Probability Theory

University of Groningen

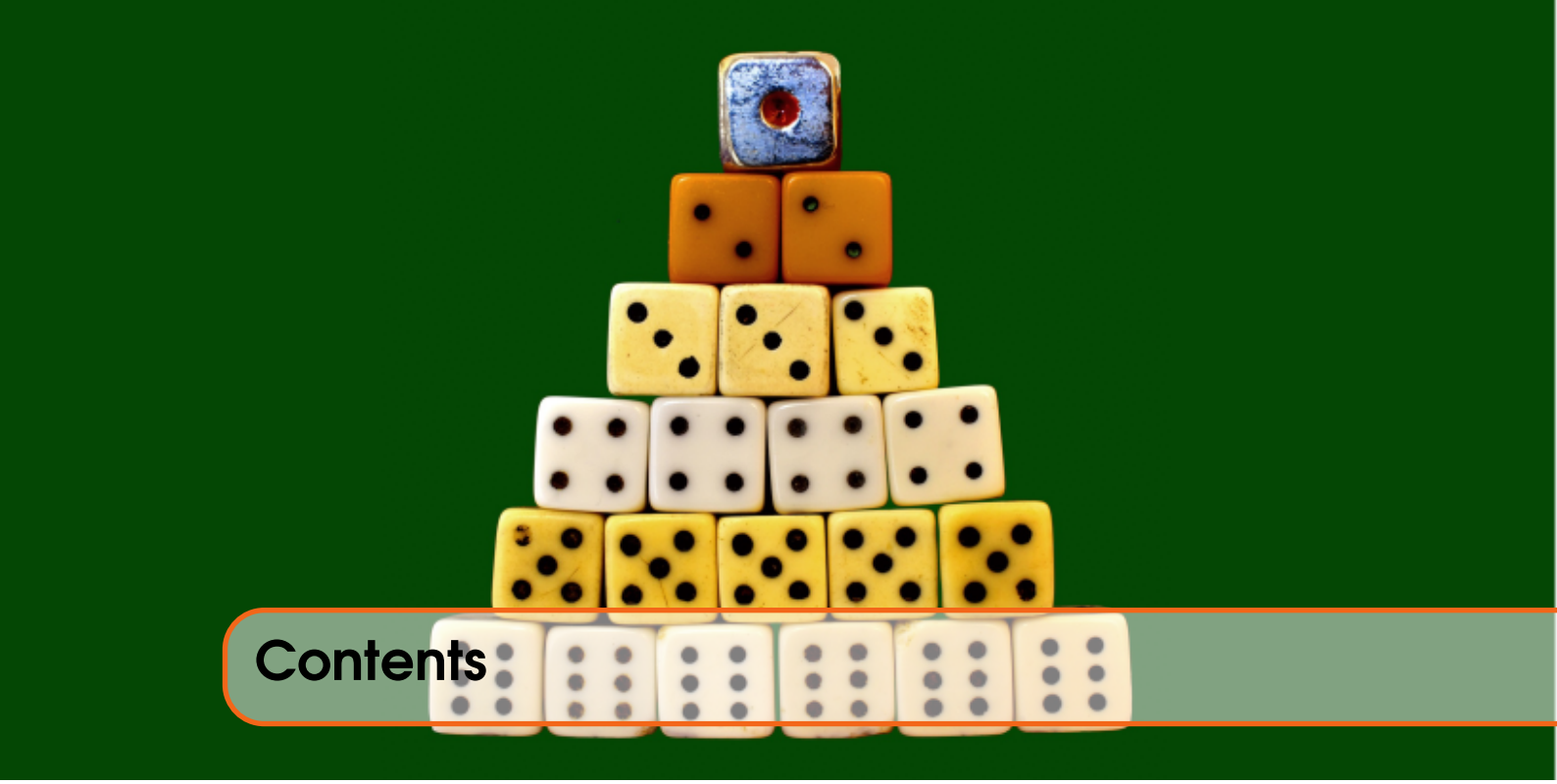
2023/2024 Period 2b

Version April 9, 2024

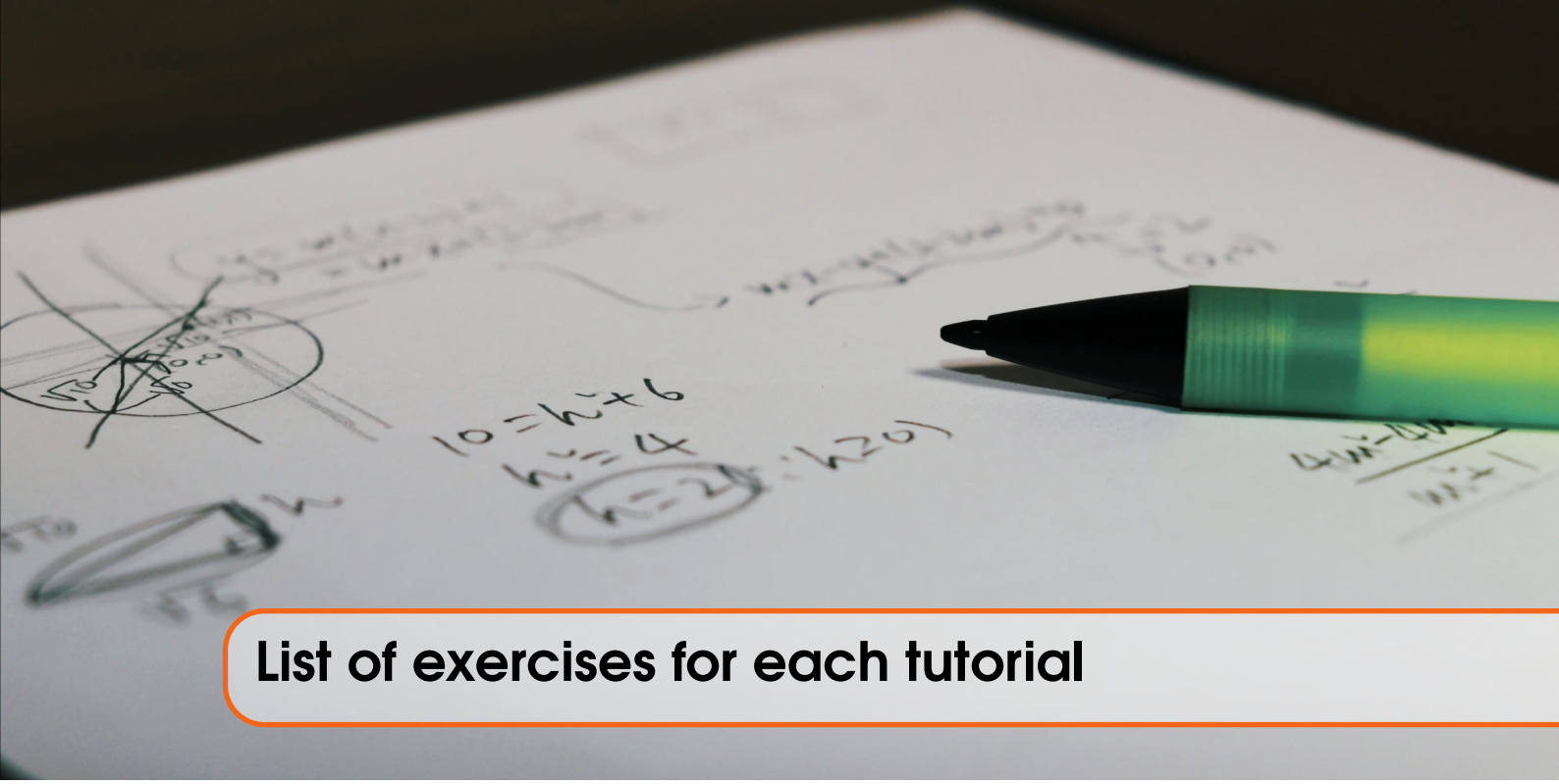
Dr. Gilles Bonnet

These lecture notes are based on the notes from earlier instances of this course taught by Daniel Valesin (14/15, 15/16 and 16/17), Tobias Müller (17/18 and 18/19), Christian Hirsch (19/20 and 20/21) and myself (21/22, 22/23). Praise should be addressed to Daniel, Tobias and Christian; complains to me.

If you notice a typo or something incorrect you are very welcome to report it at g.f.y.bonnet@rug.nl with the subject: "Probability Theory: feedback". This way you will contribute to the continuous improvement of these notes.



Contents



List of exercises for each tutorial

Remember that this document might be updated, this include the list below. Exercises listed in future tutorials is an indication of the lecturer's initial plan. This might change depending on the course's progress and the students' feedback.

Tutorial 1

Exercises 1.1 to 1.6.

Tutorial 2

Exercises 1.7 to 1.9, 2.1 and 2.5

Tutorial 3

Exercises 2.7 and 3.1 to 3.4

Tutorial 4

Exercises 4.1 and 4.3 to 4.6

Tutorial 5

Exercises 4.2, 4.7 and 5.1 to 5.5

Tutorial 6

Exercises 5.6 and 6.1 to 6.3

Tutorial 7

Exercises 6.4 to 6.6, 7.1 and 7.2

Tutorial 8

Exercises 8.1 to 8.3

Tutorial 9

Exercises 9.1, 9.3 to 9.5, 9.8 and 10.4

Tutorial 10

Exercises 9.2, 9.6 and 9.7

Tutorial 11

Exercises 9.9, 9.10, 10.1 and 10.2

Tutorial 12

Exercises 10.3, 11.1, 11.2 and 12.8

Tutorial 13

Exercises 8.4 and 12.3 to 12.6

Tutorial 14

Exercises 12.1, 12.2, 12.7, 13.1 and 13.2

Tutorial 15

Exercises 15.1 to 15.3

Tutorial 16

Q&A

Trivia

Some of the tutorials for this course are in the building “Bernoulliborg”, which also houses the Bernoulli institute for Mathematics, Computer Science and Artificial Intelligence. The Bernoulli’s were a family of Swiss (and Dutch/Belgian) mathematicians. Johann Bernoulli (1667–1748) was a professor at RuG. His older brother Jacob (1655–1705) is known for amongst other things the “discovery” of the number e and his pioneering work in probability theory. Eight years after his death his book “Ars Conjectandi” (Latin: the art of guessing) was published, one of the first works on probability.

Figure 1: Jakob Bernoulli

The very first such work on probability theory was “Van Rekeningh in Spelen van Gluck” (Dutch: of computation in games of chance) by Christiaan Huygens, a Dutch mathematician, in 1657.

So the serious study of probability theory only started in the 17th century. This is remarkable since probability is basically everywhere around us “in nature” and other subjects like geometry and number theory go back to even before the ancient Greeks. In fact games involving chance have

been played at least since around 3000BC, when the Egyptians were already playing a game where they threw “astragali”, little bones from the ankle of a goat, and were betting on which side of the bone would come up on top. Apparently it took until the 17th century before people started to realize that it could be helpful to compute the likelihood of various outcomes when betting on such “games of chance”.

1. Combinatorics

The goals of this chapter are:

- ▷ to **compute the number of possible outcomes** or various concrete experiments;
- ▷ to define the following notation: **factorial** $n!$, **falling factorial** $(n)_k$, **binomial coefficient** $\binom{n}{k}$;
- ▷ to introduce the (counting version of) the **inclusion-exclusion formula**.

1.1 So you think you can count?

As most people that have played a game involving rolling a dice or shuffling playing cards will realize, probability is intimately related to counting. We start by reviewing some basic, but crucial counting problems.

Question 1.1 How many ways are there to put the numbers $1,\dots,n$ in a sequence?

Answer: You have n choices for the first number in the sequence. Having chosen the first number, there are $n-1$ choices left for the second number. Having chosen the first and second number, there are $n-2$ numbers left for the third number, and so on. So the final answer is:

$$n! := n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1.$$

The notation $n!$ is pronounced "n-**factorial**" and the notation $:=$ means "is defined by". ■

Such an ordering of the numbers $1,\dots,n$ is called a *permutation*, by the way.

You can imagine a lotto machine with n balls inside numbered $1,\dots,n$. The machine spits out balls in a sequence. For now we only count the number of sequences that could possibly occur, we are not yet concerned with probabilities.

A lotto machine.

In lotto we usually do not draw all the balls out, but some number $k \leq n$.

Question 1.2 How many ways are there to pick a sequence of k distinct numbers chosen from $1,\ldots,n$?

Answer: This is similar to above. You have n choices for the first number in the sequence, $n-1$ choices left for the second number, etc. But now you stop once k numbers have been chosen. So the answer is:

$$(n)_k := n \cdot (n-1) \cdot \cdots \cdot (n-k+1).$$

The notation $(.)_k$ is sometimes called "falling factorial". People are sometimes confused that you stop at $n-k+1$ not $n-k$. This is of course because you started with n, so that $n, n-1, \ldots, n-k+1$ are k numbers while $n, \ldots, n-k$ would be $k+1$ numbers.
Observe that we can write $(n)_k$ in terms of factorials as

$$(n)_k = \frac{n(n-1)\ldots 1}{(n-k)(n-k-1)\ldots 1} = \frac{n!}{(n-k)!}.$$

■

Note the word "distinct" in the above question is crucial.

Question 1.3 How many ways are there to pick a sequence of k (not necessarily distinct) numbers chosen from $1,\ldots,n$?

Answer: You have n choices for the first number in the sequence, n choices left for the second number, n choices for the third number, etc. So the answer is:

$$n^k = \underbrace{n \cdot \ldots \cdot n}_{k \text{ times}}.$$

■

To make the distinction between the last two cases we sometimes say "the balls are drawn without/with replacement". This is because in the second case you can imagine that after a ball is taken out of the lotto machine, you write down its number and then put it back in the machine before the next ball is drawn.

Imagine now that (we are again drawing without replacement and) after a sequence of k balls is generated the balls are put in a bag and the bag is shaken. We can no longer tell the order in which the k balls arrived. In this case we speak of "drawing without replacement and without order". The

information we have corresponds to a subset of $\{1,\dots,n\}$ of size k.

Question 1.4 How many subsets of $\{1,\dots,n\}$ of cardinality exactly k are there?

Answer: If you have k balls in your bag then they could have arrived in $k!$ different orders (by the first question). Put differently, any subset of size k can be arranged as a sequence in $k!$ ways. On the other hand, each sequence of length k obviously belongs to exactly one subset. (A sequence determines a subset by just forgetting the order.) We can write

$$\#\text{ subsets of size } k \cdot k! = \#\text{ sequences of length } k = (n)_k.$$

(Using the Question 1.2.) So we find

$$\#\text{ subsets of size } k = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!} =: \binom{n}{k}.$$

The notation $\binom{n}{k}$ is pronounced "n choose k", and $\binom{n}{k}$ is called a *binomial coefficient*. ■

With counting problems it is sometimes useful to translate the problem to a setting of binary vectors. Recall that $\{0,1\}^n$ denotes the set of all vectors $x = (x_1,\dots,x_n)$ of length n with $x_i \in \{0,1\}$.

Question 1.5 How many vectors in $\{0,1\}^n$ have exactly k ones?

Answer: The answer is $\binom{n}{k}$, which can be seen by noting that the number of such vectors is the same as the number of subsets of $\{1,\dots,n\}$ of size k. This in turn is easily seen via the correspondence between sets $A \subseteq \{1,\dots,n\}$ and vectors $x \in \{0,1\}^n$ given by

$$x_i = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{if } i \notin A. \end{cases}$$

Every set corresponds to precisely one vector and vice-versa (in other words have constructed a bijection). Moreover, a set has cardinality k if and only if the corresponding vector has precisely k ones. ■

Finally we consider the problem of counting/sampling with replacement and without order. That is, there are n balls, we sample k of them putting each ball back before sampling the next and we just write down how many times each of the n balls was drawn. So we make a list with n numbers ranging from zero to k, that tracks how many times each balls occurred. The sum of all the numbers in the list will equal k.

Question 1.6 How many ways are there to draw k balls out of $1,\dots,n$ with replacement but without order?

Answer: The answer is

$$\binom{k+n-1}{k}.$$

One way to see this is as follows. Consider the list described above. We will turn it into a vector in $x \in \{0,1\}^m$ for suitable m. Let k_i denote the number of times ball i was drawn. The first k_1 entries of x will be ones, then we put a single zero, next we put k_2 many ones, then a zero, then k_3 many ones, and so on. We do not add a zero after the ones corresponding to k_n, because that does not add any information. (Note that there can be two consecutive zeroes which indicates

the corresponding k_i equals zero.)

How many ones does our vector have?

Answer: $k_1 + \cdots + k_n = k$.

How many zeroes does our vector have?

Answer: $n-1$, because each group of k_i ones is followed by a zero, except the last one.

For each way to draw k balls out of $1,\ldots,n$ with replacement but without order, have constructed a vector in $\{0,1\}^m = \{0,1\}^{k+n-1}$ with exactly k ones. On the other hand, each such vector corresponds to precisely one outcome of the draw: count how many ones there are until the first zero. That is the number of times ball one was drawn. Remove these initial ones and the zero that follows them, now count how many ones until the second zero, that is the number of times ball two was drawn. And so on.

So we see that the number we are interested in is precisely the number of vectors in $\{0,1\}^{k+n-1}$ having exactly k ones. The previous question tells us this is $\binom{k+n-1}{k}$ as claimed. ∎

1.2 Inclusion-exclusion (counting version)

Consider a group of N men, of which $N(\text{beard})$ have a beard, $N(\text{bald})$ are bald, and $N(\text{beard, bald})$ are both bald and bearded. A certain (choosy) lady is only interested in men that are neither bald nor bearded. How many men fit this description? Answer:

$$N - N(\text{beard}) - N(\text{bald}) + N(\text{beard, bald}).$$

(Men who have both properties are counted twice by $N(\text{beard}) + N(\text{bald})$.)

