem*:

Theorem 12.4. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. real random variables with $\mu = \mathbf{E}[X_1]$ and $\sigma^2 = \mathrm{Var}[X_1] \in (0, \infty)$. Then,*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1). \tag{12.9}$$

The notation in (12.9) *means that*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} Z, \text{ where } Z \sim \mathcal{N}(0,1). \tag{12.10}$$

There are different ways to establish the central limit theorem. In there notes, two are presented, one using moment generating functions and the other more direct one, is presented in the Appendix A.3.

End of Lecture 24

For the first method, we need the following *continuity theorem for moment generating functions*, which we state without proof:

Lemma 12.5. *Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of real random variables with Z a real random variable. Suppose the the moment generating functions $\psi_{Z_n}(t)$ and $\psi_Z(t)$ exist for all $n \geq 1$, $|t| < \varepsilon$ for some $\varepsilon > 0$. Then*

$$Z_n \xrightarrow[n\to\infty]{d} Z \qquad \Leftrightarrow \qquad \psi_{Z_n}(t) \xrightarrow[n\to\infty]{} \psi_Z(t) \qquad \text{for all } t \in (-\varepsilon, \varepsilon). \tag{12.11}$$

Proof. See, e.g., [i, Theorems 3.1, 3.2] for a proof. □

Since we will need it in the rest, we calculate the moment generating function of $Z \sim \mathcal{N}(0,1)$:

$$\begin{aligned} \psi_Z(t) = \mathbf{E}[e^{tZ}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz} e^{-\frac{z^2}{2}} \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} \mathrm{d}z = e^{\frac{t^2}{2}}, \end{aligned} \tag{12.12}$$

for every $t \in \mathbb{R}$.

Proof of Theorem 12.4. To show that claim (12.9), we show that the moment generating function of $\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}$ converges to that of $Z \sim \mathcal{N}(0,1)$. We also assume that all moment generating functions exist, at least for t in a neighborhood of 0.

Now we have:

$$\begin{aligned} \psi_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) = \psi_{\frac{1}{\sqrt{n}\sigma}(\sum_{i=1}^n (X_i-\mu))}(t) &= \prod_{i=1}^n \psi_{X_i-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right) \\ &= \prod_{i=1}^n \left(e^{-\mu\frac{t}{\sqrt{n}\sigma}} \psi_{X_i}\left(\frac{t}{\sqrt{n}\sigma}\right)\right) = e^{-\mu\sqrt{n}\frac{t}{\sigma}} \left(\psi_{X_1}\left(\frac{t}{\sqrt{n}\sigma}\right)\right)^n. \end{aligned} \tag{12.13}$$

We consider $L(u) = \log \psi_{X_1}(u)$. Note that

$$\begin{aligned} L(0) &= 0, \\ L'(0) &= \frac{\psi'_{X_1}(0)}{\psi_{X_1}(0)} = \mathbf{E}[X_1] = \mu, \\ L''(0) &= \frac{\psi_{X_1}(0)\psi''_{X_1}(0) - (\psi'_{X_1}(0))^2}{(\psi_{X_1}(0))^2} = \mathrm{Var}[X_1] = \sigma^2. \end{aligned} \tag{12.14}$$

Now we take the logarithm in (12.13):

$$\log \psi_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) = -\mu\sqrt{n}\frac{t}{\sigma} + n \log \psi_{X_1}\left(\frac{t}{\sqrt{n}\sigma}\right) = -\mu\sqrt{n}\frac{t}{\sigma} + nL\left(\frac{t}{2\sqrt{n}\sigma^2}\right). \tag{12.15}$$

One can then perform a Taylor expansion of the term on the right-hand side:

(12.16)

$$\begin{aligned} \log \psi_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) &= -\mu\sqrt{n}\frac{t}{\sigma} + n\left(L(0) + \frac{t}{\sqrt{n}\sigma}L'(0) + \frac{t^2}{n\sigma^2}L''(0) + O\left(\frac{t^3}{n^{3/2}\sigma^3}\right)\right) \\ &\overset{(12.14)}{=} \frac{t^2}{2} + O\left(\frac{t^3}{\sqrt{n}\sigma^3}\right). \end{aligned}$$

Sending $n \to \infty$ gives us

$$\log \psi_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) \to \frac{t^2}{2} \qquad \Rightarrow \qquad \psi_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) \to e^{\frac{t^2}{2}} \overset{(12.12)}{=} \psi_Z(t), \tag{12.17}$$

by the continuity of exp. The claim then follows by applying Lemma 12.5. □

Note that the proof above relied on Lemma 12.5 (which we did not prove) and does not work if the moment generating function of X_1 fails to exist in a neighborhood of 0. For completeness, a more direct (but technically more complicated) proof is sketched in the Appendix A.3.

Let us turn to some applications of the central limit theorem.

Example 12.6. (i) Recall the question (12.2) posed at the beginning of this chapter, namely suppose we roll a die $n = 100$ times and we are interested in the probability that the sum of the outcomes exceeds 400. To model this, let $X_1, X_2, \ldots$ be i.i.d. random variables with $X_1 \sim \mathcal{U}(\{1, \ldots, 6\})$. An easy calculation shows that

$$\mu = \mathbf{E}[X_1] = 3.5, \qquad \sigma^2 = \mathrm{Var}[X_1] = \frac{35}{12}. \tag{12.18}$$

Thus, the assumptions of the central limit theorem (Theorem 12.4) are fulfilled and thus for every $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbf{P}\left[\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \le x\right] = \lim_{n\to\infty} \mathbf{P}\left[\frac{\sum_{i=1}^n X_n - n\mu}{\sqrt{n}\sigma} \le x\right] = \Phi(x). \tag{12.19}$$

For $n = 100$, we therefore have the approximate probability

$$\begin{aligned}\mathbf{P}\left[\sum_{i=1}^{100} X_i > 400\right] &= \mathbf{P}\left[\frac{\sum_{i=1}^{100} X_i - 350}{\sqrt{100}\cdot\sqrt{\frac{35}{12}}} > \frac{400-350}{\sqrt{100}\cdot\sqrt{\frac{35}{12}}}\right] \\ &\approx \mathbf{P}\left[Z > \frac{400-350}{\sqrt{100}\cdot\sqrt{\frac{35}{12}}}\right], \qquad Z \sim \mathcal{N}(0,1) \\ &= 1 - \Phi\left(\frac{400-350}{\sqrt{100}\cdot\sqrt{\frac{35}{12}}}\right) = 1 - \Phi(2.927) \\ &\approx 1 - 0.9983 = 0.27\%.\end{aligned} \tag{12.20}$$

(ii) As another application, we present a general formula to approximate the binomial distribution $Bin(n,p)$ by a normal distribution, if $n \cdot p$ is not too small and n is large. We rely on the fact that for $X_1, X_2, \ldots$ i.i.d. random variables with $X_1 \sim Ber(p)$, $p \in (0,1)$, we have

$$S_n = \sum_{i=1}^n X_i \sim Bin(n,p). \tag{12.21}$$

Also recall that we have

$$\mathbf{E}[X_1] = p, \qquad \mathrm{Var}[X_1] = p(1-p). \tag{12.22}$$

Suppose now we want to approximate the probability mass function of $Bin(n,p)$ for $0 \le k \le n$, which is given by $p_{S_n}(k) = \binom{n}{k} p^k (1-p)^{n-k}$. Note that

$$\begin{aligned} \mathbf{P}[S_n = k] &= \mathbf{P}[k - 0.5 < S_n \le k + 0.5] \\ &= \mathbf{P}\Bigg[\frac{k-0.5-np}{\sqrt{np(1-p)}} < \underbrace{\frac{S_n - np}{\sqrt{np(1-p)}}}_{=\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}} \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\Bigg] \\ &\approx \mathbf{P}\left[\frac{k-0.5-np}{\sqrt{np(1-p)}} < Z \le \frac{k+0.5-np}{\sqrt{np(1-p)}}\right], \qquad Z \sim \mathcal{N}(0,1) \\ &= \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{np(1-p)}}\right). \end{aligned} \tag{12.23}$$

In the approximation we have chosen the interval of length 1, which makes the approximation better for finite n.

12.3. More properties of convergence in distribution

Proposition 12.7. *Let $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$ be sequences of real random variables with*

$$Y_n \xrightarrow[n\to\infty]{d} X, \qquad X_n - Y_n \xrightarrow[n\to\infty]{\mathbf{P}} 0. \tag{12.24}$$

Then it follows that

$$X_n \xrightarrow[n\to\infty]{d} X. \tag{12.25}$$

Proof. Let $Z_n = Y_n - X_n$. We need to show that $F_{X_n}(x) \to F_X(x)$ for all points of continuity x of $F_X(x)$. Let $x \in \mathbb{R}$ and $\varepsilon > 0$ such that x, $x \pm \varepsilon$ are points of continuity of F_X. Then

$$\begin{aligned} F_{X_n}(x) = \mathbf{P}[X_n \le x] &= \mathbf{P}[Y_n \le x + Z_n] \\ &= \mathbf{P}[Y_n \le x + Z_n, Z_n < \varepsilon] + \mathbf{P}[Y_n \le x + Z_n, Z_n \ge \varepsilon] \\ &\le \mathbf{P}[Y_n \le x + \varepsilon] + \mathbf{P}[|Z_n| \ge \varepsilon] \\ \Rightarrow \qquad \limsup_{n\to\infty} F_{X_n}(x) &\le F_X(x+\varepsilon). \end{aligned} \tag{12.26}$$

Analogously (by setting $Z_n > -\varepsilon$ instead of $Z_n < \varepsilon$):

$$\liminf_{n\to\infty} F_{X_n}(x) \ge F_X(x-\varepsilon). \tag{12.27}$$

Finally, since x is a point of continuity of F_X, we find that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x). \tag{12.28}$$

□

The following gives an overview of relations betweeen convergence in probability and convergence in distribution.

Theorem 12.8. *Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of real random variables, X a real random variable and $c \in \mathbb{R}$ a constant.*

(i) If $(X_n)_{n\in\mathbb{N}}$ converges in probability to X, then it converges also to X in distribution:

$$X_n \xrightarrow[n\to\infty]{\mathbf{P}} X \qquad \Rightarrow \qquad X_n \xrightarrow[n\to\infty]{d} X. \tag{12.29}$$

(ii) We have that

$$X_n \xrightarrow[n\to\infty]{d} X, Y_n \xrightarrow[n\to\infty]{\mathbf{P}} 0 \qquad \Rightarrow \qquad X_nY_n \xrightarrow[n\to\infty]{\mathbf{P}} 0. \tag{12.30}$$

(iii) One has Slutsky's theorem*:*

$$\begin{aligned} X_n \xrightarrow[n\to\infty]{d} X, Y_n \xrightarrow[n\to\infty]{\mathbf{P}} c \qquad \Rightarrow \qquad & X_n + Y_n \xrightarrow[n\to\infty]{d} X + c, \\ & X_nY_n \xrightarrow[n\to\infty]{d} cX, \\ & \frac{X_n}{Y_n} \xrightarrow[n\to\infty]{d} \frac{X}{c}, \ \text{if } c \neq 0. \end{aligned} \tag{12.31}$$

Proof. For (i), set $Y_n \equiv X$ in Proposition 12.7.