Now suppose there are N men and r properties $p_1,\ldots,p_r$ of interest. How many men have none of these properties? In Exercise 1.2 below, using induction, you will prove that the answer is

$$N - \sum_{j=1}^{r} N(p_j) + \sum_{1\le j_1<j_2\le r} N(p_{j_1},p_{j_2}) - \sum_{1\le j_1<j_2<j_3\le r} N(p_{j_1},p_{j_2},p_{j_3}) + \cdots + (-1)^r N(p_1,\ldots,p_r). \tag{1.1}$$

A real-life situation in which this formula is relevant, at least for those of you who are Dutch, is "Sinterklaas". When many dutch families celebrate the sinterklaas holiday, a (random) assignment is made where each family member buys a present for exactly one other family member. So we are in the setting of permutations, but of course we do not want any family members having to buy a present for themselves. What we get are *derangements*, sequences fo the numbers $1,\ldots,n$ with the property that the i number is not in position i, for every i.

Question 1.7 What is D_n, the number of derangements of $1,\ldots,n$?

Answer: Let the property p_i denote that the number i is in the i-th place. The inclusion-exclusion formula says that

$$D_n = n! - \sum_{i} N(p_i) + \sum_{1\le i_1<i_2\le n} N(p_{i_1},p_{i_2}) + \cdots + (-1)^n N(p_1,\ldots,p_n).$$

Now note that

$$N(p_{i_1},\ldots,p_{i_r}) = (n-r)!,$$

since we just fix $i_1,\ldots,i_r$ and arbitrarily rearrange the other numbers. Also the number of choices

of $1 \leq i_1 < \cdots < i_r \leq n$ is precisely $\binom{n}{r}$ – we just choose a subset of size r. Hence the expression becomes

$$\begin{aligned} D_n &= n! - \binom{n}{1}(n-1)! + \binom{n}{2}(n-2)! + \cdots + (-1)^n \binom{n}{n}(n-n)! \\ &= \sum_{r=0}^{n} (-1)^r \binom{n}{r}(n-r)! \\ &= \sum_{r=0}^{n} (-1)^r \frac{n!}{r!} \\ &= n! \cdot \left(\sum_{r=0}^{n} (-1)^r \frac{1}{r!} \right). \end{aligned}$$

Now we note that

$$\sum_{r=0}^{n} (-1)^r \frac{1}{r!} \xrightarrow[n\to\infty]{} \sum_{r=0}^{\infty} (-1)^r \frac{1}{r!} = e^{-1}.$$

(The last equality you can for instance see from the Taylor expansion for e^x.)

So

$$D_n \approx \frac{n!}{e}.$$

That is, when n is large then roughly a $1/e$-fraction of all permutations are also derangements. ■

1.3 Exercises

1.3.1 Check-up the basics

Exercise 1.1 — Decide whether or not each statement is necessarily true.

1. For sets A, B, C, we have $(A \cup B \cup C)^c = A^c \cup B^c \cup C^c$.
2. $|A \cup B| = |A| + |B|$.
3. If $A, B \subset \{1, \ldots, n\}$ satisfy $|A|, |B| > \frac{n}{2}$ then $A \cap B \neq \emptyset$.

■

1.3.2 Statement from the lecture

Exercise 1.2 — Inclusion-exclusion formula. The (counting version of the) inclusion-exclusion formula can be formulated as follow: Let X be a finite set and $X_1, \ldots, X_r \subset X$ be subsets of X. It holds that

$$\begin{aligned} \#(X \setminus \cup_{i=1}^{r} X_i) = \#X - \sum_{j=1}^{r} \#X_i + \sum_{1 \leq j_1 < j_2 \leq r} \#(X_{j_1} \cap X_{j_2}) \\ - \sum_{1 \leq j_1 < j_2 < j_3 \leq r} \#(X_{j_1} \cap X_{j_2} \cap X_{j_3}) + \ldots + (-1)^r \#(X_1 \cap \cdots \cap X_r). \end{aligned}$$

Here $\#Y$ denotes the *cardinal* of a set Y, that is how many elements Y has.

1. Explain why the right hand side is the same as Equation (1.1).
2. If this formula looks very abstract to you, check that the property holds on simple examples with $r = 1$, $r = 2$ and $r = 3$.
3. Prove the formula by induction.

■

1.3.3 Problems

Exercise 1.3 How many ways can you place a king, a horse and two rooks (all of the same colour) on a chess board? ■

Exercise 1.4 a) In how many ways can we place k balls into n urns, if the balls and urns are distinguishable?
b) Same question, but now balls are indistinguishable. ■

Exercise 1.5 How many different "words" can we form with (all) the letters: M I S S I S S I P P I? ■

Exercise 1.6
a) Suppose that a set S has n elements. Prove that the number of subsets of S is equal to 2^n.
b) Combining this with the information from the lecture, prove that $2^n = \sum_{m=0}^{n} \binom{n}{m}$. ■

Exercise 1.7 — The justification for Pascal's triangle. Show that, for all $n \in \mathbb{N}$ and $1 \leq k \leq n$ we have

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

Hint: *We count the number of ways to choose a k-element subset from the set* $\{1,\ldots,n+1\}$. *Either you include the last element* $n+1$ *or you do not.* ■

Exercise 1.8 Show that $\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2$.
Hint: *We count the number of ways to choose a n-element subset from the set* $\{1,\ldots,2n\}$. *If k elements have been chosen among* $1,\ldots,n$, *then how many elements must be chosen among* $n+1,\ldots,2n$ *if we are to choose n elements in total?* ■

Exercise 1.9 Consider n similar wagons and r similar locomotives. How many distinct ways can you make r trains if all the wagons are to be used, and
(i) The wagons and locomotives all have distinct numbers?
(ii) The wagons are anonymous (no number, same colour, etc.) and the locomotives are numbered?
(iii) The wagons and locomotives all have distinct numbers and each train must contain at least k wagons?
(iv) The wagons are anonymous and the locomotives are numbered and each train must contain at least k wagons? ■

1.3.4 Are you up for a challenge?

Exercise 1.10 n married couples are seated on a circular table so that men and women alternate. How many seatings are there in which no couple is sitting next to each other? ■

2. Basics of set theory

The goals of this chapters are:

- ▷ to learn some elementary **properties of set operations** (inclusion, exclusion...);
- ▷ to define the notion of **probability space** and **probability function**;
- ▷ to derive the most important basic **properties of a probability functions**.

2.1 Set theory

We start with a quick review of some set theory, that has already been covered in other courses. In probability theory, we are always studying some (possibly completely fictional) "experiment", with various possible "outcomes". The experiment could for instance be rolling a dice and the possible outcomes would then be $1, 2, \ldots, 6$.

Definition 2.1.1 A **sample space** is a set Ω. Any element $\omega \in \Omega$ is called an **outcome**. Any subset $A \subseteq \Omega$ is called an **event**.

R Meaning of set inclusion and equality:

$$A \subseteq B \text{ means: } \quad \omega \in A \Longrightarrow \omega \in B,$$
$$A = B \text{ means: } \quad A \subseteq B \text{ and } B \subseteq A.$$

A sample space Ω could be either countable (finite or countably infinite) or uncountable.

R Countably infinite means there exists a one-to-one function $f : \mathbb{N} \to S$. I.e. we can "count" S. The sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are all countable, but $\mathbb{R}$ is not. If you are not familiar with it you may wish to look up Cantor's beautiful "diagonalization argument" showing that the reals are uncountable.

Set operations:

▷ Union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$

▷ Intersection $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$

▷ Complement $A^c = \{\omega \in \Omega : \omega \notin A\}$

▷ Difference $A \backslash B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$

Note that here, as is usual in mathematics, "or" is taken to be the "inclusive or". That is "p or q" includes the possibility that p and q are simultaneously true.

Theorem 2.1.1 — Algebra of sets. For any $A, B, C \subseteq \Omega$ and $\omega \in \Omega$,

1. if $\omega \in A$ and $A \subseteq B$, then $\omega \in B$,
2. $A, B \subseteq A \cup B$ and $A \cap B \subseteq A, B$,
3. $A \cup B = B \cup A$ and $A \cap B = B \cap A$, (commutativity)
4. $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$, (associativity)
5. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, (distributivity)
6. $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$, (De Morgan's Laws)
7. $A \backslash B = A \cap B^c$.

Proof. We will only prove $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ to illustrate how these proofs go. The remaining proofs are left as an exercise. We need to show:

$$A \cup (B \cap C) \subset (A \cup B) \cap (A \cup C) \qquad (\clubsuit)$$

$$A \cup (B \cap C) \supset (A \cup B) \cap (A \cup C) \qquad (\spadesuit)$$

Proof of ($\clubsuit$): Let $\omega \in A \cup (B \cap C) = \{\omega' \in \Omega : \omega' \in A \text{ or } \omega' \in B \cap C\}$. There are two cases:

▷ $\omega \in A \implies \omega \in A \cup B$ and $\omega \in A \cup C \implies \omega \in (A \cup B) \cap (A \cup C)$

▷ $\omega \in B \cap C \implies \omega \in B$ and $y \in C \implies \omega \in A \cup B$ and $\omega \in A \cup C \implies \omega \in (A \cup B) \cap (A \cup C)$

Proof of ($\spadesuit$): Let $\omega \in (A \cup B) \cap (A \cup C)$. We have $\omega \in A \cup B$ and $\omega \in A \cup C$. Also, either $\omega \in A$ or $\omega \notin A$. In case $\omega \in A$, we of course have $\omega \in A \cup (B \cap C)$. If, however, $\omega \notin A$, then from $\omega \in A \cup B$ we get $\omega \in B$ and from $\omega \in A \cup C$ we get $\omega \in C$, so $\omega \in B \cap C$, so $\omega \in A \cup (B \cap C)$. ■

Definition 2.1.2

▷ A, B are **disjoint** if $A \cap B = \varnothing$.

▷ $A_1, A_2, \ldots$ are **pairwise disjoint** if $A_i \cap A_j = \varnothing$ for all $i \neq j$.

▷ $A_1, A_2, \ldots$ form a **partition** of B if they are pairwise disjoint and their union is B.

R The union and intersection of more than two sets.

$$A_1, A_2, \ldots, A_n \qquad \bigcup_{i \leqslant n} A_i = \{\omega : \omega \in A_i \text{ for some } i\} \qquad \bigcap_{i \leqslant n} A_i = \{\omega : \omega \in A_i \text{ for all } i\}$$

Similarly:

$$\bigcup_{i \geqslant 1} A_i, \quad \bigcap_{i \geqslant 1} A_i, \quad \bigcup_{\lambda \in L} A_\lambda, \quad \bigcap_{\lambda \in L} A_\lambda$$

(L can be an infinite, and possibly uncountable, set).

R De Morgan's Laws also hold:

$$\left(\cup A_i\right)^c = \cap A_i^c, \quad \left(\cap A_i\right)^c = \cup A_i^c.$$

2.2 The axioms of probability

Definition 2.2.1 A **probability space** is a triple $(\Omega, \mathscr{A}, \mathbb{P})$ where
- ▷ Ω is a *sample space* (i.e. a set, as above),
- ▷ $\mathscr{A}$ is a *collection of events* (i.e. a set of subsets of Ω):

$$\mathscr{A} = \begin{cases} \text{set of all subsets of } \Omega, & \text{if } \Omega \text{ is countable,} \\ \text{a certain set of subsets of } \Omega, & \text{if } \Omega \text{ is uncountable,} \end{cases}$$

- ▷ and $\mathbb{P}$ is **probability function** on Ω: that is a function $\mathbb{P}\colon \mathscr{A} \to [0,1]$ satisfying
  (A1) $\mathbb{P}(\Omega) = 1$,
  (A2) if $A_1, A_2, \ldots$ are pairwise disjoint, then $\mathbb{P}\left(\bigcup_{i\geqslant 1} A_i\right) = \sum_{i\geqslant 1} \mathbb{P}(A_i)$.

Technical comment: We keep the expression "a certain set" deliberately vague – because the intricacies of doing probability theory on uncountable sample spaces are too advanced for the scope of this course. In fact, the mystery will not be fully lifted until the third year of the maths degree if/when you take the course "probability and measure". If you cannot wait until then you could look up *sigma-algebras*.

The set of all subsets of Ω is also called the *power set* of Ω and denoted by $\mathscr{P}(\Omega)$ or 2^Ω.

The conditions (A1), (A2) are called (Kolmogorov's) *axioms of probability*.

Axiom (A2) is sometimes called **sigma additivity**.

■ **Example 2.1 — Probability space of a single fair coin tossing.**

$$\begin{aligned} &\Omega = \{H, T\}, \\ &\mathscr{A} = \{\varnothing, \{H\}, \{T\}, \{H,T\}\}, \\ &\mathbb{P}(\varnothing) = 0,\ \mathbb{P}(\{H\}) = \frac{1}{2},\ \mathbb{P}(\{T\}) = \frac{1}{2},\ \mathbb{P}(\{H,T\}) = 1 \end{aligned}$$

gives the probability space corresponding to tossing a fair coin. ■

This example shows one way of defining a probability $\mathbb{P}$: we simply specified the probability of *all possible events*. This is not too practical: if Ω has n elements, then $\mathscr{A}$ has 2^n elements (check this as an exercise!) and we would have to specify the probabilities of each of them. Fortunately, there is an easier way to define a probability (at least for countable Ω).

Theorem 2.2.1 Suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$ and $p_1, p_2, \ldots$ are non-negative numbers with $\sum p_i = 1$. Defining, for all $A \subseteq \Omega$,

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i,$$

the function $\mathbb{P}$ thus obtained is a probability.

(The proof is quite easy, we leave it as an exercise.)

■ **Example 2.2 — Experiment: roulette.**

$$\left.\begin{aligned} &\Omega = \{\text{green, blue, red}\} \\ &\mathbb{P}(\{\text{green}\}) = \tfrac{1}{37} \\ &\mathbb{P}(\{\text{blue}\}) = \tfrac{18}{37} \\ &\mathbb{P}(\{\text{red}\}) = \tfrac{18}{37} \end{aligned}\right\} \text{writing this is enough to define } \mathbb{P}$$

(These numbers are for the French version of roulette where there are numbers $0, 1, \ldots, 36$ with 0 green and the other numbers divided evenly between red and black. In the American version there is also a "number" 00 that is green so the denominators would have to be changed to 38.) ∎

Theorem 2.2.2 — First properties of probabilities.

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

b) if $B_1, \ldots, B_n$ are pairwise disjoint, then $\mathbb{P}\left(\bigcup_{i\leqslant n} B_i\right) = \sum_{i\leqslant n}\mathbb{P}(B_i)$;

c) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any A;

d) $\mathbb{P}(A) \leq 1$ for any A;

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

f) if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof. We will only prove a few of the statements. The rest are left as an exercise.

Proof of **a**): Let $A_1 = \varnothing, A_2 = \varnothing, A_3 = \varnothing, \ldots$. These sets are pairwise disjoint, so

$$\mathbb{P}(\varnothing) = \mathbb{P}\left(\bigcup_{i\geqslant 1} A_i\right) \overset{(A2)}{=} \sum_{i\geqslant 1}\mathbb{P}(A_i) = \sum_{i\geqslant 1}\mathbb{P}(\varnothing) \quad \Longrightarrow \quad \mathbb{P}(\varnothing) = 0.$$

(Since zero is the only one real number such that if you add it to itself infinitely many times and not get $+\infty$ or $-\infty$.)

Proof of **b**): Assume $B_1, \ldots, B_n$ are pairwise disjoint. Let $A_1 = B_1, \ldots, A_n = B_n$ and $A_{n+1} = A_{n+2} = \cdots = \varnothing$. Then, the A_i's are pairwise disjoint and

$$\mathbb{P}\left(\bigcup_{i\leqslant n} B_i\right) = \mathbb{P}\left(\bigcup_{i\geqslant 1} A_i\right) \overset{(A2)}{=} \sum_{i\geqslant 1}\mathbb{P}(A_i) = \sum_{i\leqslant n}\mathbb{P}(B_i) + \sum_{i=n+1}^{\infty}\mathbb{P}(\varnothing) = \sum_{i\leqslant n}\mathbb{P}(B_i)$$

Proof of **e**): Note: that $A\backslash B$, $B\backslash A$ and $A\cap B$ are pairwise disjoint, so we can apply b) to see:

$$\begin{aligned}&\mathbb{P}(A\cup B) = \mathbb{P}((A\backslash B)\cup(B\backslash A)\cup(A\cap B)) = \mathbb{P}(A\backslash B) + \mathbb{P}(B\backslash A) + \mathbb{P}(A\cap B)\\ &\Longrightarrow \mathbb{P}(A\cup B) + \mathbb{P}(A\cap B) = \underbrace{\mathbb{P}(A\backslash B) + \mathbb{P}(A\cap B)}_{\mathbb{P}(A)} + \underbrace{\mathbb{P}(B\backslash A) + \mathbb{P}(A\cap B)}_{\mathbb{P}(B)}.\end{aligned}$$

■

Theorem 2.2.3 — Sigma sub-additivity. For any $A_1, A_2, \ldots$,

$$\mathbb{P}\left(\bigcup_{i\geqslant 1} A_i\right) \leq \sum_{i\geqslant 1}\mathbb{P}(A_i)$$

(this also works for finitely many sets: $\mathbb{P}(\cup_{i\leqslant n} A_i) \leq \sum_{i\leqslant n}\mathbb{P}(A_i)$).

Proof. Let $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus (A_1 \cup A_2)$, …, $B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1})$.

Claim 1. $\bigcup_{i \geqslant 1} B_i = \bigcup_{i \geqslant 1} A_i$ (easy to check).

Claim 2. The B_i's are pairwise disjoint. Indeed, if $i < j$, then

$$B_j = A_j \setminus (A_1 \cup \cdots \cup A_{j-1}) = A_j \cap (A_1 \cup \cdots \cup A_{j-1})^c = A_j \cap A_1^c \cap A_2^c \cap \cdots \cap A_{j-1}^c \subseteq A_i^c,$$

so

$$B_i \cap B_j \subseteq A_i \cap A_i^c = \varnothing,$$

so $B_i \cap B_j = \varnothing$.

Conclusion. We now have:

$$\mathbb{P}\Big(\bigcup_{i \geqslant 1} A_i\Big) = \mathbb{P}\Big(\bigcup_{i \geqslant 1} B_i\Big) = \sum_{i \geqslant 1} \mathbb{P}(B_i) \leq \sum_{i \geqslant 1} \mathbb{P}(A_i),$$

where the first equality follows from Claim 1, the second equality follows from Claim 2 and (A2), and the inequality follows from the fact that $B_i \subseteq A_i$. ■

Theorem 2.2.4 — Uniform probability on a finite sample space. If $(\Omega, \mathscr{A}, \mathbb{P})$ is a probability space, Ω is finite and the outcomes $\omega \in \Omega$ all have the same probability, then, for any $A \in \mathscr{A}$,

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{\text{number of elements of } A}{\text{number of elements of } \Omega}.$$

Proof. Exercise! ■

2.3 Exercises

2.3.1 Check-up the basics

Exercise 2.1 — Decide whether or not each statement is necessarily true.