For (ii), let $k, \varepsilon > 0$. Then

$$\begin{aligned} & \mathbf{P}[|X_nY_n| \geq \varepsilon] = \mathbf{P}\left[|X_nY_n| \geq \varepsilon, |Y_n| < \tfrac{\varepsilon}{k}\right] + \mathbf{P}\left[|X_nY_n| \geq \varepsilon, |Y_n| \geq \tfrac{\varepsilon}{k}\right] \\ & \qquad \leq \mathbf{P}[|X_n| > k] + \mathbf{P}\left[|Y_n| \geq \tfrac{\varepsilon}{k}\right] \\ \Rightarrow \quad & \limsup_{n\to\infty} \mathbf{P}[|X_nY_n| \geq \varepsilon] \leq \mathbf{P}[|X| > k], \ \text{if } \pm k \text{ are points of continuity of } F_X. \end{aligned} \tag{12.32}$$

Replace k by a monotone sequence of points of continuity $k_m \in \mathbb{R}$. Since $\sum_{m=1}^{\infty} \mathbf{P}[|X| \in (k_{m-1}, k_m]] < \infty$, we have $\lim_{m\to\infty} \mathbf{P}[|X| > k_m] = 0$, thus (12.30) follows.

For (iii), we first see that

$$(X_n + Y_n) - (X_n + c) = Y_n - c \xrightarrow[n\to\infty]{\mathbf{P}} 0 \qquad \Rightarrow \qquad X_n + Y_n \xrightarrow[n\to\infty]{d} X + c, \tag{12.33}$$

by Proposition 12.7. Moreover, we have

$$X_n \xrightarrow[n\to\infty]{d} X \qquad \Rightarrow \qquad cX_n \xrightarrow[n\to\infty]{\mathbf{P}} cX. \tag{12.34}$$

On the other hand, by (ii):

$$X_nY_n - cX_n = X_n(Y_n - c) \xrightarrow[n\to\infty]{\mathbf{P}} 0 \qquad \Rightarrow \qquad X_nY_n \xrightarrow[n\to\infty]{d} cX, \tag{12.35}$$

by Proposition 12.7. The last part of (iii) follows analogously. □

We illustrate the use of Proposition 12.7 and Theorem 12.8.

Example 12.9. Let $X_1, X_2, \ldots$ be i.i.d. real random variables, $\mathbf{E}[X_1] = \mu$, $\mathrm{Var}[X_1] = \sigma^2 \in (0,\infty)$ and $\mathbf{E}[(X_1-\mu)^4] = \mu_4^{(c)} \in (\sigma^4,\infty)$ the centered fourth moment of X_1. By the weak law of large numbers (Theorem 11.3), we know that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\mathbf{P}} \mathbf{E}[X_1] = \mu. \tag{12.36}$$

If we let $Y_i = (X_i-\mu)^2$, then we have $\mathbf{E}[Y_1] = \sigma^2$ and $\mathrm{Var}[Y_1] = \mathbf{E}[Y_1^2] - \mathbf{E}[Y_1]^2 = \mu_4^{(c)} - \sigma^4 \in (0,\infty)$, so by the central limit theorem 12.4 we have that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - \sigma^2\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mu_4^{(c)} - \sigma^4). \tag{12.37}$$

As an example consider the empirical variance given by

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2. \tag{12.38}$$

We claim that

$$\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mu_4^{(c)} - \sigma^4). \tag{12.39}$$

To see this, we consider the difference between the expression in (12.37) and the estimator S_n^2 in (12.38):

$$\begin{aligned} &\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2 - \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\right) \\ &= - \underbrace{\sqrt{n}(\overline{X}_n - \mu)}_{\xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^2),\ (12.9)} \cdot \underbrace{(\overline{X}_n - \mu)}_{\xrightarrow[n\to\infty]{\mathbf{P}} 0,\ (11.8)} \\ &\xrightarrow[n\to\infty]{\mathbf{P}} 0, \qquad \text{by Theorem 12.8, (ii).} \end{aligned} \tag{12.40}$$

Since the difference between $\sqrt{n}\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ and $\sqrt{n}(\widehat{\sigma}_n^2 - \sigma^2)$ converges in probability to zero, we can combine (12.37) and Proposition 12.7 to obtain (12.39).

13. An overview of selected random processes

(Reference: [GS01, Sections 6.1–6.4, 6.8], or [Geo12, Section 3.5, Chapter 6])

In this chapter, we will introduce a stochastic process in time that models the arrival of random events of a discrete nature, such as arrival times of radioactive particles at a Geiger counter, the times at which electronic components fail, or the arrival of customers in a store.

Definition 13.1. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space, $I \neq \emptyset$ an index set and E a set equipped with a σ-algebra $\mathscr{E}$. A *stochastic process* with *time parameter set* I and *state space* E is a collection $(X_t)_{t\in I}$ of random variables $X_t : (\Omega, \mathscr{F}) \to (E, \mathscr{E})$. For $\omega \in \Omega$, we say that $t \mapsto X_t(\omega)$ is a *sample path* of the process.

For us, $I = \mathbb{N}$ or $I = [0, \infty)$ are natural choices. The easiest examples of stochastic processes would include $(X_n)_{n\in\mathbb{N}}$, where $X_n = Z : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ for all $n \in \mathbb{N}$ (a constant process, that only depends on the initial randomness $Z(\omega)$), or $(Y_n)_{n\in\mathbb{N}}$, where $Y_1, Y_2, \ldots : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ are i.i.d. random variables. In these examples, $I = \mathbb{N}$ and $(E, \mathscr{E}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

13.1. The Poisson process

We now introduce the Poisson process (with rate $\lambda > 0$), which is a stochastic process with $I = [0, \infty)$ and $(E, \mathscr{E}) = (\mathbb{N}_0, \mathscr{P}(\mathbb{N}_0))$.

Definition 13.2. A stochastic process $(N_t)_{t\geq 0}$ of random variables $N_t : (\Omega, \mathscr{F}) \to (\mathbb{N}_0, \mathscr{P}(\mathbb{N}_0))$ is called a *Poisson process of rate* $\lambda > 0$ if it fulfills the following properties:

(i) $N_0 = 0$,

(ii) the paths $t \mapsto N_t(\omega)$ are right-continuous for every $\omega \in \Omega$,

(iii) for $0 = t_0 < t_1 < \ldots < t_n$, the increments $(N_{t_k} - N_{t_{k-1}})_{k=1,\ldots,n}$ are independent,

(iv) for $0 \leq s < t$, $N_t - N_s \sim \mathit{Pois}(\lambda(t-s))$.

Example 13.3. Let N_t be the number of emitted particles of a radioactive source during the time interval $[0, t]$. Then (iii) means that the numbers of particles emitted during disjoint time intervals are independent.

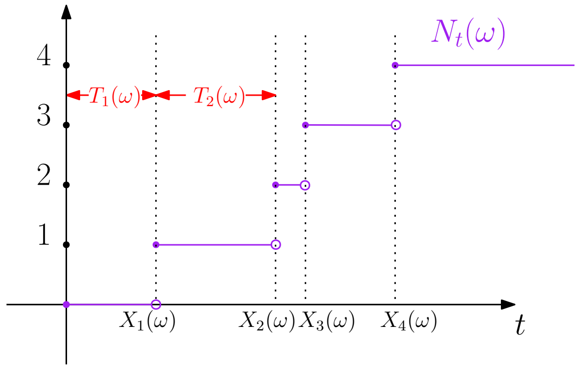

Figure 13.1.: Plot for one trajectory for $t \mapsto N_t(\omega)$.

Since $N_t = N_s + (N_t - N_s)$, where $N_t - N_s \geq 0$, we see that $t \mapsto N_t(\omega)$ is nondecreasing. Moreover, by definition, the process only attains values in $\mathbb{N}_0$. Therefore, sample paths will be piecewise constant, with "jumps" occuring at random times (see figure 13.1). How long do we have to wait for the first jump? Let us define

$$T = \inf\{t \geq 0\,;\, N_t > 0\}, \tag{13.1}$$

then we have that

$$\mathbf{P}[T > t] = \mathbf{P}[N_t = 0] = e^{-\lambda t}, \tag{13.2}$$

since $N_t = N_t - N_0 \sim Pois(\lambda t)$. More generally, we have that

$$\mathbf{P}[N_{s+t} - N_s = 0] = e^{-\lambda t}, \qquad s \geq 0, t > 0. \tag{13.3}$$

This suggests that the "random jumps" occur after independent, exponentially distributed times, the *interarrival times*. We use this idea to construct a Poisson process, and briefly explain afterwards, that this property actually characterizes Poisson processes.

Theorem 13.4. *Let $T_1, T_2, \ldots \sim \mathcal{E}(\lambda)$ be i.i.d. random variables with $\lambda > 0$. Define*

$$X_n = \sum_{k=1}^{n} T_k, \qquad N_t = |\{n \in \mathbb{N}_0\,;\, X_n \leq t\}|. \tag{13.4}$$

The family $(N_t)_{t\geq 0}$ is a Poisson process with intensity λ.

Proof. The proof follows [Geo12, Theorem 3.34]. We must show that for any $n \in \mathbb{N}$ and any sequence $0 = t_0 < t_1 < \ldots < t_n$, we have that $(N_{t_i} - N_{t_{i-1}})_{i=1}^n$ are independent and $N_{t_i} - N_{t_{i-1}} \sim Pois(\lambda(t_i - t_{i-1}))$. For simplicity, we show it only sketch the proof in the case of $n = 2$. Thus we want to show that for $0 < s < t$ and $\ell, k \in \mathbb{N}_0$:

$$\mathbf{P}[N_s = k, N_t - N_s = \ell] = \left(e^{-\lambda s}\frac{(\lambda s)^k}{k!}\right)\left(e^{-\lambda(t-s)}\frac{(\lambda(t-s))^\ell}{\ell!}\right). \tag{13.5}$$

This implies that N_s and $N_t - N_s$ are independent by Proposition 7.13, (ii). Moreover, by summing over $k \in \mathbb{N}_0$, we have $N_t - N_s \sim Pois(\lambda(t-s))$. Now the joint law of $T_1, ..., T_{k+\ell+1}$, $\mathbf{P}_{T_1,...,T_{k+\ell+1}}$, has the density

$$f_{T_1,...,T_{k+\ell+1}}(x_1, ..., x_{k+\ell+1}) = \lambda^{k+\ell+1} e^{-\lambda(x_1+...+x_{k+\ell+1})} \prod_{j=1}^{k+\ell+1} \mathbb{1}_{[0,\infty)}(x_j). \tag{13.6}$$

Let us only consider the cases where $k, \ell \geq 1$ (the other cases, i.e. where either $k = 0$ or $\ell = 0$ are similar). Now we have

$$\begin{aligned} \mathbf{P}[N_s = k, N_t - N_s = \ell] &= \mathbf{P}[X_k \leq s < X_{k+1} \leq X_{k+\ell} \leq t < X_{k+\ell+1}] \\ &= \int_0^\infty ... \int_0^\infty \mathrm{d}x_1 ... \mathrm{d}x_{k+\ell+1} \lambda^{k+\ell+1} e^{-\lambda(x_1+...+x_{k+\ell+1})} \\ &\quad \cdot \mathbb{1}_{s \in [x_1+...+x_k, x_1+...+x_{k+1})} \cdot \mathbb{1}_{t \in [x_1+...+x_{k+\ell}, x_1+...+x_{k+\ell+1})}. \end{aligned} \tag{13.7}$$

We now integrate now first over $x_{k+\ell+1}$: This yields for fixed $x_1, ..., x_{k+\ell}$

$$\int_0^\infty \mathrm{d}x_{k+\ell+1} \lambda e^{-\lambda(x_1+...+x_{k+\ell+1})} \mathbb{1}_{\{x_1+...+x_{k+\ell}+x_{k+\ell+1}>t\}} = \int_t^\infty \mathrm{d}z \lambda e^{-\lambda z} = e^{-\lambda t}. \tag{13.8}$$

We then fix $x_1, ..., x_k$ and integrate over $x_{k+1}, ..., x_{k+\ell}$:

$$\begin{aligned} &\int_0^\infty ... \int_0^\infty \mathrm{d}x_{k+1} ... \mathrm{d}x_{k+\ell} \mathbb{1}_{\{s < x_1+...+x_{k+1} \leq x_1+...x_{k+\ell} \leq t\}} \\ &\qquad = \int_0^\infty ... \int_0^\infty \mathrm{d}y_1 ... \mathrm{d}y_\ell \mathbb{1}_{\{y_1+...+y_\ell \leq t-s\}} = \frac{(t-s)^\ell}{\ell!}. \end{aligned} \tag{13.9}$$

We have used the substitution $y_1 = x_1 + ... + x_{k+1} - s$, $y_2 = x_{k+2}$, ..., $y_\ell = x_{k+\ell}$. Finally, we need to integrate over the remaining variables $x_1, ..., x_k$:

$$\int_0^\infty ... \int_0^\infty \mathrm{d}x_1 ... \mathrm{d}x_k \mathbb{1}_{\{x_1+...+x_k \leq s\}} = \frac{s^k}{k!}. \tag{13.10}$$

This means that we have

$$\mathbf{P}[N_s = k, N_t - N_s = \ell] = e^{-\lambda t} \lambda^{k+\ell} \frac{s^k}{k!} \frac{(t-s)^\ell}{\ell!}. \tag{13.11}$$

This is exactly □

Theorem 13.4 gives us the construction of a Poisson process: If we can simulate $T_1, T_2, ... \sim \mathscr{E}(\lambda)$ i.i.d. random variables, then we can define $(N_t)_{t\geq 0}$ as in (13.4) and obtain a Poisson process with rate $\lambda > 0$. In fact, the converse is also true, which we state here without proof.

Proposition 13.5. *Let $(N_t)_{t\geq 0}$ be a Poisson process of rate $\lambda > 0$. We define the jump times $(X_i)_{i\geq 1}$ by*

$$X_i(\omega) = \inf\{t \geq 0\,;\, N_t(\omega) = i\}, \qquad i \in \mathbb{N}. \tag{13.12}$$

For a set $\widetilde{\Omega} \in \mathscr{F}$ with $\mathbf{P}[\widetilde{\Omega}] = 1$, we have that $X_k(\omega) < \infty$ for $\omega \in \widetilde{\Omega}$ for every $k \in \mathbb{N}$, and the interarrival times

$$T_i = X_i - X_{i-1}, \qquad i \geq 1, \tag{13.13}$$

are i.i.d. random variables with $T_1 \sim \mathscr{E}(\lambda)$.

Let us consider a worked example.

Example 13.6. Queuing at a post office can be modelled by a Poisson process with rate $\lambda = \frac{1}{2}$ (in units $\frac{\text{customers}}{\text{min}}$).

(i) The expected number of arrivals during the first 10 minutes of an hour is given by

$$\mathbf{E}[N_{10}] = \lambda \cdot 10 = \frac{1}{2} \cdot 10 = 5, \tag{13.14}$$

since $N_{10} \sim Pois(\lambda \cdot 10)$.

(ii) The probability to have 4 or less arrivals during the first 10 minutes of an hour is

$$\mathbf{P}[N_{10} \leq 4] = \sum_{k=0}^{4} \frac{(\lambda \cdot 10)^k}{k!} e^{-\lambda \cdot 10} = e^{-5} \sum_{k=0}^{4} \frac{5^k}{k!} \approx 0.4405. \tag{13.15}$$

(iii) What is the probability that there are 3 customers in the first 10 minutes and 5 customers in the next 20 minutes? We have:

$$\begin{aligned} &\mathbf{P}[N_{10} = 3, N_{30} - N_{10} = 5] \\ &\qquad = \mathbf{P}[N_{10} = 3] \cdot \mathbf{P}[N_{30} - N_{30} = 5] \qquad \text{(independent increments)} \\ &\qquad = \left(e^{-5} \frac{5^3}{3!}\right) \cdot \left(e^{-10} \frac{10^5}{5!}\right) \qquad \text{(since } N_t - N_s \sim Pois(\lambda(t-s))\text{)} \\ &\qquad \approx 0.0053. \end{aligned} \tag{13.16}$$

13.2. Markov chains

In this section, we will define *Markov chains* and give a very brief overview over some classical properties. To keep things as simple as possible, we will only consider Markov chains

- in discrete time,
- with a finite state space.[1]

Throughout the entire section, we will set

$$E = \{1, 2, \ldots, N\}, \tag{13.17}$$

with $N \in \mathbb{N}$. This will be the *state space*. A (discrete-time) Markov chain will be a stochastic process $(X_n)_{n \in \mathbb{N}_0}$ with time parameter set $I = \mathbb{N}_0$ and state space E in the sense of Definition 13.1. Its characterizing feature is that at time n, the next position $X_{n+1} \in E$ of the process only depends on the current position X_n.

[1] Markov chains with an at most countable state space can be treated very similarly, and we refer to [Geo12, Chapter 6] for more on this.

Definition 13.7. A $\mathbb{R}^{n\times n}$-matrix $\Pi = (\Pi(x,y))_{(x,y)\in E^2}$ is called a *matrix of transition probabilities* if for every $x \in E$, the function

$$\Pi(x,\cdot) : E \to \mathbb{R}, y \mapsto \Pi(x,y) \tag{13.18}$$

is a probability mass function.

We will interpret the value $\Pi(x,y)$ as

(13.19) *the probability that the chain is at y at time $n+1$, if it was at x at time n.*

Example 13.8. Let $E = \{1,2,3\}$ and consider the three matrices

$$\Pi_1 = \begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.3 & 0.3 & 0.4 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \qquad \Pi_2 = \begin{pmatrix} 0.1 & 0.1 & 0.7 \\ 0.3 & 0.9 & 0.4 \\ 0 & 0 & 0.5 \end{pmatrix}, \qquad \Pi_3 = \begin{pmatrix} 0.2 & -0.1 & 0.9 \\ -0.3 & 0.9 & 0.4 \\ 0.6 & 0 & 0.5 \end{pmatrix}. \tag{13.20}$$

Out of the three matrices, only Π_1 is a valid matrix of transition probabilities. Indeed, in Π_2, the rows do not sum to one, and Π_3 has negative entries.

To visualize this, one often uses *transition graphs*, which are formally directed weighted graphs on the set E, and the weight of the directed edge $\overrightarrow{(x,y)}$ is exactly $\Pi(x,y)$. Edges with weight 0 are often ommited completely. An example for such a transition graph is given below.

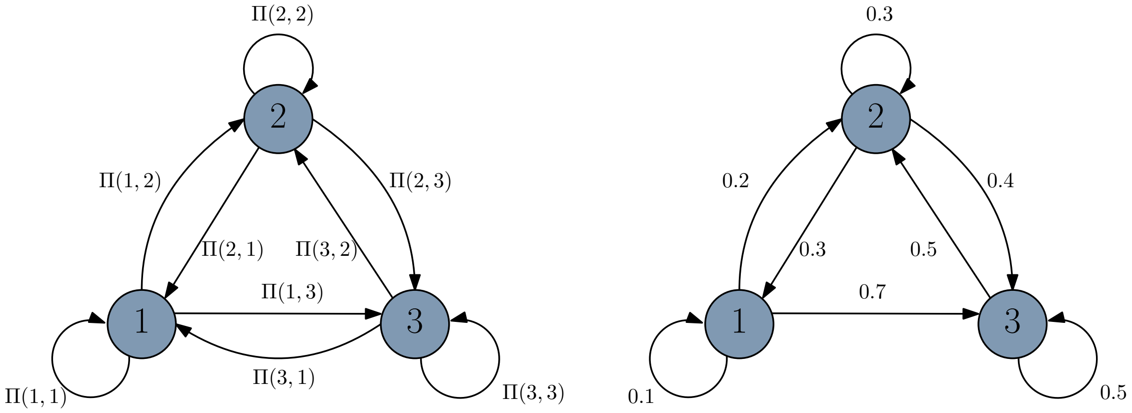

Figure 13.2.: Transition graph for a general Markov chain on $E = \{1,2,3\}$ on the left, and for the choice Π_1 from Example 13.8 on the right.

We now come to the formal definition of a Markov chain.

Definition 13.9. Let $X_0, X_1, \ldots$ be a sequence of random variables defined on a probability space $(\Omega, \mathscr{F}, \mathbf{P})$, all taking values in E. The stochastic process $(X_n)_{n\in\mathbb{N}_0}$ is called a *Markov chain* on E with transition probabilities Π if

$$\mathbf{P}[X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n] = \Pi(x_n, x_{n+1}), \tag{13.21}$$

for every $n \geq 0$ and $x_0, ..., x_{n+1} \in E$ such that $\mathbf{P}[X_0 = x_0, ..., X_n = x_n] > 0$. The law $\boldsymbol{\mu} = \mathbf{P}_{X_0}$ of X_0 is called the *initial distribution* of the Markov chain.

Remark 13.10. (i) From the preceding remarks, it is not immediate that Markov chains actually exist. One can show however, that given any initial distribution and any matrix of transition probabilities Π, one can construct a corresponding Markov chain. This construction can be found, e.g., in [Geo12, Remark 6.2 (d)].

(ii) Note that the right-hand side of (13.21) does not explicitly depend on n, and also not on $x_0, ..., x_{n-1}$. This formalizes the intuition given earlier: The next step taken by a Markov chain only depends on the *current position* $\{X_n = x_n\}$.