1. If $A_1, A_2, \ldots$ form a partition of S and $B \subset S$, then $A_1 \cap B$, $A_2 \cap B$, … form a partition of B.
2. For any sets A, B, C one has $A \setminus (B \setminus C) = (A \setminus B) \cup C$.

■

2.3.2 Statement from the lecture

Exercise 2.2 Prove the statements of Theorem 2.1.1 which were not proven during the lecture. ■

Exercise 2.3 Prove the statements of Theorem 2.2.2 which were not proven during the lecture. ■

Exercise 2.4 Prove Theorem 2.2.4. ■

2.3.3 Problems

Exercise 2.5 A fair six-faced die is rolled three times.

1. State a probability space that describes this random experiment.
2. Formally define the events
   (a) A = "the sum of all rolled dots is 11";
   (b) B = "the sum of all rolled dots is 12"
   and calculate their probability.

■

Exercise 2.6 Let $A_1, A_2, \ldots$ be events in some probability space. Prove that

1. for each $n \in \mathbb{N}$
$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - n + \sum_{i=1}^{n} \mathbb{P}(A_i).$$
2. (a) If $A_n \subseteq A_{n+1}$ for all $n \geq 1$ and $\cup_{n=1}^{\infty} A_n = A$, then,
$$\mathbb{P}(A_n) \xrightarrow{n\to\infty} \mathbb{P}(A);$$
(b) If $A_n \subseteq A_{n-1}$ for all $n \geq 2$ and $\cap_{n=1}^{\infty} A_n = A$, then,
$$\mathbb{P}(A_n) \xrightarrow{n\to\infty} \mathbb{P}(A);$$

■

Exercise 2.7 In a game of poker you get a random hand consisting of 5 (out of 4×13) cards.

1. State a suitable probability space to describe this random experiment.
2. Calculate the probability that you have *three of a kind* but not *four of a kind* or a *full house* (*three of a kind* + *pair*)
3. Calculate the probability that you have a *full house*.

■

3. Probabilities and events' relations

The goals of this chapters are:

- ▷ to define **conditional probability**;
- ▷ to derive the **law of total probability**;
- ▷ to derive **Bayes' formula**;
- ▷ to define the property of **pairwise independence** and **mutual independence** of events.

3.1 Conditional probability

The following definition formalizes the intuitive notion of "the probability that A occurs given that we know that B has occurred".

Definition 3.1.1 — Conditional probability. Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space and A, B be events; assume $\mathbb{P}(B) > 0$. Then, the **probability of A given B** is defined by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Note that $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A)$. The convention is $\mathbb{P}(B \mid A) = 0$ if $\mathbb{P}(A) = 0$.

Question 3.1 We roll three dice. Let

$$A = \{\text{We obtain 6 three times}\}, \quad B = \{\text{The sum of the values obtained is 17 or more}\}.$$

Find $\mathbb{P}(A)$ and $\mathbb{P}(A \mid B)$.

Answer:

$$\Omega = \{(i, j, k) : i, j, k \in \{1, 2, 3, 4, 5, 6\}\}.$$

Figure 3.1: Thomas Bayes

All outcomes have the same probability, so

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{1}{216}.$$

Now let us find $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. We have

$$B = \{(5,6,6),(6,5,6),(6,6,5),(6,6,6)\}, \quad A \cap B = \{(6,6,6)\},$$

so

$$\mathbb{P}(B) = \frac{4}{216}, \qquad \mathbb{P}(A \cap B) = \frac{1}{216}, \qquad \Longrightarrow \mathbb{P}(A \mid B) = \frac{1}{4}.$$

■

Theorem 3.1.1 Assume B is an event with $\mathbb{P}(B) > 0$. Assume $A_1, A_2, \ldots$ are pairwise disjoint and $\cup A_i = \Omega$. Then,

$$\mathbb{P}(B) = \sum_{i \geq 1} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i), \qquad \textbf{(Law of total probability)}$$

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)}{\sum_{j \geq 1} \mathbb{P}(B \mid A_j) \cdot \mathbb{P}(A_j)}. \qquad \textbf{(Bayes' formula)}$$

Intermezzo. Thomas Bayes (1702-1761) was a Presbyterian minister who was interested in theology, philosophy and statistics. He wrote an essay in which he proved what is now called Bayes' theorem, but it was published only after his death.

Proof. 1) Since $B = \bigcup_i (B \cap A_i)$ and $B \cap A_1$, $B \cap A_2$, ... are pairwise disjoint, we have

$$\mathbb{P}(B) = \sum \mathbb{P}(B \cap A_i) = \sum \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(A_i)} \cdot \mathbb{P}(A_i) = \sum \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i).$$

2)

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)}{\sum_{j \geq 1} \mathbb{P}(B \mid A_j) \cdot \mathbb{P}(A_j)}.$$

■

Question 3.2 About 1 percent of all women aged 40-50 have breast cancer. If a woman has breast cancer a mammogram will give a "positive" result (i.e. the test signals that there is cancer) in about 90 percent of the cases. (So the probability of a "false negative" is 10 percent.) If a woman does not have breast cancer there is nevertheless a 10 percent chance of a "false positive" result if a mammogram is done.

Consider the "experiment" where a woman between 40 and 50 has a mammogram done. Let

$$A := \{\text{breast cancer}\},$$
$$B := \{\text{positive mammogram}\}.$$

Putting ourselves in the shoes of a woman in this age group that has received the distressing news of a positive mammogram, we wish to find out the conditional probability $\mathbb{P}(A \mid B)$.

Answer: By Bayes' formula (with $A_1 := A$, $A_2 := A^c$):

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}.$$

The above data translates to

$$\mathbb{P}(A) = .01, \quad \mathbb{P}(A^c) = .99, \quad \mathbb{P}(B|A) = .9, \quad \mathbb{P}(B|A^c) = .1.$$

Filling in these numbers gives

$$\mathbb{P}(A|B) = \frac{.9 \times .01}{.9 \times .01 + .1 \times .99} = \frac{9}{108},$$

which is roughly 8 percent.

The answer may be somewhat surprising. Indeed, when 95 doctors were asked "what is the chance a woman has breast cancer given that she had a positive mammogram?", the average answer was *75 percent*!!! ■

3.2 Independence

We roll twice with a dice. Let $A = \{\text{the first roll gives six}\}$, $B = \{\text{the second roll gives six}\}$. As should be intuitively obvious, you can compute

$$\mathbb{P}(B|A) = \frac{1}{6} = \mathbb{P}(B).$$

So the knowledge that A has occurred does not make any difference to the likelihood that B occurs. This motivates the following definition.

Definition 3.2.1 — Independent. Events A, B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

In case $\mathbb{P}(A) > 0$, the definition of independence is the same as:

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

(The reason we did not use this as the definition of independence is that $\mathbb{P}(B|A)$ is undefined if $\mathbb{P}(A) = 0$.)

As a general rule, if we consider a chance experiment (e.g. "roll a dice twice") and the events A, B are determined by different "subexperiments" that do not influence each other (e.g. the roll of the first dice vs. the roll of the second dice) then A, B are independent.

Common mistake: thinking that "A and B disjoint" is the same as "A and B independent". These are very different notions! Make sure you understand the difference.

Question 3.3 A coin is tossed and a die is rolled. Let

$$A = \{\text{Coin gives Ⓗ}\}, \qquad B = \{\text{Die gives ⚄ or ⚅}\}.$$

Prove that A and B are independent.

Answer: Then, $\#\Omega = 2 \cdot 6 = 12$, all outcomes are equally likely, $\#A = 6$, $\#B = 4$ and $\#A \cap B = 2$. Hence,

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{3}, \quad \mathbb{P}(A \cap B) = \frac{1}{6} \quad \Longrightarrow A,\ B \text{ are independent.}$$

■

Definition 3.2.2 Events $A_1, \ldots, A_n$ are:

- ▷ **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j)$ for all $i \neq j$;
- ▷ **mutually independent** if, for any subcollection $A_{i_1}, \ldots, A_{i_k}$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Question 3.4 A die is rolled 4 times. Find the probability of obtaining at least one 6.
Note: You may use that if A_1, A_2, A_3, A_4 are events about dice 1, 2, 3, 4, respectively, then the events are mutually independent.

Answer:

$$\begin{aligned}\mathbb{P}(\text{At least one 6 in 4 rolls}) &= 1 - \mathbb{P}(\text{No 6 in 4 rolls}) \\ &= 1 - \mathbb{P}\left(\bigcap_{i=1}^{4} \{\text{Roll } i \text{ does not give 6}\}\right) \\ &= 1 - \left(\frac{5}{6}\right)^4 = 0.518.\end{aligned}$$

■

Important: If $A_1, \ldots, A_n$ are mutually independent then they are also pairwise independent. However, in general the converse does not hold, as the following example shows.

Question 3.5 A fair coin is tossed three times. Let

$$\begin{aligned}A_1 &:= \{\text{coin 1 has the same outcome as coin 2}\}, \\ A_2 &:= \{\text{coin 1 has the same outcome as coin 3}\}, \\ A_3 &:= \{\text{coin 2 has the same outcome as coin 3}\}.\end{aligned}$$

Are the events A_1, A_2 and A_3 independent?

Answer: We have $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{2}$. We can easily compute (do this as an exercise!)

that

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

Thus

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2), \quad \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_3), \quad \mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2)\mathbb{P}(A_3),$$

and therefore the events are *pairwise independent*. On the other hand

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 1/4 \neq \frac{1}{8} = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3)$$

(An easy way to see this is to notice that in fact $A_1 \cap A_2 = A_1 \cap A_2 \cap A_3$.), and thus the events are NOT *mutually independent*. ■

3.3 Exercises

3.3.1 Check-up the basics

Exercise 3.1 — Decide whether or not each statement is necessarily true.

1. If A and B are mutually exclusive events and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = 0$.
2. I have two dice. Die 1 is a normal die, that is, it has faces ⚀⚁⚂⚃⚄⚅ and each is obtained with probability $\frac{1}{6}$. Die 2 is unusual: it has faces ⚀⚀⚂⚂⚄⚄, and again each is obtained with probability $\frac{1}{6}$. I pick a die at random (each having probability $\frac{1}{2}$ of being picked) and roll it. Given that I obtained a 3, the probability of having chosen die 2 is $\frac{1}{2}$.

■

3.3.2 Problems

Exercise 3.2 Prove each of the following statements. (Assume that any conditioning event has positive probability).

(a) If $\mathbb{P}(B) = 1$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$ for any A.
(b) If $A \subset B$, then $\mathbb{P}(B|A) = 1$ and $\mathbb{P}(A|B) = \mathbb{P}(A)/\mathbb{P}(B)$.
(c) If A and B are mutually exclusive, then

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

(d) $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B|C)\mathbb{P}(C)$.

■

Exercise 3.3 A pair of events A and B cannot be simultaneously *mutually exclusive* and *independent*. Prove that if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then:

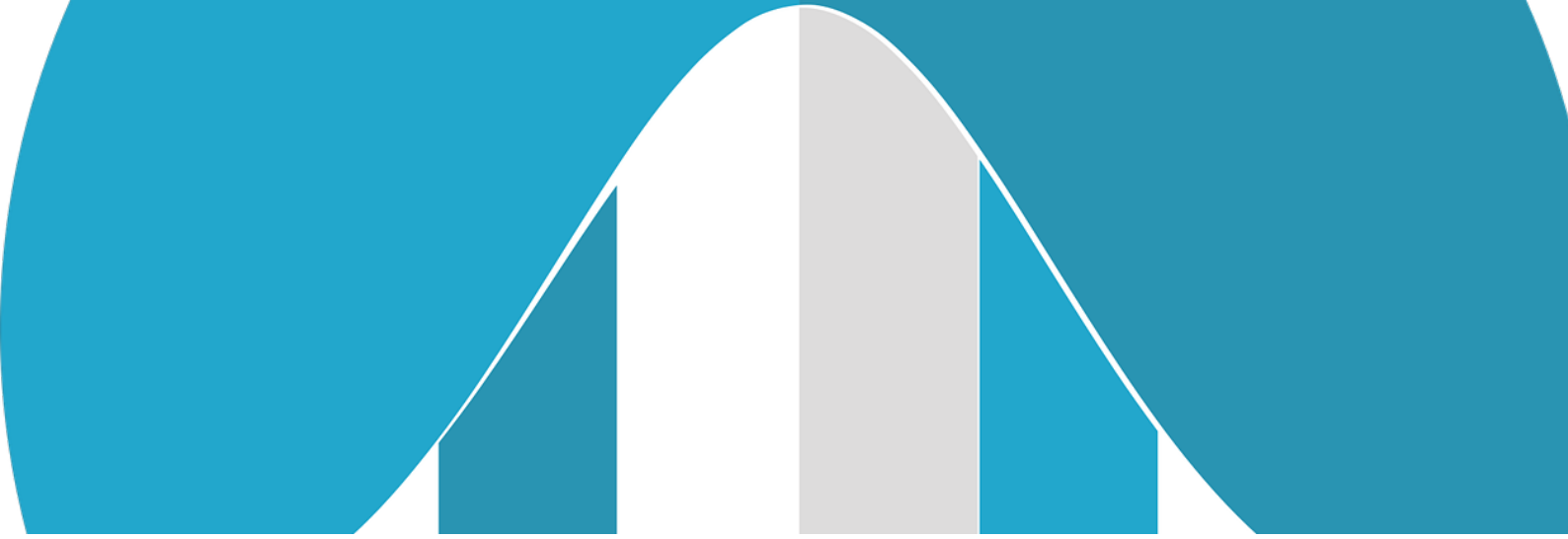
(a) If A and B are mutually exclusive, they cannot be independent.
(b) If A and B are independent, they cannot be mutually exclusive.

■

Exercise 3.4 We have three cards. One has both faces black, another has both faces white, and the third has one face of each color. We take one of the three cards at random (that is, each has probability $\frac{1}{3}$ of being taken) and reveal one of its faces, which is black. What is the probability that the other face is also black? ■

3.3.3 Are you up for a challenge?

Exercise 3.5 40% of the boys in a certain village are polite and the remaining 60% are impolite. Polite boys open doors to elders 2/3 of the time, whereas impolite boys only do so half the time. Little Paul is seen opening the door to Mrs Marple, but not opening it to Mr Poirot. What is the probability that he is polite? ■



4. Random variables and their distributions

The goals of this chapters are:

- ▷ to define **random variables** and their **distributions**;
- ▷ to define the **cumulative distribution function** (cdf) of a random variable;
- ▷ to define **discrete random variables** and their **probability mass functions** (pmf);
- ▷ to define **absolutely continuous random variables** and their **probability density functions** (pdf);
- ▷ to derive properties of cdf's, pmf's and pdf's.

4.1 Random variables

Suppose we roll a die 10^{10} times and want to look at the number of times we obtain a 6.

$$\Omega = \{(n_1, n_2, \ldots, n_{10^{10}}) : n_i \in \{1,2,3,4,5,6\} \text{ for each } i\}.$$

We find $\#\Omega = 6^{10^{10}}$ which is huge. However, much of the information contained in an outcome will be **useless** to us, since we just want the number of 6's.
Say we define a function $X : \Omega \to \mathbb{R}$ by

$$X(n_1, n_2, \ldots, n_{10^{10}}) = \text{number of 6's in the sequence.}$$

The value of X can be between 0 and 10^{10} (**much** smaller than $6^{10^{10}}$). So life will be easier if we can focus our attention on X rather than on Ω.

Definition 4.1.1 Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space (see Definition 2.2.1). A **random variable** is a function from Ω to $\mathbb{R}$ (usually denoted by X, Y or Z).

Technical comment: To be 100% rigorous, in the case where Ω is uncountable, one should add some additional conditions for a function $X : \Omega \to \mathbb{R}$ in order to be called a random variable, but this goes beyond the scope of this course. In fact, the mystery will not be fully lifted until the third year of the maths degree if/when you take the course "probability and measure". If you cannot wait until then you could look up *measurable functions*.

Notation: The set $\{\omega \in \Omega \text{ such that } X(\omega) = x\}$ is an event (since it is a subset of Ω). It is equal to $X^{-1}(x)$ (the inverse image of x), and we usually use the notation $\{X = x\}$ for that event. Instead of $\mathbb{P}(\{\omega \in \Omega \text{ such that } X(\omega) = x\})$ we write: $\mathbb{P}(X = x)$. Similarly

$$\begin{aligned}
\mathbb{P}(X \in A) &:= \mathbb{P}(\{\omega \in \Omega \text{ such that } X(\omega) \in A\}), && A \subset \mathbb{R},\\
\mathbb{P}(X \leq x) &:= \mathbb{P}(\{\omega \in \Omega \text{ such that } X(\omega) \leq x\}), && x \in \mathbb{R},\\
\mathbb{P}(a < X \leq b) &:= \mathbb{P}(\{\omega \in \Omega \text{ such that } a < X(\omega) \leq b\}), && a, b \in \mathbb{R},\\
&\vdots
\end{aligned}$$

Definition 4.1.2 The **distribution** of a random variable X is the function that maps

$$\text{each } A \subset \mathbb{R} \qquad \text{to} \qquad \mathbb{P}(X \in A) = \mathbb{P}(\{s \in S : X(s) \in A\}).$$

Technical comment: Again, we are ignoring some technicalities which are beyond the scope of the lecture. In the definition of the distribution of a random variable, one does not need to consider all subsets $A \subset \mathbb{R}$ but only nice subsets called *Borel sets*... but this is an other story.
In the rest of this set of lecture notes, we will keep ignoring this technicality.

Note that if a random variable takes only a finite number of values (as in the next example), the distribution can be very simply described by considering only probability of the form $\mathbb{P}(X = i)$.

Question 4.1 A coin is tossed 3 times. Let X be the number of heads obtained. Find the distribution of X.