A Markov chain is in fact characterized completely by its initial distribution and its transition probabilities. To understand this, we make the following observation: Suppose that for some fixed $x \in E$:

$$\mathbf{P}[X_0 = y] = \mathbb{1}_{\{x=y\}}. \tag{13.22}$$

This means that at time 0, the chain starts in the point x. We have:

(13.23)

$$\begin{aligned}\mathbf{P}[X_1 = x_1, ..., X_n = x_n] &= \mathbf{P}[X_0 = x, X_1 = x_1, ..., X_n = x_n] \\ &= \underbrace{\mathbf{P}[X_n = x_n | X_{n-1} = x_{n-1}, ..., X_1 = x_1, X_0 = x]}_{=\Pi(x_{n-1}, x_n)} \\ &\qquad \cdot \mathbf{P}[X_{n-1} = x_{n-1}, ..., X_1 = x_1, X_0 = x] \\ &= \Pi(x_{n-1}, x_n)\, \underbrace{\mathbf{P}[X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}, ..., X_1 = x_1, X_0 = x]}_{=\Pi(x_{n-2}, x_{n-1})} \\ &\qquad \cdot \mathbf{P}[X_{n-2} = x_{n-2}, ..., X_1 = x_1, X_0 = x] \\ &= ... = \Pi(x_{n-1}, x_n) \cdot ... \cdot \Pi(x_2, x_1)\mathbf{P}[X_1 = x_1 | X_0 = x] \cdot \mathbf{P}[X_0 = x] \\ &= \prod_{j=1}^{n} \Pi(x_{j-1}, x_j).\end{aligned}$$

Let us first exemplify this: Suppose in Example 13.8 (with Π_1) we also know that $\mathbf{P}[X_0 = 1]$. What is the probability that the chain first jumps to 2, then to 3 and then to 3 again? Clearly:

$$\mathbf{P}[X_1 = 2, X_2 = 3, X_3 = 3] = \Pi(1,2)\Pi(2,3)\Pi(3,3) = 0.2 \cdot 0.4 \cdot 0.5 = 0.04.$$

A similar calculation can be performed in the case where X_0 has a non-trivial distribution $\boldsymbol{\mu}$ under $\mathbf{P}$. We summarize this in the following Proposition:

Proposition 13.11. *Suppose that* $(X_n)_{n \in \mathbb{N}_0}$ *is a Markov chain on* E *with transition probabilities* Π *and initial distribution* $\boldsymbol{\mu}$*. Then we have* *(i) We have*

$$\mathbf{P}[X_0 = x_0, ...X_n = x_n] = \boldsymbol{\mu}(x_0)\Pi(x_0, x_1) \cdot ... \cdot \Pi(x_{n-1}, x_n). \tag{13.24}$$

(ii) The Chapman-Kolmogorov equation *holds: Define the n-step transition probabilities as*

$$\Pi^{(n)}(x, y) = \mathbf{P}[X_{n+m} = y | X_n = x]. \tag{13.25}$$

Then

$$\Pi^{(n)}(x, y) = \sum_{z \in E} \Pi^{(r)}(x, z)\Pi^{(n-r)}(z, y), \qquad \textit{for } 0 < r < n. \tag{13.26}$$

In particular (13.26) *means that in the n-step transition probability is the n-fold matrix product of Π, i.e. $\Pi^{(n)} = \Pi^n = \Pi \cdot ... \cdot \Pi$ (n times).*

(iii) The probability for the chain to be at y at time n is given by

$$\mathbf{P}[X_n = y] = \sum_{z \in E} \boldsymbol{\mu}(z)(\Pi^n)(z, y) \tag{13.27}$$

Is it possible that the law of X_{n+1} coincides with the law of X_n for all n, i.e. for the Markov chain to be *stationary*? The answer is provided by (13.27): It states that in order to obtain the law of X_1, we can write the probability mass function of the law of X_0 as a row vector and multiply it with Π from the left. For instance, suppose that $E = \{1, 2, 3\}$ and $\boldsymbol{\mu} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$, i.e. at time 0 the chain starts in 1. Then considering again Example 13.8 (with Π_1), clearly

$$\begin{pmatrix} \mathbf{P}[X_1 = 1] & \mathbf{P}[X_1 = 2] & \mathbf{P}[X_1 = 3] \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.3 & 0.3 & 0.4 \\ 0 & 0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.1 & 0.2 & 0.7 \end{pmatrix}.$$

This leads us to the following idea.

Definition 13.12. Let Π be a matrix of transition probabilities of a Markov chain on $E = \{1, ..., n\}$. We say that a probability mass function $(\boldsymbol{\pi}(x))_{x \in E}$ is a *stationary* distribution for Π if

$$\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \Pi \tag{13.28}$$

(as matrices), or in components:

$$\boldsymbol{\pi}(x) = \sum_{z \in E} \boldsymbol{\pi}(z)\Pi(z, x), \qquad \text{for all } x \in E. \tag{13.29}$$

In other words: A stationary distribution is an *eigenvector* to the matrix $\Pi^\top$ corresponding to eigenvalue 1, with positive entries and ℓ^1-norm equal to 1. Finding a stationary distribution is therefore a problem in linear algebra! One may now ask, whether stationary distributions always exist and whether they are unique. The answer is given by the following theorem, which we state without proof.

Theorem 13.13. *Let* Π *be a matrix of transition probabilities for a Markov chain on* $E = \{1, ..., n\}$*. Suppose that there exists* $n \in \mathbb{N}$ *with the property that*[2]

$$\Pi^n(x, y) > 0 \qquad \textit{for all } x, y \in E. \tag{13.30}$$

Then there exists a unique stationary distribution for the Markov chain.

For many further topics, including

- classification of states (absorbing, recurrent, transient, ...),
- return times and their distribution,
- convergence of Markov chains,

we refer to the literature (see [Geo12, Section 6] for a much more complete treatment).

[2] The following condition means that the chain is *irreducible*: In other words, every state y may be reached by starting from any other state x in some finite time.

A. Appendix

A.1. Riemann integration

We give a very brief account of Riemann integrals without proofs, with a focus on improper Riemann integrals.

Definition A.1. Let $\mathbb{B}([a,b])$ be the set of bounded functions $f : [a,b] \to \mathbb{R}$.

(i) $Z_n = (x_0, x_1, ..., x_n)$ is called a *partition* if $a = x_0 < x_1 < ... < x_n = b$. For a partition Z_n, we call

$$|Z_n| = \max\{|x_j - x_{j-1}| \,:\, 1 \leq j \leq n\} \tag{A.1}$$

the *norm* of Z_n. We define

$$\mathscr{S} = \{Z_n \,:\, Z_n \text{ is a partition of } [a,b]\}. \tag{A.2}$$

(ii) Let $f \in \mathbb{B}([a,b])$ and $Z_n \in \mathscr{S}$ with $Z_n = (x_0, ..., x_n)$. Writing $I_j = [x_{j-1}, x_j]$ for $1 \leq j \neq n$, we define the *lower Riemann sum*

$$\underline{S}(Z_n, f) = \sum_{j=1}^{n} (x_j - x_{j-1}) \inf_{z \in I_j} f(z), \tag{A.3}$$

and the *upper Riemann sum*

$$\overline{S}(Z_n, f) = \sum_{j=1}^{n} (x_j - x_{j-1}) \sup_{z \in I_j} f(z) \tag{A.4}$$

(note that these are finite since f is assumed to be bounded).

(iii) Let $f \in \mathbb{B}([a,b])$. We define the *lower Riemann integral* of f over $[a,b]$ as

$$\underline{A}(f) = \underline{\int_a^b} f(x)\mathrm{d}x = \sup\{\underline{S}(Z,f) \,:\, Z \in \mathscr{S}\} \tag{A.5}$$

and the *upper Riemann integral* of f over $[a,b]$ as

$$\overline{A}(f) = \overline{\int_a^b} f(x)\mathrm{d}x = \inf\{\overline{S}(Z,f) \,:\, Z \in \mathscr{S}\}. \tag{A.6}$$

We say that f is *Riemann-integrable* over $[a,b]$ if $\underline{A}(f) = \overline{A}(f)$. In this case, the value $\underline{A}(f) = \overline{A}(f)$ is denoted $\int_a^b f(x)\mathrm{d}x$.

For the following, we define

$$\mathscr{R}([a,b]) = \{f \in \mathbb{B}([a,b]) \,:\, f \text{ is Riemann-integrable over } [a,b]\}. \tag{A.7}$$

Theorem A.2. *Let $f : [a,b] \to \mathbb{R}$ be a continuous function. Then f is Riemann-integrable over $[a,b]$. In other words:*

$$\mathscr{C}([a,b]) = \{f \in \mathbb{R}^{[a,b]} \,:\, f \textit{ is continuous}\} \subseteq \mathscr{R}([a,b]). \tag{A.8}$$

Proof. Note that $\|f\|_\infty = \max\{|f(x)| \,:\, x \in [a,b]\} < \infty$ since $|f|$ is continuous on the compact[1] interval $[a,b]$, so $f \in \mathscr{B}([a,b])$. Therefore, the lower and upper Riemann sums are well-defined.

Now let $Z_n = (x_0, ..., x_n) \in \mathscr{S}$ be the partition

$$x_j = a + \frac{b-a}{n} \cdot j, \qquad 0 \leq j \leq n. \tag{A.9}$$

Recall the notation $I_j = [x_{j-1}, x_j]$ and note that $|Z_n| = \frac{b-a}{n}$. Let $\varepsilon > 0$. Since f is continuous, it is in particular uniformly continuous on the compact interval $[a,b]$, and therefore there exists $N \in \mathbb{N}$ such that

$$\sup_{1 \leq j \leq n} \left(\sup_{x \in I_j} f(x) - \inf_{x \in I_j} f(x) \right) < \varepsilon, \qquad n \geq N. \tag{A.10}$$

This implies that

$$\overline{S}(Z_n, f) - \underline{S}(Z_n, f) = \sum_{j=1}^{n} \underbrace{(x_j - x_{j-1})}_{=\frac{b-a}{n}} \left(\sup_{x \in I_j} f(x) - \inf_{x \in I_j} f(x) \right) < \varepsilon(b-a), \tag{A.11}$$

for all $n \geq N$. Now notice that $\overline{A}(f) \leq \overline{S}(Z_n, f)$, $\underline{A}(f) \geq \underline{S}(Z_n, f)$, but on the other hand by definition $\overline{A}(f) \geq \underline{A}(f)$. This means that for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for $n \geq N$,

$$0 \leq \overline{A}(f) - \underline{A}(f) \leq \overline{S}(Z_n, f) - \underline{S}(Z_n, f) < \varepsilon(b-a). \tag{A.12}$$

Letting $\varepsilon \downarrow 0$, we see that $\overline{A}(f) = \underline{A}(f)$. □

The following theorem is one of the most celebrated results of real analysis in one variable, we state it without proof.

Theorem A.3 (Fundamental theorem of calculus). *(i) Let $f : [a,b] \to \mathbb{R}$ be continuous and define*

$$F(x) = \int_a^x f(t)\mathrm{d}t, \tag{A.13}$$

for $x \in (a,b]$, and $F(0) = 0$. Then F is uniformly continuous on $[a,b]$, differentiable on (a,b) and one has $F'(x) = f(x)$ for all $x \in (a,b)$.

[1] i.e. closed and bounded

(ii) If $f : [a, b] \to \mathbb{R}$ is differentiable, then f' is Riemann-integrable on $[a, b]$ and

$$\int_a^b f'(x)\mathrm{d}x = f(b) - f(a). \tag{A.14}$$

Let us recall (without proof) some standard properties of Riemann integrals.

Lemma A.4. *(i) Let $f \in \mathscr{R}([a, b])$ with $f \geq 0$. Then also $\int_a^b f(x)\mathrm{d}x \geq 0$.*

(ii) $\mathscr{R}([a, b])$ is a linear subspace of $\mathbb{B}([a, b])$, and for $\alpha, \beta \in \mathbb{R}$, we have

$$\int_a^b (\alpha f(x) + \beta g(x))\mathrm{d}x = \alpha \int_a^b f(x)\mathrm{d}x + \beta \int_a^b g(x)\mathrm{d}x. \tag{A.15}$$

(iii) If $f, g \in \mathscr{R}([a, b])$, also $|f|$ and $f \cdot g \in \mathscr{R}([a, b])$. Moreover,

$$\left| \int_a^b f(x)\mathrm{d}x \right| \leq \int_a^b |f(x)|\mathrm{d}x. \tag{A.16}$$

We now briefly discuss improper Riemann integrals, which we have used in Chapter 4 when discussing probability density functions.