Answer:

outcome s	ⒽⒽⒽ	ⒽⒽⓉ	ⒽⓉⒽ	ⓉⒽⒽ	ⒽⓉⓉ	ⓉⒽⓉ	ⓉⓉⒽ	ⓉⓉⓉ
$X(\omega) =$	3	2	2	2	1	1	1	0

Then,

$$\mathbb{P}(X=0) = \frac{1}{8},\ \mathbb{P}(X=1) = \frac{3}{8},\ \mathbb{P}(X=2) = \frac{3}{8},\ \mathbb{P}(X=3) = \frac{1}{8},\ \underbrace{\mathbb{P}(X=x) = 0 \text{ for any other } x}_{\substack{\text{we usually omit this part.}\\ \text{It is implicit, since the others add up to 1.}}}.$$

From that description, one can derive the probability of any event. For instance,

$$\mathbb{P}(X \leq 2) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}.$$

■

4.2 Distribution functions

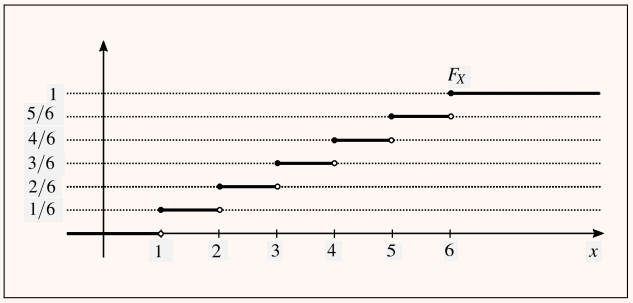
Definition 4.2.1 — Cumulative distribution function. The **cumulative distribution function** (abbreviated **cdf**; sometimes simply called distribution function) of a random variable X is the function $F_X : \mathbb{R} \to [0,1]$ given by:

$$F_X(x) = \mathbb{P}(X \leqslant x).$$

Question 4.2 Let X be the result of rolling a die. Compute its cumulative distribution function and draw its graph.

Answer:

$$F_X(x) = \begin{cases} 0, & \text{if x < 1;} \\ 1/6, & \text{if } 1 \le x < 2; \\ 2/6, & \text{if } 2 \le x < 3; \\ 3/6, & \text{if } 3 \le x < 4; \\ 4/6, & \text{if } 4 \le x < 5; \\ 5/6, & \text{if } 5 \le x < 6; \\ 1, & \text{if } x \ge 6. \end{cases}$$



■

Theorem 4.2.1 The distribution function F_X of a random variable X satisfies:

(a) $\lim_{x\to-\infty} F_X(x) = 0, \quad \lim_{x\to\infty} F_X(x) = 1.$

(b) F_X is nondecreasing.

(c) F_X is right continuous, that is, for any x, $\lim_{y\searrow x} F_X(y) = F_X(x)$.

Proof. Exercise. ■

Note: $F_X(x) = \mathbb{P}(X \le x)$ implies:

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leqslant x) = 1 - F_X(x),$$

$$\mathbb{P}(a < X \le b) = \mathbb{P}(X \le b) - \mathbb{P}(X \le a) = F_X(b) - F_X(a).$$

Moreover,

Lemma 4.2.2 $\mathbb{P}(X = x) = F_X(x) - \lim_{y\nearrow x} F_X(y)$. In particular, F_X is continuous at x if and only if $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$.

Proof. Observe first that $\mathbb{P}(X = x) = F_X(x) - \mathbb{P}(X < x)$. Thus we have to show that $\mathbb{P}(X < x) \overset{!}{=} \lim_{y\nearrow x} F_X(y)$.

Note that saying that $\lim_{y\nearrow x} F_X(y)$ exists is the same as saying that $\lim_{n\to\infty} F_{y_n}$ exists for any increasing sequence y_n approaching x from below, and that this limit is independent from the choice of the sequence itself. Here this is insured by the fact that F_X is non-decreasing and bounded. In particular, we can pick a arbitrary sequence $y_n \nearrow x$, for example $y_n = x - \frac{1}{n}$; and we have to show that $\mathbb{P}(X < x) \overset{!}{=} \lim_{n\to\infty} F_X(x - \frac{1}{n})$.

Consider the events $E_1 = \{X \le x - 1\}$, and $E_i = \{x - \frac{1}{i-1} < X \le x - \frac{1}{i}\}$, $i \ge 2$, and observe that

$$E_i \cap E_j = \emptyset, \quad \text{for any } 1 \le i < j \text{ (pairwise disjoint)},$$
$$\cup_{i=1}^{n} E_i = \left\{X \in (-\infty, x-1] \cup (x-1, x-\tfrac{1}{2}] \cup \cdots \cup (x - \tfrac{1}{n-1}, x - \tfrac{1}{n}]\right\} = \left\{X \le x - \tfrac{1}{n}\right\},$$
$$\cup_{i=1}^{\infty} E_i = \{X < x\}.$$

Therefore

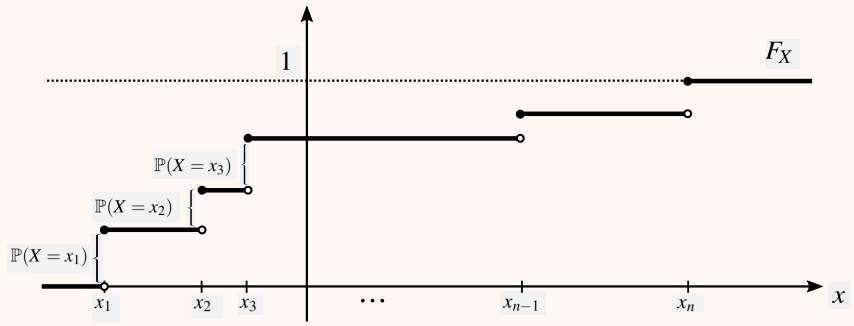
$$\begin{aligned}
\mathbb{P}(X<x) &= \mathbb{P}\left(\cup_{i=0}^{\infty} E_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(E_i) && \text{(by sigma additivity, because the sets are pairwise disjoint)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(E_i) && \text{(by def. of a series of positive terms)} \\
&= \lim_{n\to\infty} \mathbb{P}\left(\cup_{i=1}^{n} E_n\right) && \text{(sigma additivity again)} \\
&= \lim_{n\to\infty} \mathbb{P}\left(X \le x - \frac{1}{n}\right) && \text{(see above)} \\
&= \lim_{n\to\infty} F_X\left(x - \frac{1}{n}\right),
\end{aligned}$$

which proves the claim. ■

Definition 4.2.2 A random variable is **discrete** if it takes finitely many values, or more generally values in a discrete subset of $\mathbb{R}$.

Question 4.3 Suppose X is discrete. Sketch its cdf.

Answer:



■

Definition 4.2.3 A random variable is **continuous** if its cdf is continuous on $\mathbb{R}$.

By the above lemma, we immediately get:

F_x continuous on $\mathbb{R}$ if and only if $\mathbb{P}(X=x)=0$ for every $x\in\mathbb{R}$.

For this reason, if X is continuous, the following are equal:

$$\mathbb{P}(a<X<b), \quad \mathbb{P}(a\le X<b), \quad \mathbb{P}(a<X\le B), \quad \mathbb{P}(a\le X\le b).$$

R Discrete and continuous random variables are two extreme cases. If X is discrete, F_X only increases through jumps (as in the examples above). If X is continuous, F_X has **no** jumps. Naturally, there exist random variables that are neither continuous nor discrete (just think of a cdf which increases continuously at some points and also has some jumps).

Definition 4.2.4 Two random variables are **identically distributed** if they have the same distribution.

■ **Example 4.1** Toss 10 coins. Let $X = \#(\text{H})$ obtained and $Y = \#(\text{T})$ obtained. Then X and Y are identically distributed. ■

Theorem 4.2.3 *X and Y are identically distributed if and only if $F_X = F_Y$.*

Proof. If X and Y are identically distributed, then

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x) = F_Y(x) \quad \text{for all } x.$$

The other direction depends on Measure Theory and is outside the scope of this course. ■

4.3 Probability mass function (pmf) and probability density function (pdf)

Definition 4.3.1 The **probability mass function** (abbreviated **pmf**) of a *discrete* random variable X is

$$f_X(x) = \mathbb{P}(X = x) \quad \text{for all } x.$$

Question 4.4 We roll a die repeatedly until we obtain a 6. Let X be the number of rolls. Find f_X.

Answer: Suppose we roll the die infinitely many times (regardless of results) and let $E_i = \{\text{Roll } i \text{ gives a } 6\}$. Then, for $x \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{P}(X = x) &= \mathbb{P}(E_1^c \cap E_2^c \cap \cdots \cap E_{x-1}^c \cap E_x) \\
&= \mathbb{P}(E_1^c) \cdots \mathbb{P}(E_{x-1}^c)\mathbb{P}(E_x) \\
&= \left(1 - \frac{1}{6}\right)^{x-1} \cdot \frac{1}{6} \\
&= \left(\frac{5}{6}\right)^{x-1} \cdot \frac{1}{6}.
\end{aligned}$$

Hence, $f_X(x) = \left(\frac{5}{6}\right)^{x-1} \cdot \frac{1}{6}$ for any integer x, and $f_X(x) = 0$ for any x which is not an integer. As we will see later, random variables with pmf's of this form are called *geometric* random variables. ■

Note: If X is discrete, then

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{y \leq x} \mathbb{P}(X = y) = \sum_{y \leq x} f_X(y).$$

Also, as we've seen,

$$F_X(x) - \lim_{y \nearrow x} F_X(y) = f_X(x),$$

so the pmf f_X describes the jumps of the cdf F_X.

Definition 4.3.2 The **probability density function** (abbreviated **pdf**) of a continuous random variable X is a function $f_X : \mathbb{R} \to [0,\infty)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy \quad \text{for all } x \in \mathbb{R}.$$

By the Fundamental Theorem of Calculus,

$$X \text{ has continuous density } f_X \implies \frac{dF_X}{dx} = f_X.$$

Also note: for $a < b$,

$$\mathbb{P}(a < X \le b) = F_X(b) - F_X(a) = \int_a^b f_X(y)\,dy.$$

The cdf always exists.
The pmf and the pdf only exist in particular cases.

Technical comment: As seen above, if f_X is a continuous pdf of X, then it is the derivative of the cdf F_X, and therefore is unique (meaning that if g_X is also a continuous pdf of X, then $f_X = g_X$). However, in general a pdf is not unique. Indeed, if you change the value of $f_X(y)$ for countably many y, then the integral $\int_{-\infty}^{x} f_X(y)\,dy$ is not affected. We will tend to ignore this technicality and make statements of the form *"... **the** pdf of X satisfies ..."* when it might be more rigorous to say something like *"... **a** pdf of X satisfies..."*.

Definition 4.3.3 A random variable is **absolutely continuous** if it is continuous and has a probability density function.

Notation:

we write	meaning:
$X \sim F_X$	X has cdf $= F_X$
$X \sim f_X$ (X discrete)	X has pmf $= f_X$
$X \sim f_X$ (X absolutely continuous)	X has pdf $= f_X$

Question 4.5 Assume X has pdf

$$f_X(x) = \begin{cases} cxe^{-3x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find c.

Answer: Note that

$$1 = \mathbb{P}(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^{\infty} cxe^{-3x}\,dx,$$

so

$$\int_0^{\infty} cxe^{-3x}\,dx = 1,$$

so

$$c = \frac{1}{\int_0^\infty x e^{-3x}\,\mathrm{d}x}.$$

Integrating by parts, $c = 9$. ∎

4.4 Function of a random variable

In this section, we will have a random variable X and a function $g\colon \mathbb{R} \to \mathbb{R}$; we will study the random variable $Y = g(X)$.
For any $A \subset \mathbb{R}$, we have

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in g^{-1}(A)).$$

We will first review properties of the inverse mapping $g^{-1}(A)$; second consider the cdf of $g(X)$; third specialize to the discrete setting; and finally specialize to the absolutely continuous setting.

Review on inverse mappings

If $g\colon \mathbb{R} \to \mathbb{R}$, the inverse mapping g^{-1} is defined by:

$$g^{-1}(A) = \{x \in \mathbb{R} \text{ such that } g(x) \in A\}, \quad A \subset \mathbb{R}.$$

g^{-1} maps sets into sets. For sets A with a single element, say $A = \{x\}$, we write $g^{-1}(x)$ instead of $g^{-1}(\{x\})$.

Facts:

1. $g(x) = y$ for every $x \in g^{-1}(y)$.
2. If g is one-to-one, then $g^{-1}(y)$ either is empty or has a single element.
3. If g is strictly increasing or strictly decreasing, then it is one-to-one. Additionally,

$$g^{-1}((-\infty, y]) = \begin{cases} (-\infty, g^{-1}(y)], & \text{if } g \text{ is strictly increasing,} \quad (\star) \\ [g^{-1}(y), \infty), & \text{if } g \text{ is strictly decreasing.} \quad (\star\star) \end{cases}$$

cdf of $Y = g(X)$

In general we have:

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(g(X) \in (-\infty, y]) = \mathbb{P}(X \in g^{-1}((-\infty, y])). \tag{4.1}$$

If X is discrete, this becomes: $F_Y(y) = \sum_{x \in g^{-1}((-\infty, y])} f_X(x)$.

If X is continuous, $F_Y(y) = \int_{\{x \in g^{-1}((-\infty, y]\}} f_X(u)\,\mathrm{d}u$.

Specializing (4.1) to the cases where g is strictly increasing or strictly decreasing, we get:

Theorem 4.4.1

1. If $Y = g(X)$ and g is strictly increasing, then $F_Y(y) = F_X(g^{-1}(y))$.
2. If $Y = g(X)$, g is strictly decreasing and X is a continuous random variable, then $F_Y(y) = 1 - F_X(g^{-1}(y))$.

Proof.

1. By (4.1) we have $F_Y(y) = \mathbb{P}(X \in g^{-1}((-\infty, y]))$. Using $(\star)$ this gives
$$F_Y(y) = \mathbb{P}(X \in (-\infty, g^{-1}(y)]) = F_X(g^{-1}(y)).$$
2. Similarly, using (4.1) and $(\star\star)$,
$$F_Y(y) = \mathbb{P}(X \in [g^{-1}(y), \infty)) = \mathbb{P}(X \geq g^{-1}(y)) = \mathbb{P}(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y));$$
in the third equality we have used the fact that X is a continuous random variable.

■

pmf of $Y = g(X)$

Assume X is discrete. Then, Y is also discrete and

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(X \in g^{-1}(y)) = \sum_{x \in g^{-1}(y)} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(y)} f_X(x).$$

If g is one-to-one (in particular, if it is strictly increasing), this gives

$$f_Y(y) = f_X(g^{-1}(y)).$$

pdf of $Y = g(X)$

For continuous X, we have:

Theorem 4.4.2 Assume X has pdf f_X and $Y = g(X)$ with g differentiable and strictly increasing or strictly decreasing. Then,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \tfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|.$$

Proof. Assume first that g is strictly increasing. Then,

$$f_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \cdot \tfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y).$$

If g is strictly decreasing,

$$f_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} \left(1 - F_X(g^{-1}(y))\right) = f_X(g^{-1}(y)) \cdot \left(-\tfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right).$$

■

Question 4.6 Let X be a continuous random variable with pdf $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be the function defined by $g(x) = x^3$. Compute the pdf of $Y = g(X)$.

Answer: Note that g is differentiable and strictly increasing, with $g^{-1}(y) = y^{1/3}$ and $\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = \frac{1}{3}|y|^{-2/3}$. Thus

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \tfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| = \frac{1}{\sqrt{2\pi}} e^{-(y^{1/3})^2/2} \cdot \left| \frac{1}{3}|y|^{-2/3} \right| = \frac{1}{3\sqrt{2\pi}} e^{-y^{2/3}/2} |y|^{-2/3}.$$

■

4.5 Exercises

4.5.1 Check-up the basics

Exercise 4.1 — Decide whether or not each statement is necessarily true.

1. Any random variable is either discrete or continuous.
2. If F_X is the cdf of a random variable X, then $F_X(x) = 0$ for all $x \leq 0$.
3. Let X be a random variable corresponding to the result of rolling a fair die. Then, $f_X(3) = \frac{1}{6}$ and $F_X(3) = \frac{1}{2}$.
4. If X and Y are two random variables defined on a probability space $(S, \mathscr{B}, \mathbb{P})$, then $F_{X+Y} = F_X + F_Y$.

■

Exercise 4.2 — Decide whether or not each statement is necessarily true.

1. If X is a continuous random variable, $g\colon \mathbb{R} \to \mathbb{R}$ and $Y = g(X)$, then Y cannot be discrete.
2. If X has pdf $f_X(x) = 3x^2$, $0 < x < 1$ and $Y = \sqrt{X}$, then $f_Y(x) = 1$, $0 < x < 1$.

■

4.5.2 Statement from the lecture

Exercise 4.3 Prove that the distribution function F_X of a random variable X satisfies:

(a) $\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1.$

(b) F_X is nondecreasing.

(c) F_X is right continuous, that is, for any x, $\lim_{y \searrow x} F_X(y) = F_X(x)$.

■

4.5.3 Problems

Exercise 4.4 An appliance store receives a shipment of 30 ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens at random, without replacement, and tests to see if they are defective. Let X be the number of defectives found. Calculate the pmf and cdf of X and plot the cdf. ■

Exercise 4.5 A random variable X is said to follow an **exponential distribution** with *rate* parameter $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

This distribution is well known to be “memoryless" because it satisfies:

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t) \text{ for any } s,t > 0.$$

Prove this property. ■

Exercise 4.6 An electronic device has lifetime denoted by T, which follows an exponential distribution with scale parameter $\lambda = 1.5$. The device has value $V = 5$ if it fails before time $t = 3$; otherwise, it has value $V = 2T$. Find the cdf of V. ■

Exercise 4.7 In each of the following find the pdf of Y.

a) $Y = X^2$ and $f_X(x) = 1$, $0 < x < 1$.

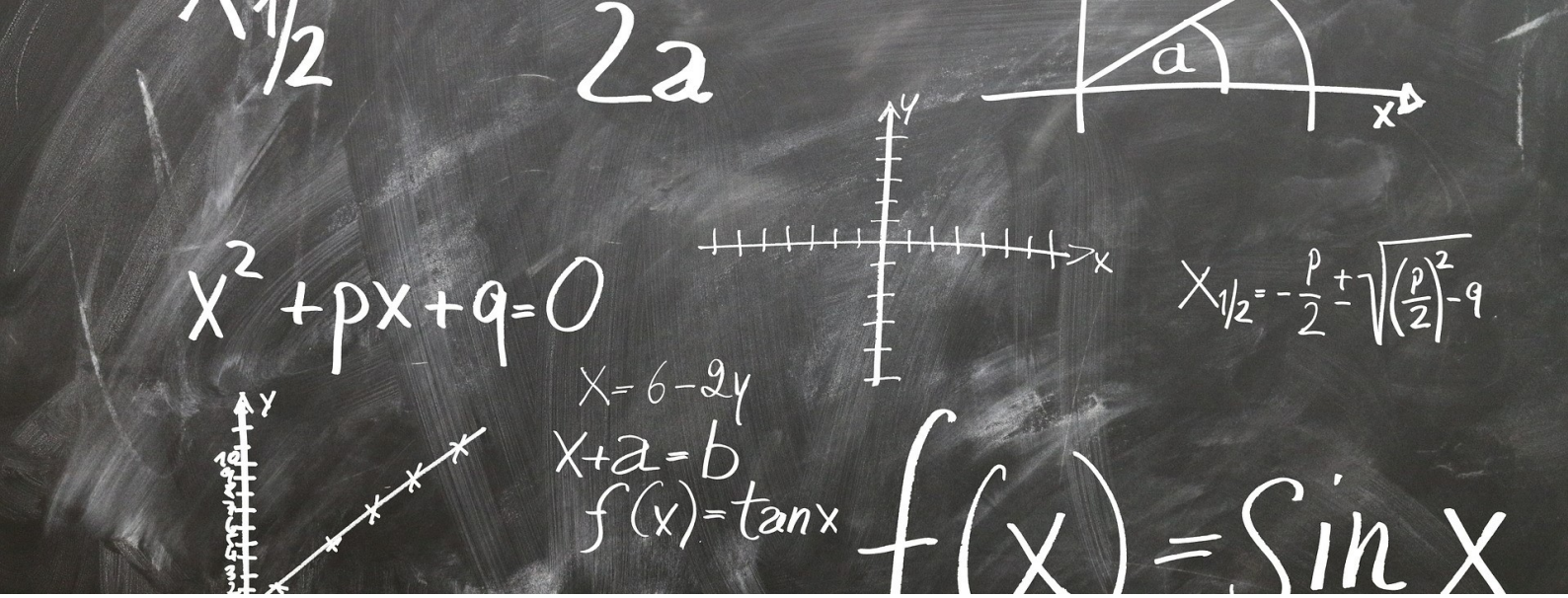
b) $Y = -\log X$ and X has pdf

$$f_X(x) = \frac{(n+m+1)!}{n!\,m!} x^n (1-x)^m, \quad 0 < x < 1.$$

■

4.5.4 Are you up for a challenge?

Exercise 4.8 — Banach's matches problem. At first, Banach has n matches in his left pocket and n matches in his right pocket. Whenever he needs a match, he chooses a pocket (each with probability $\frac{1}{2}$) and takes a match from there. Consider the first instant in which he reaches for a pocket and finds it empty. Let X be the number of matches in the *other* pocket at this instant. Find the distribution of X (that is, find $\mathbb{P}(X = k)$ for $k = 0, 1, \ldots, n$). ■



5. Expectation and variance

The goals of this chapters are:

- ▷ to define the **expectation**, **variance** and **standard deviation** of a random variable which is either discrete or absolutely continuous;
- ▷ to derive a few properties of the expectation and variance.

5.1 Expectation

Definition 5.1.1 The **expectation** (or *expected value* or *mean*) of a random variable X is

$$\mathbb{E}(X) = \begin{cases} \sum_x x \cdot f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x, & \text{if } X \text{ is absolutely continuous,} \end{cases}$$

provided that the sum or integral exists.

Technical comment: The expectation can also be defined for random variables that are neither discrete nor continuous. As this definition requires Measure Theory, we will not see it, and only treat discrete or absolutely continuous random variables.
In the rest of the lecture will always assume (often implicitly) that the random variables we consider are either discrete or absolutely continuous; even if most of (if not all) the properties we will see hold without this restriction.

Question 5.1 Let X be the result of rolling a die. Compute the expectation of X.

Answer:

$$\mathbb{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

(Note that we cannot say that this is a "typical" result of rolling a die!) ■

Question 5.2 Suppose $(\Omega, \mathscr{A}, \mathbb{P})$ is a probability space, A is an event and X is defined by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}$$

X is called the **indicator function** of A (sometimes denoted by $\mathbb{1}_A$). Compute the expectation of X.

Answer: We have

$$\mathbb{E}(X) = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot \mathbb{P}(A^c) + 1 \cdot \mathbb{P}(A) = \mathbb{P}(A).$$

■

Question 5.3 Let X be an absolutely continuous random variable with pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where λ is a positive parameter. Compute the expectation of X.

Answer:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x = \int_0^{\infty} x \cdot \lambda e^{-\lambda x}\,\mathrm{d}x = \underbrace{[-xe^{-\lambda x}]_{x=0}^{\infty}}_{=0} + \underbrace{\int_0^{\infty} e^{-\lambda x}\,\mathrm{d}x}_{=[\frac{1}{\lambda}e^{-\lambda x}]_{x=0}^{\infty} = \frac{1}{\lambda}} = \frac{1}{\lambda}.$$

■

R A random variable as in the previous example is said to be *exponentially distributed* with rate parameter λ. We will talk about it again in the next chapter.

Question 5.4 Let X be an absolutely continuous random variable with pdf

$$f_X(x) = \begin{cases} 1/2x^2, & \text{if } |x| \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Does the expectation of X exist? If yes, what is its value?

Answer: If the expectation of X would exists, we would have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x = \int_{-\infty}^{-1} \frac{1}{2x}\,\mathrm{d}x + \int_1^{\infty} \frac{1}{2x}\,\mathrm{d}x = -\infty + \infty.$$

But this quantity is undefined. Thus, the expectation does not exist. ■

Note that a linear combination $a_1X_1 + \cdots + a_nX_n$ of random variables defined on the same probability space, is also a random variable. The next theorem provides a very useful tool to compute its expectation.

Theorem 5.1.1 — Linearity of the expectation. Let $X_1, \dots, X_n$ be random variables defined on the same probability space, i.e. X_i is a function $X_i \colon \Omega \to \mathbb{R}$, for any i. Let $a_1, \dots, a_n \in \mathbb{R}$. Then

$$\mathbb{E}[a_1X_1 + \cdots + a_nX_n] = a_1\mathbb{E}X_1 + \cdots + a_n\mathbb{E}X_n,$$

provided that the expectations involved are all well defined.

Proof. (discrete case only) First we consider $n = 2$. To shorten notation, we set $X = a_1X_1 + a_2X_2$. Note that

$$\begin{aligned} f_X(x) = \mathbb{P}(X = x) = \mathbb{P}(a_1X_1 + a_2X_2 = x) &= \mathbb{P}(\cup_{x_1,x_2:a_1x_1+a_2x_2=x}\{X_1 = x_1 \text{ and } X_2 = x_2\}) \\ &= \sum_{x_1,x_2:a_1x_1+a_2x_2=x} \mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2\}). \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}X = \sum_x x f_X(x) &= \sum_x \sum_{x_1,x_2:a_1x_1+a_2x_2=x} x\,\mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) \\ &= \sum_{x_1,x_2} (a_1x_1 + a_2x_2)\mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) \\ &= a_1 \sum_{x_1} x_1 \sum_{x_2} \mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) + a_2 \sum_{x_2} x_2 \sum_{x_1} \mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2). \end{aligned}$$

But, for any x_1,

$$\sum_{x_2} \mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) = \mathbb{P}(\cup_{x_2}\{X_1 = x_1 \text{ and } X_2 = x_2\}) = \mathbb{P}(X_1 = x_1) = f_{X_1}(x_1),$$

and similarly $\sum_{x_1} \mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) = f_{X_2}(x_2)$ for any x_2. Therefore

$$\mathbb{E}X = a_1 \sum_{x_1} x_1 f_{X_1}(x_1) + a_2 \sum_{x_2} x_2 f_{X_2}(x_2) = a_1\mathbb{E}X_1 + a_2\mathbb{E}X_2.$$

The case $n \geq 3$ follows with an induction. ■

Question 5.5 Roll one hundred dies and let X be the sum of the results. Compute the expectation of X.

Answer: Using the linearity of the expectation, we have

$$\mathbb{E}X = \mathbb{E}X_1 + \ldots + \mathbb{E}X_{100} = 3.5 + \ldots + 3.5 = 350,$$

where X_i denotes the result of the die i, and the second equality was computed in Example 5.1. Note that this computation is much faster to do than if we would like to apply directly the definition of the expectation. Indeed, it would require first to compute $\mathbb{P}(X = x)$ for all possible results of the experiment, which are all integer numbers between $100 \cdot 1 = 100$ and $100 \cdot 6 = 600$. ■

Theorem 5.1.2 For $g \colon \mathbb{R} \to \mathbb{R}$ and a random variable X,

$$\mathbb{E}(g(X)) = \begin{cases} \sum_x g(x) \cdot f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\,\mathrm{d}x, & \text{if } X \text{ is absolutely continuous,} \end{cases}$$

provided that the sum or integral exists.

Technical comment: Once more, we decide to ignore some technical details that are behind the scope of this course. The theorem above holds for most functions g you can think of (e.g. continuous functions, piecewise continuous functions,...) but not for very very weird functions, which one calls *non-measurable functions*. For the rest of this course we will always assume (without mentioning it) that functions are sufficiently nice (or in technical terms, are *measurable*).

Proof. (discrete case only). Putting $Y = g(X)$,

$$\mathbb{E}(Y) = \sum_y y \cdot f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} g(x) \cdot f_X(x) = \sum_x g(x) \cdot f_X(x).$$

■

Theorem 5.1.3 If $X, Y : \Omega \to \mathbb{R}$ be random variables, a, b, $c \in \mathbb{R}$ and $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(g_1(X))$, $\mathbb{E}(g_2(X))$ and $\mathbb{E}(g_2(Y))$ exist, then,

1. $\mathbb{E}(ag_1(X) + bg_2(Y) + c) = a\mathbb{E}(g_1(X)) + b\mathbb{E}(g_2(Y)) + c.$ (linearity)
2. If $g_1 \geq 0$, then $\mathbb{E}(g_1(X)) \geq 0$.
3. If $g_1 \geq g_2$, then $\mathbb{E}(g_1(X)) \geq \mathbb{E}(g_2(X))$.

   } (monotonicity)
4. If $a \leq g_1(X) \leq b$, then $a \leq \mathbb{E}(g_1(X)) \leq b$.

Proof. The proof is very easy, using Theorems 5.1.2 and 5.1.3. ■

We conclude with a theorem that gives alternate formulas for the expectation of **non-negative** random variables.

Theorem 5.1.4

1. If X is a discrete random variable that only assumes values on $\{0, 1, 2, \ldots\}$, then

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} (1 - F_X(n)).$$

2. If X is a continuous and non-negative random variable, then

$$\mathbb{E}(X) = \int_0^{\infty} (1 - F_X(x))\, \mathrm{d}x.$$

Proof. **Discrete case:**

$$\begin{aligned}\mathbb{E}(X) &= \sum_{n=1}^{\infty} n \cdot f_X(n) = \sum_{n=1}^{\infty} \sum_{m=1}^{n} f_X(n) = \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} f_X(n) = \sum_{m=1}^{\infty} \mathbb{P}(X \geq m) = \sum_{m=0}^{\infty} \mathbb{P}(X > m) \\ &= \sum_{m=0}^{\infty} (1 - F_X(m)).\end{aligned}$$

Continuous case:

$$\begin{aligned}\mathbb{E}(X) &= \int_0^{\infty} x \cdot f_X(x)\, \mathrm{d}x = \int_0^{\infty} \int_0^{x} f_X(x)\, \mathrm{d}y\, \mathrm{d}x = \int_0^{\infty} \int_y^{\infty} f_X(x)\, \mathrm{d}x\, \mathrm{d}y = \int_0^{\infty} \mathbb{P}(X > y)\, \mathrm{d}y \\ &= \int_0^{\infty} (1 - F_X(y))\, \mathrm{d}y.\end{aligned}$$

■

5.2 Variance

Definition 5.2.1 For a random variable X and an integer n, the **n-th moment of** X is $\mathbb{E}(X^n)$, and the **n-th central moment of** X is $\mathbb{E}((X-\mathbb{E}X)^n)$.

The second central moment, is also called the **variance** of X and is denoted by:

$$\mathrm{Var}(X) = \mathbb{E}((X-\mathbb{E}X)^2).$$

The positive square root $\sigma = \sqrt{\mathrm{Var}(X)}$ is called the **standard deviation** of X.

Technical comment: The (central) n-th moment is only defined if the corresponding integral (continuous case) or sum (discrete case) can be assigned a value unambiguously (which might be $+\infty$ or $-\infty$).

By developing the square in the expectation in the definition of variance, and writing $\mu := \mathbb{E}X$ for convenience, we see that:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}((X-\mu)^2)\\ &= \mathbb{E}(X^2 - 2\mu X + \mu^2)\\ &= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2\\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2\\ &= \mathbb{E}(X^2) - \mu^2.\end{aligned}$$

This gives rise to the following alternate formula for the variance:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

The variance of X is a measure of the *dispersion* of the value of X from its expected value μ. To get some intuition, let us consider an example where the variance is zero. Assume that X is discrete and $\mathrm{Var}(X) = 0$. Then, we have

$$0 = \mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \sum_{x: f_X(x)>0} (x-\mu)^2 \cdot f_X(x) = 0.$$

This is only possible if $f_X(\mu) = 1$, that is, $\mathbb{P}(X=\mu) = 1$, that is, X is with probability 1 equal to the number μ. We see that in this example the fact that the variance is zero implies that there is no dispersion at all.

Question 5.6 Let $n \in \mathbb{N}$ and $p \in (0,1)$. Let X be a discrete random variable with pmf

$$f_X(-n) = f_X(n) = \tfrac{p}{2}, \qquad f_X(0) = 1-p.$$

(Recall that, since these three probabilities sum up to one, they define the whole probability function and in particular X takes only three values, $-n$, 0 and n.) Compute the variance of X.

Answer: We have

$$\mathbb{E}X = \frac{p}{2}\cdot(-n) + (1-p)\cdot 0 + \frac{p}{2}n = 0,$$

and

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = (-n)^2\cdot\tfrac{p}{2} + 0^2\cdot(1-p) + n^2\cdot\tfrac{p}{2} = n^2 p.$$

Note that the variance is increasing in both n and p. ■

Theorem 5.2.1 If X is a random variable and a, b are constants,

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X).$$

Proof.

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

■

5.3 Exercises

5.3.1 Check-up the basics

Exercise 5.1 — Decide whether or not each statement is necessarily true.

1. If X is a random variable and $g\colon \mathbb{R}\to\mathbb{R}$ satisfies $g(x)\geq x$ for all x, then $\mathbb{E}(g(X))\geq\mathbb{E}(X)$.
2. If X is a random variable, $g\colon \mathbb{R}\to\mathbb{R}$ and $\mathbb{E}(g(X))\geq\mathbb{E}(X)$, then $g(x)\geq x$ for all x.
3. If X is a random variable and $g\colon \mathbb{R}\to\mathbb{R}$, then $\mathbb{E}(g(X)) = g(\mathbb{E}(X))$.
4. There exists a random variable X such that F_X is continuous and $\mathrm{Var}(X) = 0$.

■

5.3.2 Problems

Exercise 5.2 Suppose X has pmf

$$f_X(x) = \frac{1}{3}\left(\frac{2}{3}\right)^x, \quad x\in\{0,1,2,\ldots\}.$$

Suppose $Y = q^X$, where q is a non-negative real number. Find the pmf of Y, determine the values of q for which $\mathbb{E}(Y)$ exists and, in case this expectation exists, compute it. ■

Exercise 5.3 A random variable X has pdf

$$f_X(x) = c\frac{1}{x^2-x}, \qquad 2<x<5.$$

Find c and $\mathbb{E}(\lfloor X\rfloor)$ ($\lfloor\cdot\rfloor$ denotes the *floor function*. For a non-negative real number x, $\lfloor x\rfloor$ is the largest integer n such that $n\leq x$). ■

Exercise 5.4 Let $g\colon [0,1]\to\mathbb{R}$ be a continuous function. For each $n\in\mathbb{N}$, let X_n be a random variable with pmf

$$f_{X_n}(i/n) = n^{-1}, \quad i\in\{1,\ldots,n\}.$$

What is $\lim_{n\to\infty} \mathbb{E}(g(X_n))$? ■

Exercise 5.5 A *median* of a distribution is a value m such that $\mathbb{P}(X \le m) \ge \frac{1}{2}$ and $\mathbb{P}(X \ge m) \ge \frac{1}{2}$. (If X is continuous, m satisfies $\int_{-\infty}^{m} f_X(x)\,dx = \int_{m}^{\infty} f(x)\,dx = \frac{1}{2}$). Find the median of the distribution

$$f_X(x) = 3x^2, \quad 0 < x < 1.$$

■

Exercise 5.6 Compute $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ for each of the following probability distributions.

(a) $f_X(x) = ax^{a-1}$, $0 < x < 1$, $a > 0$.

(b) $f_X(x) = \frac{1}{n}$, $x = 1, 2, \ldots, n$, $n > 0$ an integer. Use the formulas:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

(c) $f_X(x) = \frac{3}{2}(x-1)^2$, $0 < x < 2$.