Definition A.5. (i) Let $f : [a, \infty) \to \mathbb{R}$ with $a \in \mathbb{R}$. Assume that f is Riemann-integrable on $[a, c]$ for any $c > a$. We say that f is Riemann-integrable over $[a, \infty)$ if the limit

$$\int_a^\infty f(x)\mathrm{d}x = \lim_{c \to \infty} \int_a^c f(x)\mathrm{d}x \tag{A.17}$$

exists and is finite. Similarly, we define integrability of $f : (-\infty, a] \to \mathbb{R}$. Finally, we say $f : \mathbb{R} \to \mathbb{R}$ is Riemann-integrable over $\mathbb{R}$ if

$$\int_{-\infty}^\infty f(x)\mathrm{d}x = \lim_{s \to \infty} \lim_{r \to \infty} \int_{-r}^s f(x)\mathrm{d}x \tag{A.18}$$

exists and is finite. These integrals are *improper Riemann integrals of the first kind.*

(ii) Let $f : (a, b] \to \mathbb{R}$ and assume that $f|_{[c,b]} : [c, b] \to \mathbb{R}$ is Riemann-integrable for every $a < c < b$. We say that f is Riemann-integrable over $(a, b]$ if the limit

$$\lim_{\varepsilon \downarrow 0} \int_{a+\varepsilon}^b f(x)\mathrm{d}x \tag{A.19}$$

exists and is finite. Integrals over $(a, b]$ are defined in a similar way. These integrals are *improper Riemann integrals of the second kind.*

Remark A.6. Suppose that $f : \mathbb{R} \to [0, \infty)$ is a piecewise continuous Riemann-integrable function. We explain in detail why for all $a \in \mathbb{R}$, one has

$$\lim_{\delta \downarrow 0} \int_a^{a+\delta} f(x)\mathrm{d}x = 0. \tag{A.20}$$

- Case I: Suppose f is continuous in a. There exists $\delta > 0$ such that $|f(x) - f(a)| < 1$ for all $|x - a| < \delta$. Therefore

  $$\left|\int_a^{a+\delta} f(x)\mathrm{d}x\right| \leq \int_a^{a+\delta} |f(a)|\mathrm{d}x + \int_a^{a+\delta} \underbrace{|f(x) - f(a)|}_{<1}\mathrm{d}x < |f(a)|\delta + \delta. \tag{A.21}$$

  Letting $\delta \downarrow 0$ shows the claim.

- Case II: Suppose that f is not continuous in a. Since f is assumed to be piecewise continuous, either $f|_{(a,b]}$ is continuous for some $b > a$ or $f|_{[b,a)}$ is continuous for some $b < a$. We only treat the first case, since the second case can be argued for in the exact same way. By definition, we know that

  $$\lim_{n\to\infty} \int_{a+1/n}^{b} f(x)\mathrm{d}x \tag{A.22}$$

  exists. By the Cauchy criterion for convergent sequences, we see that for any $\eta > 0$, there exists $N \in \mathbb{N}$ such that whenever $n \geq N$, we have

  $$\left|\int_{a+1/n}^{b} f(x)\mathrm{d}x - \int_{a+1/N}^{b} f(x)\mathrm{d}x\right| = \int_{a+1/n}^{a+1/N} f(x)\mathrm{d}x < \eta. \tag{A.23}$$

  Now note that for any $\delta < \frac{1}{N}$ (depending on η) we have that (since f is non-negative)

  $$\int_{a+1/n}^{a+\delta} f(x)\mathrm{d}x \leq \int_{a+1/n}^{a+1/N} f(x)\mathrm{d}x < \eta, \qquad \text{for all } n \geq N. \tag{A.24}$$

  Passing to the limit yields

  $$\int_{a}^{a+\delta} f(x)\mathrm{d}x = \lim_{n\to\infty} \int_{a+1/n}^{a+\delta} f(x)\mathrm{d}x \leq \eta. \tag{A.25}$$

  This precisely means that $\lim_{\delta\downarrow 0} \int_a^{a+\delta} f(x)\mathrm{d}x = 0$.

A.2. Multiple integrals

We give some details and intuition about multiple integrals. Such multiple integrals appear in Chapter 7. The focus of this appendix is *not* to study the most general multiple integrals, but rather to give some examples and calculation rules, if multiple integrals are unfamiliar.

We first present a rigorous approach to multidimensional Riemann-integrals on rectangular sets, analogously to the one-dimensional case in the previous section.

Definition A.7. Let $-\infty < a_j < b_j < \infty$ for all $1 \leq j \leq n$, and abbreviate $[\boldsymbol{a}, \boldsymbol{b}] = \prod_{i=1}^n [a_i, b_i]$. We call a *rectangular set* any set Q of the form $[\boldsymbol{a}, \boldsymbol{b}]$. For such $Q = [a_1, b_1] \times \ldots \times [a_n, b_n]$, the *diameter* of Q is

$$\operatorname{diam}(Q) = \sqrt{\sum_{j=1}^n (b_j - a_j)^2}, \tag{A.26}$$

and the *volume* of Q is

$$\operatorname{Vol}(Q) = \prod_{j=1}^n (b_j - a_j). \tag{A.27}$$

Let $\mathbb{B}([\boldsymbol{a}, \boldsymbol{b}])$ be the set of bounded functions $f : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$.

(i) A *partition* Z_N of the rectangular set $Q = [\boldsymbol{a}, \boldsymbol{b}]$ is a family $(Q_\ell)_{\ell=1}^N$ of rectangular sets $Q_\ell \subseteq Q$ with

$$Q = \bigcup_{\ell=1}^N Q_\ell, \qquad \text{with } \mathring{Q}_\ell \cap \mathring{Q}_m = \varnothing, \text{ for } \ell \neq m. \tag{A.28}$$

We define for

$$|Z_N| = \max\{\operatorname{diam}(Q_\ell) \,:\, 1 \leq \ell \leq N\} \tag{A.29}$$

the *norm* of Z_N. We define

$$\mathscr{S} = \{Z_N \,:\, Z_N \text{ is a partition of } [\boldsymbol{a}, \boldsymbol{b}]\}. \tag{A.30}$$

(ii) Let $f \in \mathbb{B}([\boldsymbol{a}, \boldsymbol{b}])$ and $Z_N \in \mathscr{S}$ with $Z_N = (Q_\ell)_{\ell=1}^N$. We define the *lower Riemann sum*

$$\underline{S}(Z_N, f) = \sum_{\ell=1}^N \operatorname{Vol}(Q_\ell) \inf_{z \in Q_\ell} f(z), \tag{A.31}$$

and the *upper Riemann sum*

$$\overline{S}(Z_N, f) = \sum_{\ell=1}^N \operatorname{Vol}(Q_\ell) \sup_{z \in Q_\ell} f(z) \tag{A.32}$$

(note that these are finite since f is assumed to be bounded).

(iii) Let $f \in \mathbb{B}([\boldsymbol{a}, \boldsymbol{b}])$. We define the *lower Riemann integral* of f over $[\boldsymbol{a}, \boldsymbol{b}]$ as

$$\underline{A}(f) = \underline{\int_{[\boldsymbol{a},\boldsymbol{b}]}} f(x) \mathrm{d}^n x = \sup\{\underline{S}(Z, f) \,:\, Z \in \mathscr{S}\} \tag{A.33}$$

and the *upper Riemann integral* of f over $[\boldsymbol{a}, \boldsymbol{b}]$ as

$$\overline{A}(f) = \overline{\int_{[\boldsymbol{a},\boldsymbol{b}]}} f(x)\mathrm{d}^n x = \inf\{\overline{S}(Z, f) \,:\, Z \in \mathscr{S}\}. \tag{A.34}$$

We say that f is *Riemann-integrable* over $Q = [\boldsymbol{a}, \boldsymbol{b}]$ if $\underline{A}(f) = \overline{A}(f)$. In this case, the value $\underline{A}(f) = \overline{A}(f)$ is denoted $\int_Q f(x)\mathrm{d}^n x$ or $\int_{[\boldsymbol{a},\boldsymbol{b}]} f(x)\mathrm{d}^n x$.

We note that the statements in Theorem A.2 and Lemma A.4 hold in the multidimensional case as well. We now face the following problems:

- How can one calculate a multidimensional Riemann integral?
- How do we define improper multidimensional Riemann integrals?
- How can we define an integral like $\int_A f(x)\mathrm{d}^n x$ if $A \subseteq \mathbb{R}^n$ is *not* a rectangular set.

For the first issue, we note the following:

Theorem A.8. *Let $f : Q \to \mathbb{R}$ be a continuous[2] function on $Q = [\boldsymbol{a}, \boldsymbol{b}]$. Then*

$$\int_Q f(x)\mathrm{d}^n x = \int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} \dots \left(\int_{a_n}^{b_n} f(x_1, \dots, x_n)\mathrm{d}x_n \right) \dots \mathrm{d}x_2 \right) \mathrm{d}x_1. \tag{A.35}$$

We omit the proof of the above theorem. Instead, we give some more intuition on the practical use of multidimensional integrals.

Remark A.9. (i) Let $f : [a_1, b_1] \to [0, \infty)$, $a_1 < b_1$, a real-valued function. If $f \in \mathscr{R}([a_1, b_1])$, we know from the previous subsection that the integral

$$\int_{[a_1,b_1]} f(x_1)\mathrm{d}x_1 = \int_{a_1}^{b_1} f(x_1)\mathrm{d}x_1 \tag{A.36}$$

gives us the area that is enclosed by the graph of the function f, the two lines $x_1 = a_1$ and $x_2 = b_1$ and the x_1-axis (which can be expressed as the line $y = 0$). Moreover, we remark that in (A.36), we could also have expressed the integral $\int_{a_1}^{b_1} f(x_1)\mathrm{d}x_1$ as $\int_{[a_1,b_1)} f(x_1)\mathrm{d}x_1$, $\int_{(a_1,b_1]} f(x_1)\mathrm{d}x_1$ or $\int_{(a_1,b_1)} f(x_1)\mathrm{d}x_1$, since the boundary points do not contribute to the area.

(ii) Let $f : [a, b] \times [a', b'] \to [0, \infty)$, for some $a < b$, $a' < b'$, be a real function depending on two variables. One can visualize the expression $z = f(x_1, x_2)$ as a kind of "mountainous landscape", where the value z describes the height with respect to zero over the point $(x_1, x_2) \in [a, b] \times [a', b']$. If f is sufficiently regular we have

$$\int_{[a_1,b_1]\times[a_2,b_2]} f(x_1, x_2)\mathrm{d}^2 x = \int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} f(x_1, x_2)\mathrm{d}x_2 \right) \mathrm{d}x_1. \tag{A.37}$$

[2] Continuity means that for every sequences $(x_1^{(k)})_{k\geq 1} \subseteq [a_1, b_1]$,..., $(x_n^{(k)})_{k\geq 1} \subseteq [a_n, b_n]$ that fulfill $x_j^{(k)} \to x_j$, for all $1 \leq j \leq n$, as $k \to \infty$, we necessarily have $\lim_{k\to\infty} f(x_1^{(k)}, \dots, x_n^{(k)}) = f(x_1, \dots, x_n)$. In $n = 2$, this intuitively this means that the graph of f does not have any "fault lines".