■

5.3.3 Are you up for a challenge?

Exercise 5.7 Prove that, if X is a continuous and non-negative random variable, then

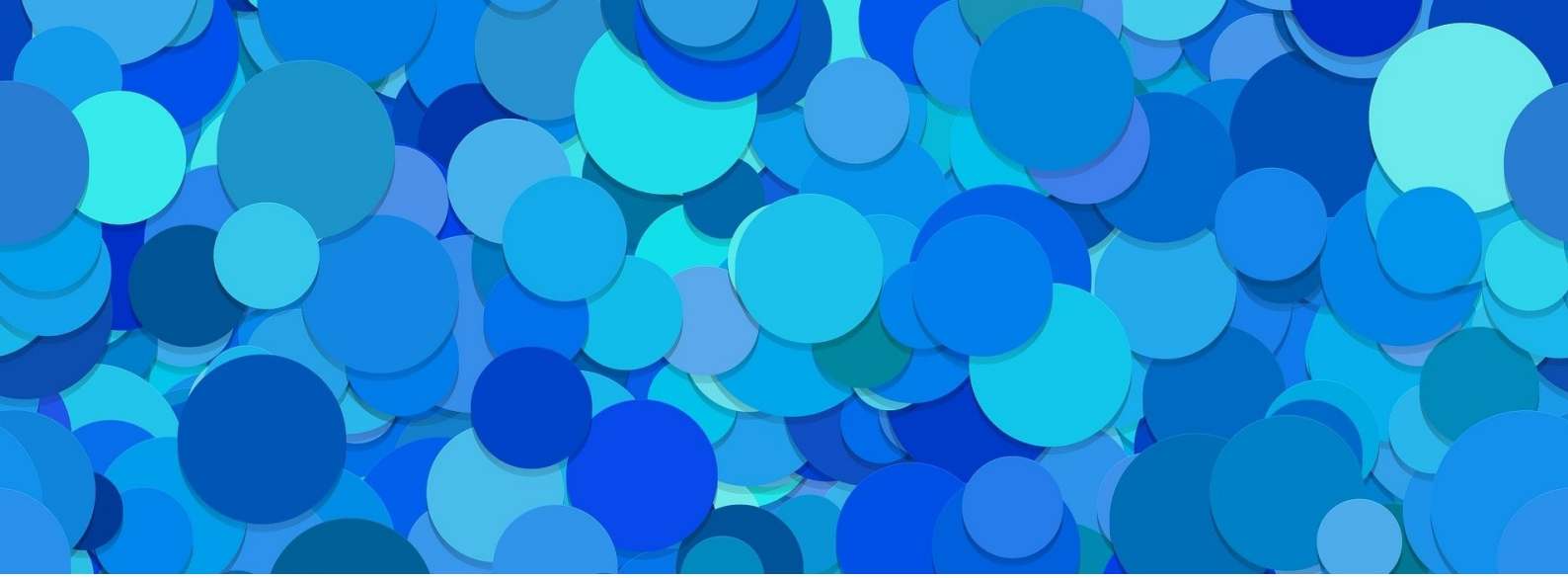
$$\mathbb{E}(X^k) = k \int_0^{\infty} x^{k-1}(1 - F_X(x))\,dx.$$

Hint. Start with

$$\mathbb{E}(X^k) = \int_0^{\infty} \mathbb{P}(X^k > x)\,dx.$$

■

Exercise 5.8 Assume X is a continuous random variable such that the cdf F_X is strictly increasing. Let $Y = F_X(X)$. Find the cdf of Y.
Hint. If F_X is strictly increasing, it is invertible. ■



6. Classical discrete distributions

The goals of this chapters are:

- ▷ to introduce **important families of discrete distributions**: uniform, Bernoulli, binomial, geometric and Poisson;
- ▷ to **compute their expectations and variances**.

6.1 Discrete uniform distribution

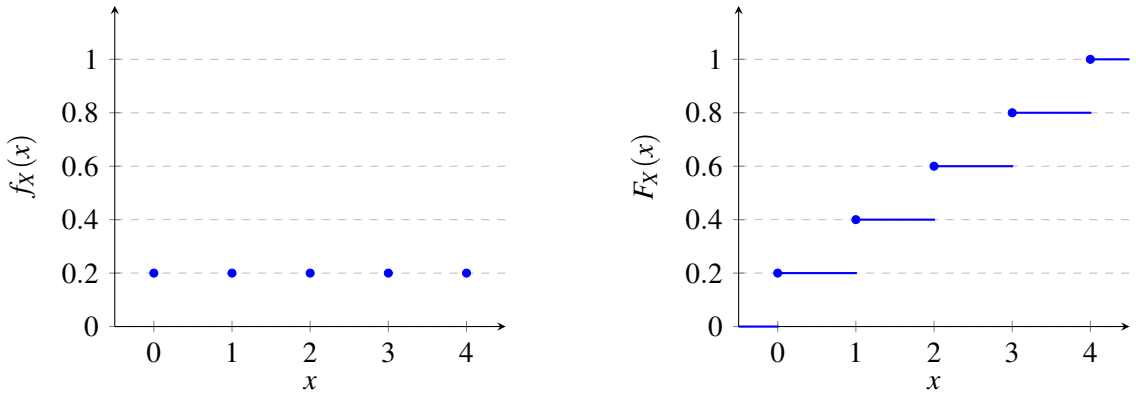


Figure 6.1: PMF and CDF of a discrete uniform distribution on $\{0,1,2,3,4\}$.

Let a and b integers with $a < b$.
A random variable X follows a **discrete uniform distribution** with parameters a and b (abbreviated $X \sim \text{Unif}(a,b)$) if

$$f_X(x) = \frac{1}{b-a+1}, \quad x = a,\ a+1,\ldots,\ b$$

(in words: X is equally likely to be equal to any of the integers between (and including) a and b).

▷ Expectation $\mathbb{E}(X) = \dfrac{a+b}{2}$

▷ Variance $\mathrm{Var}(X) = \dfrac{(b-a+1)^2-1}{12}$

(These formulas will be verified in the tutorial).

6.2 Bernoulli distribution

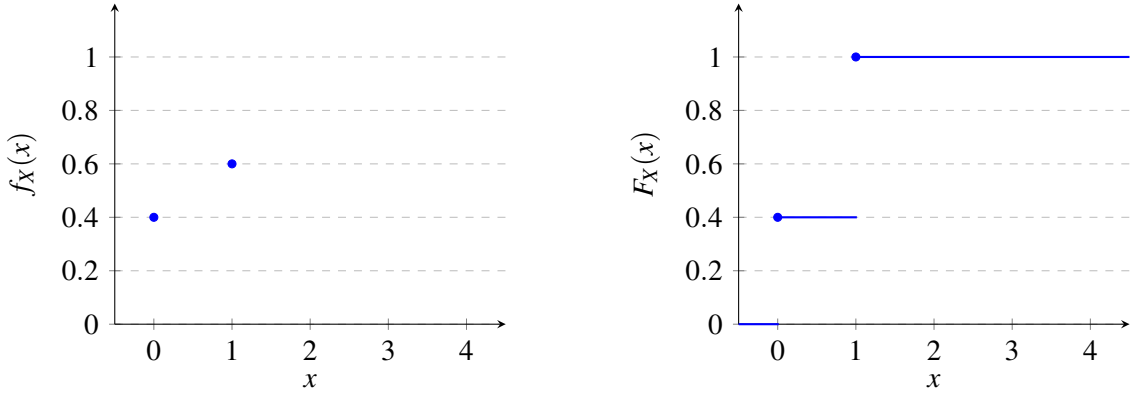


Figure 6.2: PMF and CDF of a Bernoulli with parameter $p = 0.6$.

Let $p \in [0,1]$.

A random variable X follows a **Bernoulli distribution** with parameter p (abbreviated $X \sim \mathrm{Ber}(p)$) if

$$f_X(1) = p, \qquad f_X(0) = 1-p.$$

▷ $\mathbb{E}(X) = 0 \cdot (1-p) + 1 \cdot p = p.$

▷ $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2 = 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 = p - p^2 = p(1-p).$

A **Bernoulli trial** is an experiment which results in success with probability p and failure with $1-p$. X is then 1 when there is success and 0 when there is failure.

6.3 Binomial distribution

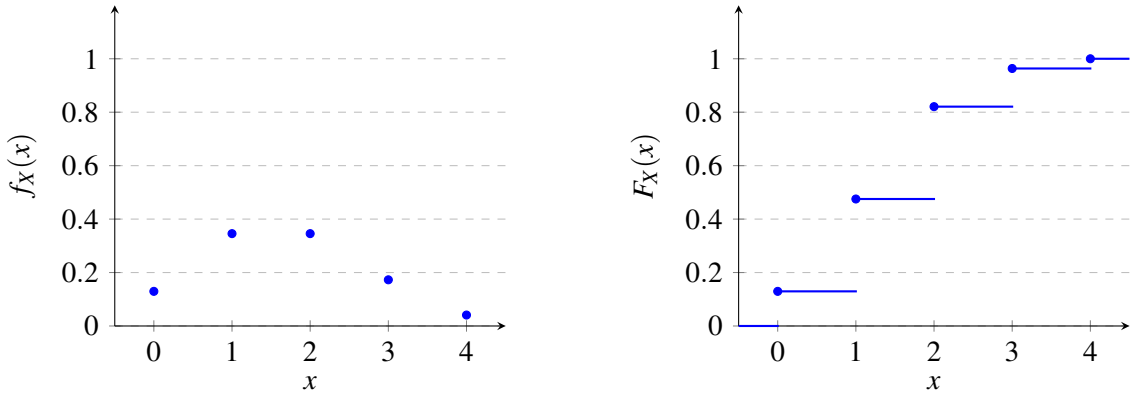


Figure 6.3: PMF and CDF of a Binomial distribution with parameter $n = 4$ and $p = 0.6$.

Let n be positive integer and $p \in [0,1]$.
Suppose we perform n independent Bernoulli trials with probability p of success (each). Let

$$X = \text{Number of successes obtained.}$$

Let us carefully define the probability space that corresponds to this situation.

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \{0,1\} \text{ for each } i\}$$

(1 means success and 0 means failure). This sample space is given the probability function

$$\mathbb{P}(\{\omega\}) = p^{\#1\text{'s in } \omega} \cdot (1-p)^{\#0\text{'s in } \omega} = p^{\sum \omega_i} \cdot (1-p)^{\sum(1-\omega_i)} = p^{\sum \omega_i} \cdot (1-p)^{n-\sum \omega_i}.$$

The random variable X is defined as

$$X(\omega) = \sum_{i=1}^{n} \omega_i,$$

so

$$\mathbb{P}(X=x) = \sum_{\substack{\omega \text{ such that} \\ X(\omega)=x}} \mathbb{P}(\{\omega\}) = \sum_{\substack{\omega \text{ such that} \\ X(\omega)=x}} p^x(1-p)^{n-x} = \#\{\omega \text{ such that } \sum \omega_i = x\} \cdot p^x \cdot (1-p)^{n-x}.$$

The number of ω such that $\sum \omega_i = x$ is the number of "words" that we can form using x times the symbol '1' and $n-x$ times the symbol '0'. This number is equal to $\frac{n!}{x!(n-x)!} = \binom{n}{x}$. This gives the formula

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

For X with this pmf, we say that X follows a **binomial distribution** with parameters n and p (abbreviated $X \sim \text{Bin}(n,p)$).

▷ Expectation. Similarly as in Example 5.5, we can use the linearity of the expectation to get:

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}X_i = np,$$

where X_i denotes the result of the i-th Bernoulli trial, that is $X_i(\omega) = \omega_i$; it is a Bernoulli random variable of parameter p and thus its expectation is p, which explains the second equality of the last displayed equation.
(Alternative lengthy computation) It is also possible to compute this expectation without using the linearity property as follow.

$$\mathbb{E}(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x}. \tag{6.1}$$

We now observe the identity

$$x\binom{n}{x} = \frac{xn!}{x!(n-x)!} = \frac{n!}{(x-1)!(n-x)!} = n \cdot \frac{(n-1)!}{(x-1)!(n-x)!} = n\binom{n-1}{x-1}.$$

The right-hand side of (6.1) is thus equal to

$$n \sum_{x=1}^{n} \binom{n-1}{x-1} p^x (1-p)^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}.$$

With the change of variable $y = x-1$, this becomes

$$\mathbb{E}(X) = np \sum_{y=0}^{n-1} \underbrace{\binom{n-1}{y} p^y (1-p)^{(n-1)-y}}_{\mathbb{P}(Y=y) \text{ for } Y \sim \mathrm{Bin}(n-1,p)} = np,$$

and we reach the same conclusion as above.

▷ Variance $\mathrm{Var}(X) = np(1-p)$. This can be proved with a similar computation as the one carried out for the expectation. We won't do this now because later we will learn a much, much simpler method.

6.4 Geometric distribution

Let $p \in [0,1]$.
Suppose we perform Bernoulli trials with probability p of success until the first success is obtained. Let

$$X = \#\text{trials we end up performing.}$$

Then, for $x \in \{1,2,\dots\}$,

$$\mathbb{P}(X = x) = \mathbb{P}(\text{failure in trials } 1,\dots,x-1, \text{ success in trial } x) = (1-p)^{x-1}p.$$

For X with

$$f_X(x) = (1-p)^{x-1}p, \quad x = 1,2,\dots,$$

we say that X follows a **geometric distribution** with parameter p (abbreviated $X \sim \mathrm{Geo}(p)$).

▷ Expectation

$$\mathbb{E}[X] = \sum_{x \geqslant 1} x \cdot (1-p)^{x-1} \cdot p. \tag{6.2}$$

We now use a trick: note that

$$x \cdot (1-p)^{x-1} = -\tfrac{\mathrm{d}}{\mathrm{d}p}\left[(1-p)^x\right],$$

so the sum on the right-hand side of (6.2) becomes

$$-p \sum_{x \geqslant 1} \tfrac{\mathrm{d}}{\mathrm{d}p}\left[(1-p)^x\right] \overset{(\star)}{=} -p \tfrac{\mathrm{d}}{\mathrm{d}p} \sum_{x \geqslant 1} (1-p)^x \overset{(\star\star)}{=} -p \tfrac{\mathrm{d}}{\mathrm{d}p}\left(\frac{1-p}{p}\right) = -p \cdot \frac{-p-1+p}{p^2} = \frac{1}{p},$$

so $\mathbb{E}[X] = \frac{1}{p}$. We need to justify steps $(\star)$ and $(\star\star)$. $(\star)$ is an exchange of infinite sum with integral and is seen in Analysis, so we won't say more here. $(\star\star)$ follows from the formula for geometric series:

$$\sum_{n \geqslant 1} q^n = -1 + \sum_{n \geqslant 0} q^n = -1 + \frac{1}{1-q} = \frac{q}{1-q}.$$

▷ Variance: to be done in the tutorial.

6.5 Poisson distribution

Let $\lambda > 0$.
A random variable X follows the **Poisson distribution** with parameter λ (abbreviated $X \sim \text{Poi}(\lambda)$) if

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots.$$

Note:

$$\sum_{x \geqslant 0} f_X(x) = e^{-\lambda} \sum_{x \geqslant 0} \frac{\lambda^x}{x!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

▷ Expectation

$$\mathbb{E}[X] = \sum_{x \geqslant 0} x \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x \geqslant 1} \frac{\lambda^x}{(x-1)!} = e^{-\lambda} \cdot \lambda \cdot \sum_{x \geqslant 1} \frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda} \cdot \lambda \cdot \sum_{y \geqslant 0} \frac{\lambda^y}{y!}$$
$$= e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda.$$

▷ Variance: We use the trick of first computing

$$\mathbb{E}(X(X-1)) = \sum_{x \geqslant 0} x(x-1) \cdot f_X(x) = \sum_{x \geqslant 0} x(x-1) \cdot \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \lambda^2 \sum_{x \geqslant 2} \frac{\lambda^{x-2}}{(x-2)!}$$
$$= e^{-\lambda} \lambda^2 \sum_{y \geqslant 0} \frac{\lambda^y}{y!} = \lambda^2.$$

Hence,

$$\lambda^2 = \mathbb{E}(X(X-1)) = \mathbb{E}(X^2 - X) = \mathbb{E}(X^2) - \mathbb{E}[X] = \mathbb{E}(X^2) - \lambda,$$
$$\implies \mathbb{E}(X^2) = \lambda^2 + \lambda,$$
$$\implies \text{Var}[X] = \mathbb{E}(X^2) - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Poisson approximation to Binomial: this is an approximation which can be summarized by:

$\text{Bin}(n, p)$ is close to $\text{Poi}(\lambda)$ when n is large, p is small and np is close to λ.

Formally, we have:

Proposition 6.5.1 — Poisson limit theorem. Assume $(p_n)_{n \geqslant 1}$ is a sequence such that

$$p_n \in [0, 1] \text{ for each } n \quad \text{and} \quad \lim_{n \to \infty} n \cdot p_n = \lambda > 0.$$

Then, for each $k \geqslant 1$,

$$\underbrace{\binom{n}{k} (p_n)^k (1 - p_n)^{n-k}}_{f_X(k) \text{ for } X \sim \text{Bin}(n, p_n)} \xrightarrow{n \to \infty} \underbrace{\frac{\lambda^k}{k!} e^{-\lambda}}_{f_X(k) \text{ for } X \sim \text{Poi}(\lambda)}$$

Proof.

$$\begin{aligned}\binom{n}{k}(p_n)^k(1-p_n)^{n-k} &= \frac{n!}{k!(n-k)!}\cdot(p_n)^k\cdot\frac{1}{(1-p_n)^k}\cdot(1-p_n)^n\\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}(p_n)^k\cdot\frac{1}{(1-p_n)^k}\cdot(1-p_n)^n\\ &= \frac{1}{k!}\underbrace{(np_n)((n-1)p_n)((n-2)p_n)\cdots((n-k+1)p_n)}_{A_n}\cdot\underbrace{\frac{1}{(1-p_n)^k}}_{B_n}\cdot\underbrace{(1-p_n)^n}_{C_n}\end{aligned}$$

Since $np_n \xrightarrow{n\to\infty} \lambda > 0$, we have $p_n \xrightarrow{n\to\infty} 0$, so $\lim_{n\to\infty} B_n = 1$.
We also have

$$A_n = \underbrace{(np_n)}_{\to\lambda}\cdot\underbrace{(np_n - p_n)}_{\to\lambda\,\to 0}\cdot\underbrace{(np_n - 2p_n)}_{\to\lambda\,\to 0}\cdots\underbrace{(np_n - (k+1)p_n)}_{\to\lambda\,\to 0},$$

so $\lim_{n\to\infty} A_n = \lambda^k$.
To deal with C_n, we will need to use the Taylor expansion of $g(x) = \log(x)$ at $x = 1$, which gives

$$g(1+h) = g(1) + g'(1)h + \varepsilon(h), \qquad \text{with } \lim_{h\to 0}\frac{\varepsilon(h)}{|h|} = 0,$$

so

$$\log(1+x) = 0 + 1\cdot h + \varepsilon(h) = h + \varepsilon(h).$$

With this at hand,

$$\begin{aligned}&C_n = (1-p_n)^n = e^{n\log(1-p_n)}\\ &\implies \lim_{n\to\infty} C_n = e^{\lim_{n\to\infty}[n\log(1-p_n)]} = e^{\lim_{n\to\infty}[n(-p_n+\varepsilon(-p_n))]} = e^{\lim_{n\to\infty}[-np_n]+\lim_{n\to\infty}n\varepsilon(-p_n)}.\end{aligned}$$

Now,

$$\lim_{n\to\infty} np_n = \lambda \quad \text{and} \quad \lim_{n\to\infty} n\varepsilon(-p_n) = \lim_{n\to\infty} np_n\cdot\frac{\varepsilon(-p_n)}{p_n} = 0,$$

so $\lim_{n\to\infty} C_n = e^{-\lambda}$, completing the proof. ■

6.6 Exercises

6.6.1 Problems

Exercise 6.1 Find expressions for $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ if X is a random variable following the discrete uniform distribution on the set $\{a, a+1, \ldots, b\}$, with $a, b \in \mathbb{Z}$, $a < b$ (that is, $f_X(x) = \frac{1}{b-a+1}$, $x \in \{a, a+1, \ldots, b\}$). ■

Exercise 6.2 — Indicator functions and the number of events that occur.
(a) Let E be an event and X be a random variable defined by

$$X(s) = \begin{cases} 1 & \text{if } s \in E,\\ 0 & \text{if } s \in E^c. \end{cases}$$

Show that $X \sim \mathrm{Ber}(p)$. What is p?
Note. X is usually called the **indicator function** of the event E, and is denoted 1_E.

(b) Assume $E_1, E_2, \ldots, E_n$ are events. Explain why the random variable $X = \sum_{i=1}^{n} 1_{E_i}$ can be interpreted as "the number of events that occur" among $E_1, \ldots, E_n$.
(c) In part (b), assuming the E_i's are independent and each has the same probability p, what is the distribution of $\sum_{i=1}^{n} E_i$?

■

Exercise 6.3
(a) A die is rolled 4 times; find the probability that at least one 6 is obtained.
(b) A pair of dice is rolled 24 times; find the probability that at least one double 6 is obtained.

■

Exercise 6.4 For $X \sim \text{Geometric}(p)$, show that $\text{Var}(X) = \frac{1-p}{p^2}$.
Hint. Instead of computing $\mathbb{E}(X^2)$, it will be easier to compute $\mathbb{E}(X(X+1))$. From this, you can find $\mathbb{E}(X^2)$ by using $\mathbb{E}(X(X+1)) = \mathbb{E}(X^2+X) = \mathbb{E}(X^2)+\mathbb{E}(X)$. ■

Exercise 6.5 The probability that a car will have a flat tire while crossing a certain bridge is 0.00005. Use the Poisson approximation to the binomial distribution to estimate the probability that: among 10,000 cars crossing the bridge, more than two will have a flat tire. You will need a calculator. ■

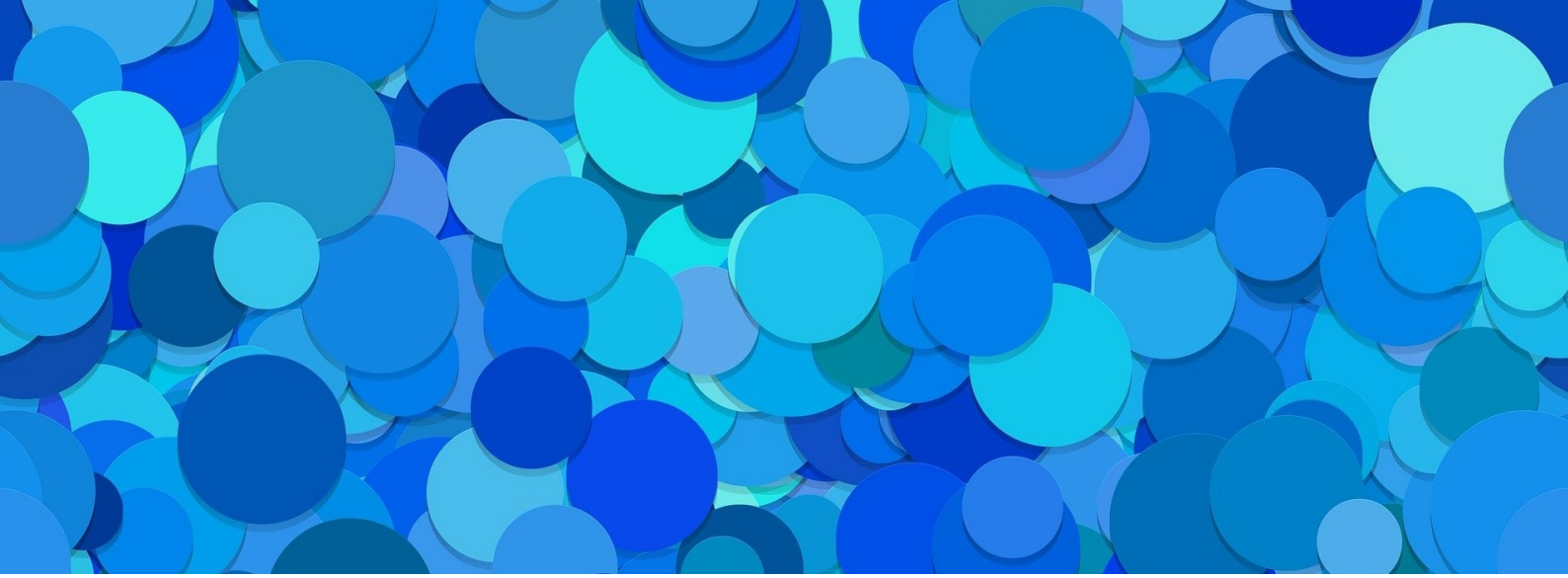
Exercise 6.6 Show that, if $X \sim \text{Geometric}(p)$, then

$$\mathbb{P}(X > x+n \mid X > n) = \mathbb{P}(X > x)$$

for any positive integers x and n. Due to this property, we say that the geometric distribution is *memoryless*. We have seen in an earlier tutorial that the exponential distribution has the same property. See Exercise 7.3 (Challenge Problem)for more on the relationship between the geometric and the exponential distributions. ■

6.6.2 Are you up for a challenge?

Exercise 6.7 A coin with probability of heads equal to p is tossed n times; let X be the number of heads obtained. For each $m \in \{0, 1, \ldots, n\}$, find the value of p that maximizes $\mathbb{P}(X = m)$. ■



7. Classical continuous distributions

The goals of this chapters are:

- ▷ to introduce **important families of continuous distributions**: continuous uniform, exponential, gamma and normal;
- ▷ to **compute their expectations and variances**.

7.1 Uniform distribution

Let $a, b \in \mathbb{R}$ with $a < b$.
A random variable X follows a **continuous uniform distribution** between a and b (abbreviated $X \sim \text{ContUnif}(a,b)$) if it has pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in (a,b), \\ 0, & \text{otherwise.} \end{cases}$$

- ▷ Expectation

$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\,\mathrm{d}x = \frac{b+a}{2}.$$

- ▷ Variance

$$\text{Var}[X] = \int_a^b \left(x - \frac{b+a}{2}\right)^2 \cdot \frac{1}{b-a}\,\mathrm{d}x = \frac{(b-a)^2}{12}.$$

7.2 Exponential distribution

Let $\lambda > 0$.
A random variable X follows an **exponential distribution** of *rate* parameter λ (abbreviated $X \sim \text{Exp}(\lambda)$) if it has pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

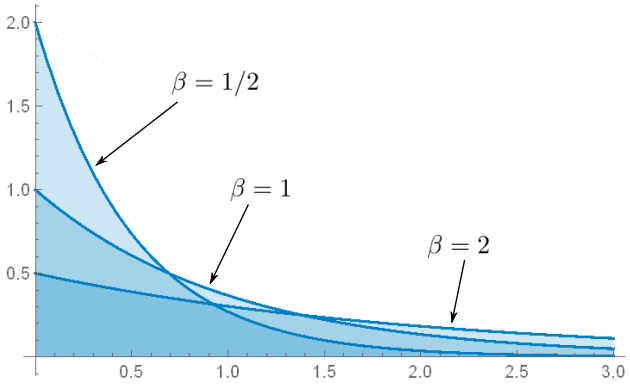


Figure 7.1: The probability density function of $\mathrm{Exp}(1/\beta)$ for different values of β.

Note:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This is a *memoryless* distribution:

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{e^{-(s+t)}}{e^{-s}} = e^{-t} = \mathbb{P}(X > t), \qquad \text{for any } s,t > 0.$$

Expectation and variance are given by what we show more generally for the Gamma distribution below:

- ▷ $\mathbb{E}[X] = 1/\lambda$,
- ▷ $\mathrm{Var}[X] = 1/\lambda^2$.

Sometime the exponential distribution is charaterized by $\beta = 1/\lambda = \mathbb{E}[X]$; this parameter is called the *scale* parameter of the exponential distribution.

7.3 Gamma distribution

Before we introduce the gamma distributed random variable, we need to define the Gamma function $\Gamma\colon (0,\infty) \to \mathbb{R}$ and recall some of its properties.

$$\Gamma(a) = \int_0^\infty t^{a-1} \cdot e^{-t}\,\mathrm{d}t.$$

Fact: $\Gamma(a+1) = a\Gamma(a)$ for any $a > 0$.

Proof.

$$\Gamma(a+1) = \int_0^\infty t^{a} \cdot e^{-t}\,\mathrm{d}t;$$

integrating by parts with $u = t^a$ and $\mathrm{d}v = e^{-t}\,\mathrm{d}t$, this is equal to:

$$(-t^a \cdot e^{-t})\,|_0^\infty + a\int_0^\infty t^{a-1} \cdot e^{-t}\,\mathrm{d}t = 0 + a\Gamma(a).$$

■

Also, $\Gamma(1) = \int_0^\infty e^{-t}\,\mathrm{d}t = 1$, so

$$\Gamma(2) = 1\cdot\Gamma(1) = 1, \qquad \Gamma(3) = 2\cdot\Gamma(2) = 2\cdot 1, \qquad \Gamma(4) = 3\cdot\Gamma(3) = 3\cdot 2\cdot 1\cdots,$$

and more generally

Fact: $\Gamma(n) = (n-1)!$ for any positive integer n.

Let $\alpha,\beta > 0$.
A random variable X follows a **Gamma distribution** of parameters α and β (abbreviated $X \sim \mathrm{Gamma}(\alpha,\beta)$) if it has pdf

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that if one sets $\alpha = 1$ and $\beta = \lambda$, then $f_X(x) = \lambda e^{-\lambda x}$ for any $x > 0$, and we recognize the expression of the pdf of an exponentially distributed random variable with parameter λ. In other words, the exponential distribution with parameter λ is the $\Gamma(1,\lambda)$ distribution.

But, why is f_X a pdf? Doing the change of variable $y = x\beta$, so that $\mathrm{d}x = \beta^{-1}\,\mathrm{d}y$, we have

$$\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}\,\mathrm{d}x = \frac{\beta^{\alpha-1}}{\Gamma(\alpha)} \int_0^\infty (y/\beta)^{\alpha-1}\, e^{-y}\,\mathrm{d}y = \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1}\, e^{-y}\,\mathrm{d}y = 1,$$

so f_X is indeed a pdf. This computation also showed that

$$\int_0^\infty x^{\alpha-1}\, e^{-x\beta}\,\mathrm{d}x = \Gamma(\alpha)\beta^{-\alpha}. \tag{♣}$$

▷ Expectation

$$\mathbb{E}[X] = \int_0^\infty x\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-x\beta}\,\mathrm{d}x = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^\alpha\cdot e^{-x\beta}\,\mathrm{d}x.$$

Thus, by (♣), we get

$$\mathbb{E}[X] = \frac{\Gamma(\alpha+1)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+1}} = \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)\beta} = \frac{\alpha}{\beta}.$$

▷ Variance

$$\begin{aligned}\mathrm{Var}[X] = \mathbb{E}(X^2) - \left(\frac{\alpha}{\beta}\right)^2 &= \int_0^\infty x^2\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-x\beta}\,\mathrm{d}x - \left(\frac{\alpha}{\beta}\right)^2 \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+1}\cdot e^{-x\beta}\,\mathrm{d}x - \frac{\alpha^2}{\beta^2}.\end{aligned}$$

Thus,

$$\mathrm{Var}(X) \overset{(♣)}{=} \frac{\Gamma(\alpha+2)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+2}} - \frac{\alpha^2}{\beta^2} = \frac{\alpha(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}.$$

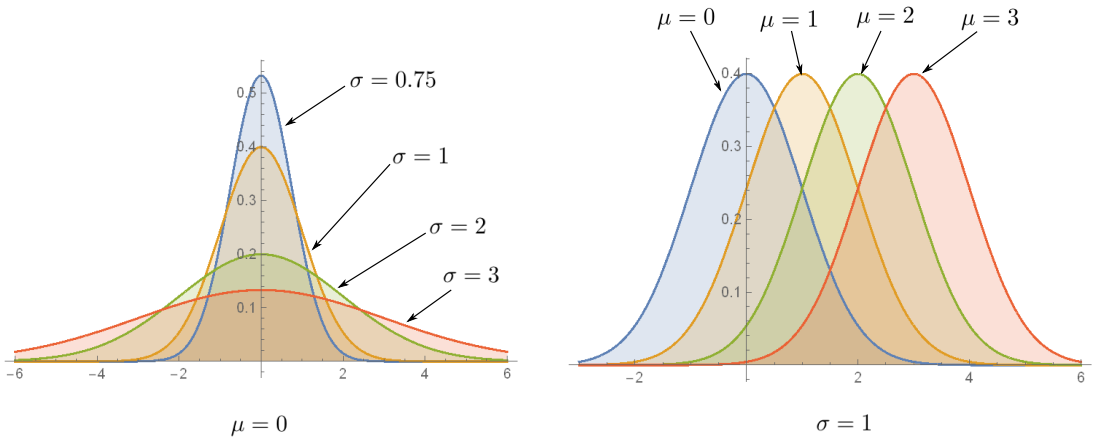


Figure 7.2: The probability density function of $\mathcal{N}(\mu, \sigma^2)$ for different values of μ and σ.

7.4 Normal (or Gaussian) distribution

Let $\mu \in \mathbb{R}$ and $\sigma > 0$.
A random variable X follows a **normal distribution** of parameters μ and σ^2 (abbreviated $X \sim \mathcal{N}(\mu, \sigma^2)$) if it has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

This density function gives a bell-shaped curve which is symmetric about μ. The larger the value of σ, the more spread-out the curve. As we will show below, μ represents the expectation and σ represents the square root of the variance (that is, the standard deviation). But first we need to check that f_X is indeed a pdf. The only thing to prove is that its integral over $\mathbb{R}$ equals 1.

Proof of $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

▷ First step: using the change of variables $y = \frac{x-\mu}{\sigma}$,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-y^2/2}\,\mathrm{d}y,$$

so it suffices to prove that

$$\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y = \sqrt{2\pi}.$$

▷ Second step:

$$\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right)^2 = \left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(y^2+z^2)/2}\,\mathrm{d}y\,\mathrm{d}z.$$

We now change to polar coordinates:

$$y = r\cos(\theta),\ z = r\sin(\theta) \quad \Longrightarrow \quad \mathrm{d}y\,\mathrm{d}z = r\mathrm{d}r\mathrm{d}\theta,\ y^2+z^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2.$$

so that the above integral becomes

$$\int_0^{\infty}\int_0^{2\pi} e^{-r^2/2}\, r\mathrm{d}r\mathrm{d}\theta = 2\pi\int_0^{\infty} re^{-r^2/2}\,\mathrm{d}r \overset{(\star)}{=} 2\pi\int_0^{\infty} e^{-u}\,\mathrm{d}u = 2\pi,$$

where the equality marked with $(\star)$ follows from the change of variable $u = r^2/2$. We have showed that

$$\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right)^2 = 2\pi,$$

as desired.

■

Before we turn to the expectation and variance of the normal distribution, we show the following.

Proposition 7.4.1 If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Proof. Letting $g(x) = ax + b$, we have $g^{-1}(y) = (y-b)/a$, so

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \tfrac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} = \frac{1}{a} \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}} = \underbrace{\frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-(a\mu+b))^2}{2\sigma^2 a^2}}}_{f_Y(y) \text{ for } Y \sim \mathcal{N}(a\mu+b, a^2\sigma^2)}.$$

■

Corollary 7.4.2 If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

$\mathcal{N}(0,1)$ is called the **standard normal** distribution.
We now turn to the expectation and variance of $\mathcal{N}(\mu, \sigma^2)$. Note that, if $Z \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\,\mathrm{d}x = 0$$

and

$$\mathbb{E}(Z^2) = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\,\mathrm{d}x = 1 \quad \text{(integrating by parts)}.$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then we know that $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, so

$$\mathbb{E}\left(\frac{X-\mu}{\sigma}\right) = 0 \Longrightarrow \frac{1}{\sigma}\,\mathbb{E}[X] - \frac{\mu}{\sigma} = 0 \Longrightarrow \mathbb{E}[X] = \mu,$$

$$\mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = 1 \Longrightarrow \frac{1}{\sigma^2}\mathrm{Var}[X] = 1 \Longrightarrow \mathrm{Var}[X] = \sigma^2.$$

7.5 Exercises

7.5.1 Problems

Exercise 7.1 Suppose the random variable T is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The **hazard function** $h_T(t)$ associated with the random variable T is defined by

$$h_T(t) = \lim_{\delta \to 0} \frac{\mathbb{P}(t \leq T < t + \delta \mid T \geq t)}{\delta}.$$

Thus, we can interpret $h_T(t)$ as the rate of change of the probability that the object survives a little past time t, given that the object survives to time t. Show that if T is a continuous random

variable, then

$$h_T(t) = \frac{f_T(t)}{1-F_T(t)} = -\tfrac{d}{dt}\log(1-F_T(t)).$$

■

Exercise 7.2 The **Pareto distribution** with parameters α and β has pdf

$$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty,\ \alpha > 0,\ \beta > 0.$$

(a) Verify that f is a pdf.
(b) Find the expectation and variance of this distribution.
(c) Prove that the variance does not exist if $\beta \leq 2$.