(again, if f is continuous, this is fulfilled by Theorem A.8), and the expression in (A.37) can be considered as the volume enclosed by the graph of the function f, the four planes $x_1 = a_1$, $x_1 = b_1$, $x_2 = a_2$, $x_2 = b_2$ and the x_1-x_2-plane (which can be expressed as $z = 0$). A graphical representation is given in Figure A.9 below. Similarly as in the one-dimensional case, we could also write the expression in (A.37) as $\int_{[a_1,b_1)\times[a_2,b_2)} f(x_1,x_2)\mathrm{d}^2x$,$\int_{(a_1,b_1]\times(a_2,b_2)} f(x_1,x_2)\mathrm{d}^2x$, etc., since planes like $x_1 = a_1$, $x_1 = b_1$, $x_2 = a_2$ and $x_2 = b_2$ give no contribution to the volume.

(iii) The expression in (A.37) is evaluated from the inner integral to the outer integral. This means that we first perform the integration with respect to x_2, treating x_1 as a constant. The resulting function $x_1 \mapsto \int_{a_2}^{b_2} f(x_1,x_2)\mathrm{d}x_2$ is then integrated with respect to x_1, over the interval $[a_1,b_1]$. We demonstrate this in Example A.10 below.

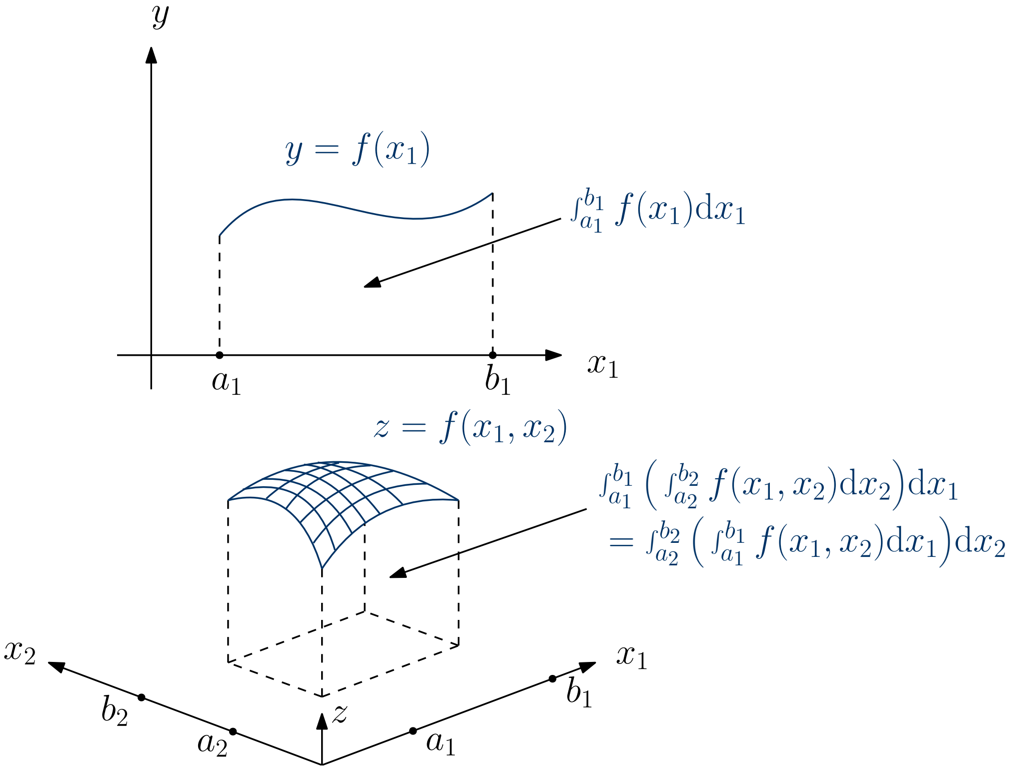
y
$y = f(x_1)$
$\int_{a_1}^{b_1} f(x_1)\mathrm{d}x_1$
x_1
a_1
b_1
$z = f(x_1,x_2)$
$\int_{a_1}^{b_1}\left(\int_{a_2}^{b_2} f(x_1,x_2)\mathrm{d}x_2\right)\mathrm{d}x_1$
$= \int_{a_2}^{b_2}\left(\int_{a_1}^{b_1} f(x_1,x_2)\mathrm{d}x_1\right)\mathrm{d}x_2$
x_2
x_1
b_2
a_2
z
a_1
b_1

Figure A.1.: Upper panel: The area under a curve for a function f depending on one variable. Lower panel: The volume under the graph for a function f depending on two variables.

(iv) It can be shown that if either $f \geq 0$, or the double integral of $|f|$ over $[a,b]\times[a',b']$ exists, the expression in (A.37) can also rewritten as follows:

$$\int_{a_1}^{b_1}\left(\int_{a_2}^{b_2} f(x_1,x_2)\mathrm{d}x_2\right)\mathrm{d}x_1 = \int_{a_2}^{b_2}\left(\int_{a_1}^{b_1} f(x_1,x_2)\mathrm{d}x_1\right)\mathrm{d}x_2. \tag{A.38}$$

In other words, the order in which we perform the integration does not matter. This fact is known as *Fubini's theorem.* Some care is required if we do not integrate f over a rectangle, but over more complicated domains (see the discussion below).

(v) Now let $f : \mathbb{R}^n \to \mathbb{R}$. For $a_i < b_i$, we can look at the expression[3]

$$\begin{aligned}\int_{[a_1,b_1]\times[a_2,b_2]\times\ldots\times[a_n,b_n]} & f(x_1,\ldots,x_n)\mathrm{d}^n x \\ &= \int_{a_1}^{b_1}\left(\int_{a_2}^{b_2}\ldots\left(\int_{a_n}^{b_n} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1.\end{aligned} \tag{A.39}$$

Again, this expression is evaluated from the inside to the outside. Note that is nothing else but (A.36) in the case of $n = 1$ or (A.37) in the case of $n = 2$. In the latter cases, we can also makes sense of negative values of the function f, by assigning the area below the x_1-axis a negative value and the volume below the x_1-x_2-plane a negative value. As in (A.38), we can perform the intregrals in any order we like. Mathematically, one has for every permutation $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ (a bijective map from $\{1, 2, \ldots, n\}$ to itself, which is nothing else but a reordering of the numbers from 1 to n):

$$\begin{aligned}&\int_{a_1}^{b_1}\left(\int_{a_2}^{b_2}\ldots\left(\int_{a_n}^{b_n} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_{a_{\sigma(1)}}^{b_{\sigma(1)}}\left(\int_{a_{\sigma(2)}}^{b_{\sigma(2)}}\ldots\left(\int_{a_{\sigma(n)}}^{b_{\sigma(n)}} f(x_1,\ldots,x_n)\mathrm{d}x_{\sigma(n)}\right)\ldots\mathrm{d}x_{\sigma(2)}\right)\mathrm{d}x_{\sigma(1)}.\end{aligned} \tag{A.40}$$

Finally, we can define the improper integral

$$\begin{aligned}\int_{\mathbb{R}^n} f(x_1,\ldots,x_n)\mathrm{d}^n x &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\ldots\left(\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1 \\ = \lim_{a_1\to\infty}\lim_{b_1\to\infty}\ldots\lim_{a_n\to\infty}\lim_{b_n\to\infty} & \int_{-a_1}^{b_1}\left(\int_{-a_2}^{b_2}\ldots\left(\int_{-a_n}^{b_n} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1\end{aligned} \tag{A.41}$$

Improper integrals in which only one boundary is infinity (which appear for instance when calculating joint cumulative distribution functions, see Proposition 7.9) are defined similarly. For instance:

$$\begin{aligned}&\int_{(-\infty,b_1]\times\ldots\times(-\infty,b_n]} f(x_1,\ldots,x_n)\mathrm{d}^n x \\ &= \int_{-\infty}^{b_1}\left(\int_{-\infty}^{b_2}\ldots\left(\int_{-\infty}^{b_n} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \lim_{a_1\to\infty}\ldots\lim_{a_n\to\infty}\int_{-a_1}^{b_1}\left(\int_{-a_2}^{b_2}\ldots\left(\int_{-a_n}^{b_n} f(x_1,\ldots,x_n)\mathrm{d}x_n\right)\ldots\mathrm{d}x_2\right)\mathrm{d}x_1\end{aligned} \tag{A.42}$$

[3] Or equivalently $\int_{[a_1,b_1)\times[a_2,b_2]\times\ldots\times[a_n,b_n]} f(x_1,\ldots,x_n)\mathrm{d}^n x$, $\int_{[a_1,b_1)\times[a_2,b_2)\times\ldots\times[a_n,b_n)} f(x_1,\ldots,x_n)\mathrm{d}^n x$ or any other number of expressions, where we only modify whether boundary values a_j or b_j are included.

Example A.10. (i) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x_1, x_2) = x_1 x_2^2$. We want to integrate f over the rectangle $[0,2] \times [0,1] \subseteq \mathbb{R}^2$. This integral is

$$\begin{aligned}\int_{[0,2]\times[0,1]} f(x_1,x_2)\mathrm{d}^2x &= \int_0^2 \left(\int_0^1 f(x_1,x_2)\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^2 \left(\int_0^1 x_1x_2^2\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^2 x_1\left[\frac{1}{3}x_2^3\right]_0^1 \mathrm{d}x_1 \\ &= \frac{1}{3}\int_0^2 x_1\mathrm{d}x_1 = \frac{1}{3}\left[\frac{1}{2}x_1^2\right]_0^2 = \frac{2}{3}.\end{aligned} \tag{A.43}$$

We have marked in the relevant inner integral all terms that play a role (x_1 is treated as a constant when integrating over x_2).

(ii) Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x_1, x_2, x_3) = \cos(x_1)x_2\exp(x_2x_3)$. We integrate f over the cuboid $[0,1] \times [2,3] \times [1,4]$.

$$\begin{aligned}\int_{[0,1]\times[2,3]\times[1,4]} f(x_1,x_2,x_3)\mathrm{d}^3x &= \int_0^1\left(\int_2^3\left(\int_1^4 f(x_1,x_2,x_3)\mathrm{d}x_3\right)\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^1\left(\int_2^3\left(\int_1^4 \cos(x_1)x_2\exp(x_2x_3)\mathrm{d}x_3\right)\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^1\left(\int_2^3 \cos(x_1)x_2\left[\frac{1}{x_2}\exp(x_2x_3)\right]_{x_3=1}^4 \mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^1 \cos(x_1)\left(\int_2^3 (\exp(4x_2)-\exp(x_2))\mathrm{d}x_2\right)\mathrm{d}x_1 \\ &= \int_0^1 \cos(x_1)\left[\frac{1}{4}\exp(4x_2)-\exp(x_2)\right]_{x_2=2}^3 \mathrm{d}x_1 \\ &= \left(\frac{1}{4}e^8(e^4-1)-(e-1)e^2\right)[\sin(x_1)]_{x_1=0}^1 \\ &= \left(\frac{1}{4}e^8(e^4-1)-(e-1)e^2\right)\sin(1).\end{aligned} \tag{A.44}$$

Note that in the innermost integral in the second line, although x_2 was treated as constant, one has to be a bit careful since we cannot divide by 0. This is however excluded, since $x_2 \in [2,3]$.

So far, we only discussed integrals over rectangular sets $Q = [a_1, b_1] \times \ldots \times [a_n, b_n]$. In one dimension, this is what we really need. However, in dimension $n \geq 2$, we may want to integrate over more general sets $A \subseteq \mathbb{R}^n$, i.e. calculate integrals of the form

$$\int_A f(x_1, \ldots, x_n)\mathrm{d}^n x. \tag{A.45}$$

To make progress, we recall for $A \subseteq \mathbb{R}^d$ the definition of the indicator function

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases} \tag{A.46}$$

We will define the integral of a function f, defined on $Q = [\boldsymbol{a}, \boldsymbol{b}]$, over a subset $A \subseteq Q$ as

$$\int_A f(x) \mathrm{d}^n x = \int_Q \mathbb{1}_A(x) f(x) \mathrm{d}^n x. \tag{A.47}$$

The problem is that even if f is continuous, the function $x \mapsto \mathbb{1}_A(x) f(x)$ is typically not continuous. However, if the set of discontinuities is sufficiently “sparse”, we may still be able to perform this integral. This leads to another definition.

Definition A.11. A set $M \subseteq \mathbb{R}^n$ is called a *Jordan nullset* if for every $\varepsilon > 0$, there are finitely many rectangular sets $(Q_j)_{j=1}^r$ in $\mathbb{R}^n$ such that

$$M \subseteq \bigcup_{j=1}^r Q_j, \qquad \text{and} \qquad \sum_{j=1}^r \mathrm{Vol}(Q_k) < \varepsilon. \tag{A.48}$$

A set $S \subseteq \mathbb{R}^n$ is called *Jordan-measurable* if its topological boundary ∂S is a Jordan nullset.

Example A.12. (i) Every finite subset of $\mathbb{R}^n$ is a Jordan nullset (every point can be covered by a rectangular set with arbitrarily small volume).

(ii) Any subset of a Jordan nullset is a Jordan nullset.

(iii) In $\mathbb{R}^n$, every $(n-1)$-dimensional rectangular set A is a Jordan-nullset. Indeed, without loss of generality suppose that $A = \{0\} \times [a_2, b_2] \times \ldots \times [a_n, b_n]$, then $A \subseteq [-\varepsilon, \varepsilon] \times [a_2, b_2] \times \ldots \times [a_n, b_n] = Q_\varepsilon$, with $\mathrm{Vol}(Q_\varepsilon) = 2\varepsilon \prod_{j=2}^n (b_j - a_j)$, which can be made arbitrarily small if $\varepsilon > 0$ is made small enough.

(iv) Finite unions of Jordan nullsets are Jordan nullsets.

(v) Let f be a continuous function on a compact set $A \subseteq \mathbb{R}^n$. Then the graph of f, defined as

$$\Gamma(f) = \{(x, f(x)) \, : \, x \in A\}, \subseteq \mathbb{R}^{n+1} \tag{A.49}$$

is a Jordan nullset. This follows essentially from the fact that f is uniformly continuous on A.

Theorem A.13. *Let $Q = [\boldsymbol{a}, \boldsymbol{b}] \subseteq \mathbb{R}^n$ be a rectangular set and suppose that $f \in \mathbb{B}([\boldsymbol{a}, \boldsymbol{b}])$. If the set*

$$D_f = \{x \in Q \, : \, f \text{ is not continuous at } x\} \tag{A.50}$$

is a Jordan nullset, then f is Riemann-integrable over Q.

With this in mind, we can turn (A.47) into a definition.

Definition A.14. Let $A \subseteq Q$ for some rectangular set Q and $f : A \to \mathbb{R}$ be a bounded function. We say that f is *Riemann-integrable over A* if $x \mapsto \mathbb{1}_A f(x)$ is Riemann-integrable over Q (in this product, f is extended arbitrarily outside of A). In this case, we define

$$\int_A f(x) \mathrm{d}^n x = \int_Q \mathbb{1}_A(x) f(x) \mathrm{d}^n x. \tag{A.51}$$

With the previous theorem, one can now show

Corollary A.15. *Let $S \subseteq \mathbb{R}^n$ be a bounded, Jordan measurable set and $f : S \to \mathbb{R}$ a bounded function such that D_f is a Jordan nullset. Then f is Riemann-integrable over S.*

Proof. The set of discontinuities of the function $x \mapsto \mathbb{1}_A(x) f(x)$ (defined on $Q \supseteq A$) is contained in $\partial S \cup D_f$. The latter is a Jordan nullset, since both ∂S and D_f are Jordan nullsets by assumption. □

Remark A.16. (i) Suppose we are given the function $f : \mathbb{R}^2 \to \mathbb{R}$, assumed to be sufficiently regular, which we want to integrate over the following area A, given by

$$A = \{(x_1, x_2)\,;\, x_1 \in [a_1, b_1], x_2 \in [u(x_1), v(x_1)]\}. \tag{A.52}$$

Here $u : [a_1, b_1] \to \mathbb{R}$ and $v : [a_1, b_1] \to \mathbb{R}$ are picewise continuous functions that fulfill $u(x_1) \leq v(x_1)$ for every $x_1 \in [a_1, b_1]$. We can define the integral

$$\int_A f(x_1, x_2) \mathrm{d}^2 x = \int_{a_1}^{b_1} \left(\int_{u(x_1)}^{v(x_1)} f(x_1, x_2) \mathrm{d}x_2 \right) \mathrm{d}x_1. \tag{A.53}$$

In the case where $f \geq 0$, we can interpret this integral again as a volume, namely of the set bounded by the graph of the function $f|_A : A \to [0, \infty)$, the planes $x_1 = a_1$, $x_1 = b_1$, the x_1-x_2-plane and the “curved planes” given by the relations $x_2(x_1) = u(x_1)$ and $x_2(x_1) = v(x_1)$. Figure A.16 below gives a graphical representation of the situation.

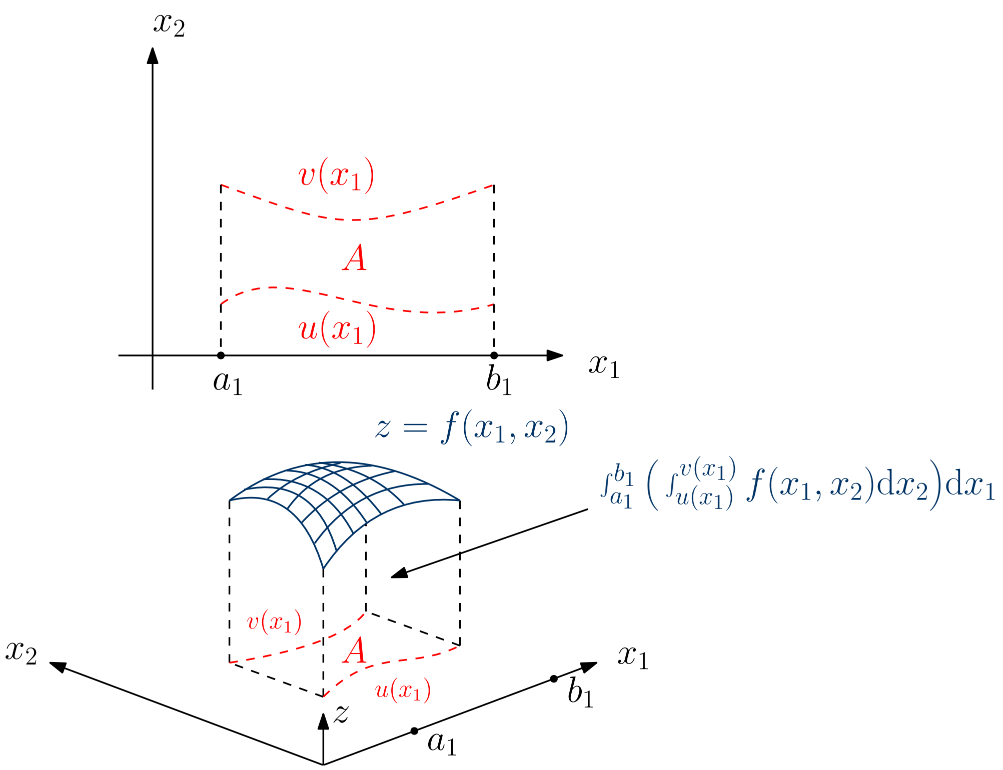

Figure A.2.: Upper panel: The area A over which we want to integrate over. Lower panel: The volume under the graph for a function f depending on two variables, integrated over A. The upper panel is the picture seen in the x_1-x_2-plane of the lower panel.

(ii) More generally, we can consider integrals of (sufficiently regular) functions $f : \mathbb{R}^n \to \mathbb{R}$ over sets of the form

$$A = \{(x_1, ..., x_n) \, ; \, x_1 \in [a_1, b_1], x_2 \in [u_2(x_1), v_2(x_2)], \\ ..., x_n \in [u_n(x_1, ..., x_{n-1}), v_n(x_1, ..., x_{n-1})]\}, \tag{A.54}$$

where the functions $u_2, v_2, u_3, v_3, ..., u_n, v_n$ have to be sufficiently regular themselves and fulfill $u_j \leq v_j$ for every $j = 2, ..., n$. We set

$$\begin{aligned} &\int_A f(x_1, x_2, ..., x_n) \mathrm{d}^n x \\ &= \int_{a_1}^{b_1} \left(\int_{u_2(x_1)}^{v_2(x_1)} \left(... \left(\int_{u_n(x_1,...,x_{n-1})}^{v_n(x_1,...,x_{n-1})} f(x_1, x_2, ..., x_n) \mathrm{d}x_n \right) ... \right) \mathrm{d}x_2 \right) \mathrm{d}x_1 . \end{aligned} \tag{A.55}$$

(iii) Let us remark that in dimension $n = 2$ the expression $\int_A 1 \mathrm{d}^2 x = \int_{a_1}^{b_1} \left(\int_{u(x_1)}^{v(x_1)} \mathrm{d}x_2 \right) \mathrm{d}x_1$ (corresponding to setting $f = 1$ in (A.53)) is exactly the area of the set A in (A.52). Formally this corresponds to the volume of a prism with base area A and height 1. In a similar manner, in dimension $n = 3$, the expression $\int_V 1 \mathrm{d}^3 x = \int_{a_1}^{b_1} \left(\int_{u_2(x_1)}^{v_2(x_2)} \left(\int_{u_3(x_1,x_2)}^{v_3(x_1,x_2)} f(x_1, x_2, x_3) \mathrm{d}x_3 \right) \mathrm{d}x_2 \right) \mathrm{d}x_1$

gives us the volume of the set

(A.56)
$$V = \{(x_1, x_2, x_3)\,;\, x_1 \in [a_1, b_1], x_2 \in [u_2(x_1), v_2(x_1)], x_3 \in [u_3(x_1, x_2), v_3(x_1, x_2)]\}.$$

Example A.17. (i) Let us first calculate the integral (we write x, y, z instead of x_1, x_2, x_3):

(A.57)
$$\begin{aligned}\int_0^1 \left(\int_0^{2x} \left(\int_0^{x+y} 1\mathrm{d}z\right) \mathrm{d}y\right) \mathrm{d}x &= \int_0^1 \left(\int_0^{2x} (x+y)\mathrm{d}y\right) \mathrm{d}x \\ &= \int_0^1 \left[xy + \frac{1}{2}y^2\right]_{y=0}^{2x} \mathrm{d}x \\ &= 4\int_0^1 x^2 \mathrm{d}x \\ &= 4\left[\frac{1}{3}x^3\right]_0^1 = \frac{4}{3}.\end{aligned}$$

This calculation gives us the volume of the set

(A.58)
$$V = \{(x, y, z)\,;\, 0 \le x \le 1, 0 \le y \le 2x, 0 \le z \le x + y\}.$$

(ii) We give some more details in the calculations appearing in (7.31) and (7.34). Here, we are given the information that X, Y are random variables with a joint density $f_{X,Y}$. This means that for a given set $B \in \mathscr{B}(\mathbb{R}^2)$, one has

(A.59)
$$\mathbf{P}[(X, Y) \in B] = \mathbf{P}_{(X,Y)}[B] = \int_B f_{X,Y}(x_1, x_2)\mathrm{d}^2x,$$

see (7.19). We wanted to find the value $\mathbf{P}[X^2 \le Y]$. Define the set

(A.60)
$$A = \{(x, y)\,;\, x \in [0, 1], y \in [x^2, 1]\},$$

which is of the form (A.52) for $u(x) = x^2$ and $v(x) = 1$. Moreover, we know that $\mathbf{P}\big[(X, Y) \in [0, 1]^2\big] = 1$. This means that

(A.61)
$$\begin{aligned}\mathbf{P}[X^2 \le Y] &= \mathbf{P}[\{X^2 \le Y\} \cap \{(X, Y) \in [0, 1]^2\}] \\ &= \mathbf{P}_{(X,Y)}[\{(x, y)\,;\, x^2 \le y \text{ and } 0 \le x, y \le 1\}] \\ &= \mathbf{P}_{(X,Y)}[A] \\ &= \int_0^1 \left(\int_{x^2}^1 f_{X,Y}(x, y)\mathrm{d}y\right) \mathrm{d}x.\end{aligned}$$

The rest of the calculation proceeds as in (7.31) or (7.34).

A.3. Alternative Proof of the central limit theorem 12.4

Lemma A.18. *Let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbf{E}[X_1] = 0$, $\mathrm{Var}[X_1] = 1$ and $g : \mathbb{R} \to \mathbb{R}$ twice differentiable such that g'' is uniformly continuous and bounded. For $Z_n = \sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ and $Z \sim \mathcal{N}(0, 1)$, one has*

(A.62)
$$\mathbf{E}[g(Z_n)] \xrightarrow[n\to\infty]{} \mathbf{E}[g(Z)].$$

Proof. The idea is to replace $X_1, ..., X_n$ step-by-step by i.i.d. random variables that have a standard normal distribution, and to use Taylor's theorem.

Consider $Y_1, ..., Y_n \sim \mathcal{N}(0,1)$ i.i.d. random variables such that $X_1, ..., X_n, Y_1, ..., Y_n$ are stochastically independent. This means that

$$\sqrt{n}\overline{Y}_n \sim \mathcal{N}(0,1) \qquad \Rightarrow \qquad \mathbf{E}\left[g(\sqrt{n}\overline{Y}_n)\right] = \mathbf{E}[g(Z)]. \tag{A.63}$$

Now we replace X_i by Y_i as follows:

$$\begin{aligned} g(Z_n) - g(\sqrt{n}\overline{Y}_n) &= g\left(\frac{X_1 + ... + X_n}{\sqrt{n}}\right) - g\left(\frac{Y_1 + ... + X_n}{\sqrt{n}}\right) \\ &+ g\left(\frac{Y_1 + ... + X_n}{\sqrt{n}}\right) - g\left(\frac{Y_1 + Y_2 + ... + X_n}{\sqrt{n}}\right) \\ &\vdots \\ &+ g\left(\frac{Y_1 + ... + Y_{n-1} + X_n}{\sqrt{n}}\right) - g\left(\frac{Y_1 + ... + Y_n}{\sqrt{n}}\right) \\ &= V_1 + ... + V_n, \end{aligned} \tag{A.64}$$

where we set

$$V_i = g\left(U_i + \frac{X_i}{\sqrt{n}}\right) - g\left(U_i + \frac{Y_i}{\sqrt{n}}\right), \tag{A.65}$$

and

$$U_i = \frac{Y_1 + ... + Y_{i-1} + X_{i+1} + ... + X_n}{\sqrt{n}}, \qquad 1 \le i \le n. \tag{A.66}$$

A Taylor expansion of g around the point U_i yields

$$V_i = \frac{X_i - Y_i}{\sqrt{n}} g'(U_i) + \frac{1}{2n} X_i^2 g''\left(U_i + \frac{\theta_1 X_i}{\sqrt{n}}\right) - \frac{1}{2n} Y_i^2 g''\left(U_i + \frac{\theta_2 Y_i}{\sqrt{n}}\right), \tag{A.67}$$

where $\theta_1, \theta_2 \in [0,1]$ are random variables. Consider the modulus of continuity

$$\begin{aligned} &\delta(h) = \sup_{|x-y| \le h} |g''(x) - g''(y)| \\ \Rightarrow \qquad &V_i = \frac{X_i - Y_i}{\sqrt{n}} g'(U_i) + \frac{1}{2n}(X_i^2 - Y_i^2) g''(U_i) + R_i, \end{aligned} \tag{A.68}$$

and we have

$$|R_i| \le \frac{1}{2n} X_i^2 \delta\left(\frac{|X_i|}{\sqrt{n}}\right) + \frac{1}{2n} Y_i^2 \delta\left(\frac{|Y_i|}{\sqrt{n}}\right). \tag{A.69}$$

Note that the random variables U_i and $X_i - Y_i$, resp. U_i and $X_i^2 - Y_i^2$ are independent. By (A.67), we find that

$$\mathbf{E}[V_i] = \frac{1}{\sqrt{n}} \underbrace{\mathbf{E}\left[(X_i - Y_i) g'(U_i)\right]}_{= \mathbf{E}[X_i - Y_i]\mathbf{E}[g'(U_i)] = 0} + \frac{1}{2n} \underbrace{\mathbf{E}\left[(X_i^2 - Y_i^2) g''(U_i)\right]}_{= \mathbf{E}[X_i^2 - Y_i^2]\mathbf{E}[g''(U_i)]} + \mathbf{E}[R_i] \tag{A.70}$$

Now it holds that

$$|\mathbf{E}[R_i]| \leq \mathbf{E}[|R_i|] \leq \frac{1}{2n}\mathbf{E}\left[X_i^2\delta\left(\frac{|X_i|}{\sqrt{n}}\right)\right] + \frac{1}{2n}\mathbf{E}\left[X_i^2\delta\left(\frac{|Y_i|}{\sqrt{n}}\right)\right]. \tag{A.71}$$

Now we estimate

$$\begin{aligned}\mathbf{E}\left[X_i^2\delta\left(\frac{|X_i|}{\sqrt{n}}\right)\right] &= \mathbf{E}\left[X_i^2\delta\left(\frac{|X_i|}{\sqrt{n}}\right)\left(\mathbb{1}_{\{|X_i|\leq\lambda\sqrt{n}\}} + \mathbb{1}_{\{|X_i|>\lambda\sqrt{n}\}}\right)\right] \\ &\leq \delta(\lambda)\mathbf{E}[X_i^2] + C\mathbf{E}[X_i^2\mathbb{1}_{\{|X_i|>\lambda\sqrt{n}\}}],\end{aligned} \tag{A.72}$$

where $C > 0$ is some constant. We have used that δ is finite (since g'' is bounded). Moreover, since g'' is uniformly continuous, we know that $\lim_{\lambda\downarrow 0}\delta(\lambda) = 0$. On the other hand, for fixed $\lambda > 0$, we have

$$\lim_{n\to\infty}\mathbf{E}[X_i^2\mathbb{1}_{\{|X_i|>\lambda\sqrt{n}\}}] = 0, \tag{A.73}$$

(this follows from the *dominated convergence theorem*). Combining everything, we see that

$$\mathbf{E}[R_i] \leq C\frac{1}{n}\cdot(\lambda + u_n), \tag{A.74}$$

where $\lim_{n\to\infty} u_n = 0$. Thus:

$$\lim_{n\to\infty}|g(Z_n) - g(\sqrt{n}Y_n)| \leq \lim_{n\to\infty} n|\mathbf{E}[V_1]| = \lim_{n\to\infty} u_n = 0. \tag{A.75}$$

□

Proof of Theorem 12.4. We first observe that since $X_1, X_2, \ldots$ are i.i.d. with $\mathbf{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2$, the random variables $X_i' = \frac{X_i-\mu}{\sigma}$ fulfill

$$\mathbf{E}[X_1'] = 0, \qquad \mathrm{Var}[X_1'] = 1, \qquad \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \sqrt{n}\overline{X'}_n. \tag{A.76}$$

Therefore, it suffices to show the statement of Lemma A.18 also for the functions $g = \mathbb{1}_{(-\infty,x]}$, $x \in \mathbb{R}$. Indeed, if (A.62) holds for such g, we have

$$F_{\frac{\overline{X}_n-\mu}{\sigma}}(x) = \mathbf{E}\left[\mathbb{1}_{(-\infty,x]}\left(\sqrt{n}\overline{X'}_n\right)\right] \xrightarrow[n\to\infty]{} \mathbf{E}\left[\mathbb{1}_{(-\infty,x]}(Z)\right] = F_Z(x) = \Phi(x). \tag{A.77}$$

So fix $g = \mathbb{1}_{(-\infty,x]}$ and $\varepsilon > 0$. We choose g_1, g_2 such that g_1'' and g_2'' are both uniformly continuous and bounded and fulfill

$$\begin{gathered} g_1(t) \leq g(t) \leq g_2(t) \\ \int_{-\infty}^{\infty}(g_2(t) - g_1(t))\mathrm{d}t \leq \varepsilon. \end{gathered} \tag{A.78}$$

It follows that with $Z_n = \sqrt{n}\overline{X'}_n$, one has

$$\mathbf{E}[g_1(Z_n)] \leq \mathbf{E}[g(Z_n)] \leq \mathbf{E}[g_2(Z_n)]. \tag{A.79}$$

The left- and rightmost items in the previous equation converge according to Lemma A.18 towards $\mathbf{E}[g_1(Z)]$ and $\mathbf{E}[g_2(Z)]$ respectively, and we have

$$\mathbf{E}[g_1(Z)] \leq \mathbf{E}[g(Z)] \leq \mathbf{E}[g_2(Z)]. \tag{A.80}$$

By our choice of g_1 and g_2, we also have

$$\mathbf{E}[g_2(Z) - g_1(Z)] = \int_{-\infty}^{\infty} (g_2(t) - g_1(t))\varphi(t)\mathrm{d}t \leq \varphi(0)\varepsilon. \tag{A.81}$$

It follows that

$$\limsup_{n\to\infty} |\mathbf{E}[g(Z_n)] - \mathbf{E}[g(Z)]| \leq \varphi(0)\varepsilon. \tag{A.82}$$

Since the latter holds for every $\varepsilon > 0$, we have proved the convergence (A.77). □

Textbook references

[Bré88] P. Brémaud, *An Introduction to Probabilistic Modeling*, Springer, 1988.

[DS10] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, Addison-Wesley, 2010.

[Dur09] R. Durrett, *Elementary Probability for Applications*, Cambridge University Press, 2009.

[Geo12] H.-O. Georgii, *Stochastics: Introduction to Probability and Statistics, 2nd ed.* De Gruyter, 2012.

[GS01] G. Grimmett and D. Stirzaker, *Probability and Random Processes, 3rd ed.*, Oxford, 2001.

Other references

[i] X. Geng, *The Mathematical Theory of Moment Generating Functions.* Lecture Notes at University of Melbourne, available at https://researchers.ms.unimelb.edu.au/∼xgge/notes.html.