■

7.5.2 Are you up for a challenge?

Exercise 7.3 Assume that $\lambda > 0$ and, for each $n \in \mathbb{N}$, let $p_n = \frac{1}{\lambda n}$ and $X_n \sim \text{Geometric}(p_n)$. Show that

$$F_{X_n/n}(x) \xrightarrow{n\to\infty} F_Y(x), \qquad x > 0,$$

where $Y \sim \text{Exponential}(1/\lambda)$.

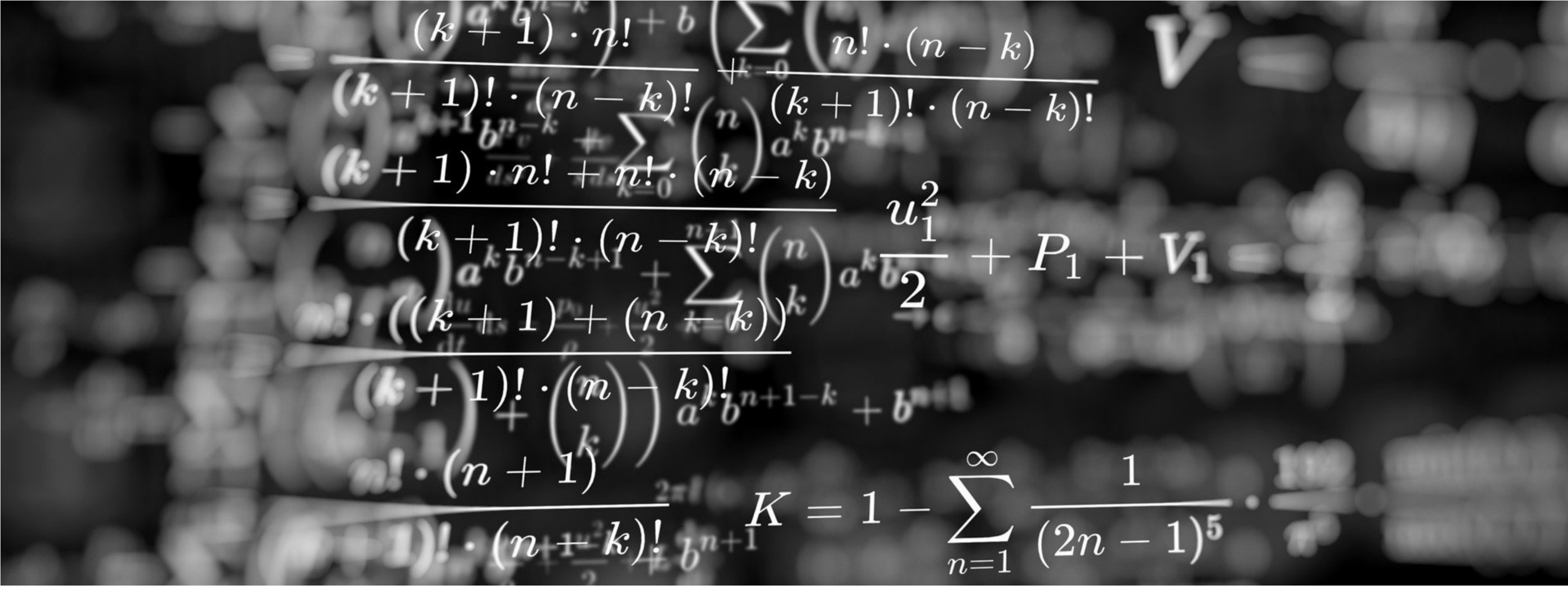
Hint. Use the fact that, for any $a \in \mathbb{R}$, $\lim_{x\to 0}\dfrac{\log(1+ax)}{x} = \tfrac{d}{dx}\log(1+ax)\big|_{x=0} = a.$ ■

Exercise 7.4 Show that

$$\int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1}e^{-z}\,dz = \sum_{y=0}^{\alpha-1}\frac{x^y}{y!}e^{-x}, \qquad \alpha = 1,2,3,\ldots.$$

Express this formula as a probabilistic relationship between Poisson and gamma random variables.

■



8. Random vectors

The goals of this chapters are:

- ▷ to define (discrete/continuous) **random vectors**;
- ▷ to define the **joint cumulative distribution function** (joint cdf) of a random vector;
- ▷ to define the **joint probability mass function** (joint pmf) of a discrete random vector;
- ▷ to define the **joint probability density function** (joint pdf) of a continuous random vector;

8.1 Joint and marginal distributions

Recall that we defined a random variable as a function from the sample space Ω into $\mathbb{R}$. We now have the similar notion:

Definition 8.1.1 Let $n \in \mathbb{N}$. An **n-dimensional random vector** is a function from a sample space Ω into $\mathbb{R}^n$.

■ **Example 8.1** Roll two dice. Let

$$X = \text{Absolute value of the difference between results,}$$
$$Y = \text{Maximum of two results.}$$

Then, the sample space is

$$\Omega = \left\{ \begin{array}{l} (⚀,⚀),(⚀,⚁),(⚀,⚂),(⚀,⚃),(⚀,⚄),(⚀,⚅) \\ (⚁,⚀),(⚁,⚁),(⚁,⚂),(⚁,⚃),(⚁,⚄),(⚁,⚅) \\ (⚂,⚀),(⚂,⚁),(⚂,⚂),(⚂,⚃),(⚂,⚄),(⚂,⚅) \\ (⚃,⚀),(⚃,⚁),(⚃,⚂),(⚃,⚃),(⚃,⚄),(⚃,⚅) \\ (⚄,⚀),(⚄,⚁),(⚄,⚂),(⚄,⚃),(⚄,⚄),(⚄,⚅) \\ (⚅,⚀),(⚅,⚁),(⚅,⚂),(⚅,⚃),(⚅,⚄),(⚅,⚅) \end{array} \right\},$$

and $(X,Y)\colon \Omega \to \mathbb{R}^2$ is a two-dimensional random vector.
Note that, obviously, X and Y taken individually are random variables. For $\omega = (⚂,⚁)$, we have $X(\omega) = 1$, $Y(\omega) = 3$ and $(X,Y)(\omega) = (1,3)$. ■

As for random variables, the **distribution of a random vector** is the function $A \mapsto \mathbb{P}((X_1,\dots,X_n) \in A)$ for any $A \subset \mathbb{R}^n$. Recall that the distribution of a random variable can be described by its cdf. This concept generalizes to higher dimension.

Definition 8.1.2 The **joint cumulative distribution** of the random vector $(X_1,\dots,X_n)$ is

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}(X_1 \leqslant x_1,\dots,\ X_n \leqslant x_n).$$

Similarly as we did with random variables, we will focus on two kinds of random vectors: discrete random vectors and continuous random vectors.

8.1.1 Discrete random vectors

Definition 8.1.3 A random vector $(X_1,\dots,X_n)$ is **discrete** if it can only attain countable many values. In that case, the function

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}((X_1,\dots,X_n) = (x_1,\dots,x_n))$$

is called the **joint probability mass function** of $(X_1,\dots,X_n)$.
(Sometimes we omit the word "joint" and simply say that $f_{X_1,\dots,X_n}(x_1,\dots,x_n)$ is the pmf of the random vector.)

Fact:

$$\mathbb{P}((X_1,\dots,X_n) \in A) = \sum_{(x_1,\dots,x_n)\in A} f_{X_1,\dots,X_n}(x_1,\dots,x_n). \qquad (\clubsuit)$$

In particular, with $A = \mathbb{R}^n$, we have

$$1 = \sum_{(x_1,\dots,x_n)} f_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

■ **Example 8.2** For the example 8.1 given above, the pmf $f_{X,Y}$ of (X,Y) is:

$x\backslash y$	1	2	3	4	5	6
0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
2	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
3	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
4	0	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$
5	0	0	0	0	0	$\frac{1}{18}$

We can do the "sanity check" that the above fact holds:

$$\sum_{(x,y)} f_{X,Y}(x,y) = 6\times\frac{1}{36} + 15\times\frac{1}{18} = \frac{6+30}{36} = 1.$$

■

Definition 8.1.4 Let $(X_1,\dots,X_n)$ be a *discrete* random vector with joint pmf $f_{X_1,\dots,X_n}$. The pmf's $f_{X_1},\dots,f_{X_n}$ of the (univariate) random variables $X_1,\dots,X_n$ are called the **marginal probability mass function**, or "marginal pmf's" or simply "marginals".

If one knows the values of the joint pmf of a discrete random vector, it is easy to get the marginal pmf's with the use of (♣) applied to the sets where one coordinate is fixed and all the others are free. For example, for the first marginal this gives the following expression.

Fact:

$$f_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2,\dots,x_n} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \sum_{x_2,\dots,x_n} f_{X_1,\dots,X_n}(x_1,\dots,x_n). \quad (8.1)$$

In particular for a two dimensional random vector:

Fact:

$$f_X(x) = \sum_y f_{X,Y}(x,y), \qquad (\diamondsuit)$$

$$f_Y(y) = \sum_x f_{X,Y}(x,y). \qquad (\spadesuit)$$

■ **Example 8.3** When we represent $f_{X,Y}$ in a table like in example 8.2, (♢) and (♠) correspond to summing over rows and columns (respectively). Consider the joint pmf $f_{x,y}$ of the previous example.

$x\backslash y$	1	2	3	4	5	6	$f_X(x)$
0	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{6}{36}$
1	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{5}{18}$
2	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{4}{18}$
3	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{3}{18}$
4	0	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{2}{18}$
5	0	0	0	0	0	$\frac{1}{18}$	$\frac{1}{18}$
$f_Y(y)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	

Did you notice that if you sum up the value of the last column (or of the last row) you get one? Why is that?[1] ■

The joint pmf cannot be obtained from the marginals (in general). Consider for instance the following two joint pmf's for a random vector (X,Y):

$x\backslash y$	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

$x\backslash y$	0	1
0	$\frac{1}{3}$	$\frac{1}{6}$
1	$\frac{1}{6}$	$\frac{1}{3}$

Both have the same marginals $f_X(0) = f_X(1) = f_Y(0) = f_Y(1) = \frac{1}{2}$.

8.1.2 Continuous random vectors

[1] f_X is the pmf of X, thus the sum of all the values of the last column is $\sum_x f_X(x) = 1$.

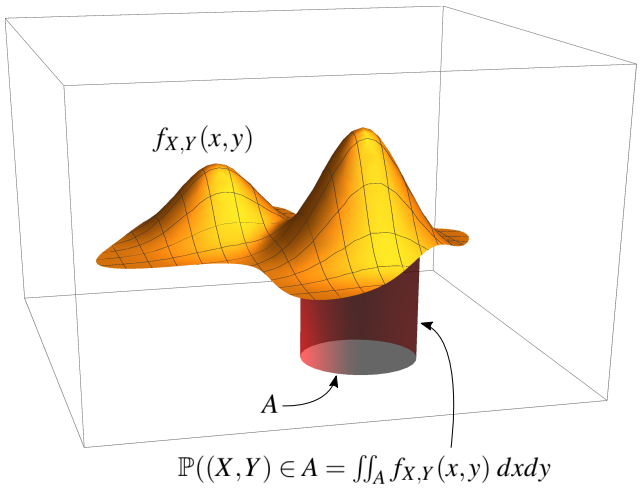
Definition 8.1.5 A random vector $(X_1,\dots,X_n)$ is **continuous** if there exists a function $f_{X_1,\dots,X_n}:\mathbb{R}^n\to[0,\infty)$ such that

$$\mathbb{P}((X_1,\dots,X_n)\in A)=\int\cdots\int_A f_{X_1,\dots,X_n}(x_1,\dots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

$f_{X_1,\dots,X_n}$ is called the **joint probability density function** of $(X_1,\dots,X_n)$.

Technical comment: If we would like to be completely coherent with the one dimensional terminology, we should say that $(X_1,\dots,X_n)$ is *absolutely* continuous if it has a joint pdf as in Definition 8.1.5. However in the literature it is common to omit the word "absolutely" for random vectors, and we will stick to this convention.

Like in the discrete case, from the joint pdf $f_{X_1,\dots,X_n}$ we can identify the marginal pdf's $f_{X_1},\dots,f_{X_n}$. In order to see how this is done, let us specialize to the 2-dimensional case. Let (X,Y) be a random vector. We have, for $A\subset\mathbb{R}^2$: In particular, taking $A=\mathbb{R}^2$,



$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y=1.$$

We claim that the marginal pdf's can be obtained through the formulas:

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y,\qquad f_Y(y)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$

Let us prove the first formula (the second follows by symmetry). We have

$$\begin{aligned}f_X(x)=\tfrac{\mathrm{d}}{\mathrm{d}x}F_X(x)=\tfrac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X\leqslant x)=\tfrac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X\leqslant x,\ -\infty<Y<\infty)&=\tfrac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(s,y)\,\mathrm{d}y\,\mathrm{d}s\\&=\int_{-\infty}^{\infty} f_{X,Y}(y)\,\mathrm{d}y,\end{aligned}$$

where in the last equality we have used the Fundamental Theorem of Calculus, $\frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} g(s)\,\mathrm{d}s=g(x)$.

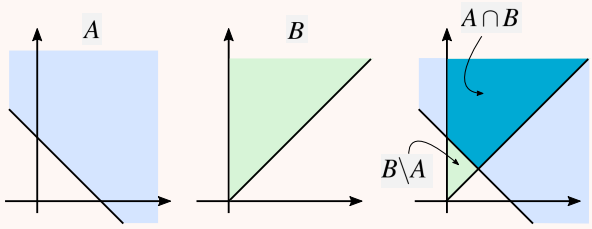
Similar formulas to obtain the marginals hold for higher dimensional random vectors.

Question 8.1 Let (X,Y) be a (continuous) random vector with joint pdf

$$f_{X,Y}(x,y) = \begin{cases} e^{-y}, & \text{if } 0 < x < y, \\ 0, & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(X+Y \geq 1)$.

Answer: Consider the regions:



We want $\iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$. B is the region where $f_{X,Y}(x,y) > 0$. We thus have

$$\begin{aligned} \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \iint_{A\cap B} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint_B f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y - \iint_{B\setminus A} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= 1 - \int_0^{1/2}\int_x^{1-x} e^{-y}\,\mathrm{d}y\,\mathrm{d}x = \frac{2}{\sqrt{e}} - \frac{1}{e}. \end{aligned}$$

■

Recall that the **joint cumulative distribution** of the random vector $(X_1,\ldots,X_n)$ is

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}(X_1 \leqslant x_1, \cdots, X_n \leqslant x_n).$$

In the continuous case, we have for continuous $f_{x_1,\ldots,x_n}$:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}}{\partial x_1 \cdots \partial x_n}(x_1,\ldots,x_n).$$

R Sometime one considers a mixture of discrete and random variables. It can be, for example, a random vector (X,Y) where X is a continuous random variable and Y a discrete random variable. In this case, one can show that there is a function $f_{X,Y}\colon \mathbb{R}^2 \to [0,\infty)$ which we call *joint probability density function* (as in the purely continuous setting) satisfying

$$\mathbb{P}((X,Y) \in A) = \int \sum_{y:(x,y)\in A} f_{X,Y}(x,y)\,\mathrm{d}x.$$

8.2 Expectation

For a random vector $(X_1,\ldots,X_n)\colon \Omega \to \mathbb{R}^n$ and a function $g\colon \mathbb{R}^n \to R$, we observe that $Y := g(X_1,\ldots,X_n)\colon \Omega \to \mathbb{R}$ is a random variable. The next theorem provides formulas to compute its expectation when the random vector is either discrete or continuous.

Theorem 8.2.1 — Expectation of a function of a random vector. If $(X_1,\dots,X_n)$ is a random vector and $g\colon \mathbb{R}^n \to \mathbb{R}$, then

$$\mathbb{E}(g(X_1,\dots,X_n)) = \begin{cases} \sum_{x_1}\cdots\sum_{x_n} g(x_1,\dots,x_n) f_{X_1,\dots,X_n}(x_1,\dots,x_n), & \text{(discrete case)},\\ \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\dots,x_n) f_{X_1,\dots,X_n}(x_1,\dots,x_n)\,\mathrm{d}x_1\dots\mathrm{d}x_n, & \text{(continuous case)}. \end{cases}$$

Proof. (Discrete setting) We proceed similarly as in the proof of Theorem 5.1.1 (linearity of the expectation).

$$\begin{aligned} \mathbb{E}g(X_1,\dots,X_n) &= \sum_z z\mathbb{P}(g(x_1,\dots,x_n)=z)\\ &= \sum_z z \sum_{x_1,\dots,x_n:g(x_1,\dots,x_n)=z} \mathbb{P}(X_1=x_1,\dots,X_n=x_n)\\ &= \sum_{x_1,\dots,x_n,z:g(x_1,\dots,x_n)=z} g(x_1,\dots,x_n) f_{X_1,\dots,X_n}(x_1,\dots,x_n)\\ &= \sum_{x_1,\dots,x_n} g(x_1,\dots,x_n) f_{X_1,\dots,X_n}(x_1,\dots,x_n) \end{aligned}$$

■

As a direct corollary we get the following theorem. Its first point is a repetition of Theorem 5.1.3, but it is so important that it does not hurt to repeat it here.

Theorem 8.2.2

1. If X and Y are random variables and $a,b\in\mathbb{R}$, then
$$\mathbb{E}[aX+bY] = a\mathbb{E}[X]+b\mathbb{E}[Y] \qquad \text{(linearity of expectation)}.$$
2. If $\mathbb{P}(X\geqslant Y)=1$, then $\mathbb{E}(X)\geqslant\mathbb{E}(Y)$. (monotonicity of expectation)

Proof.

1. Apply Theorem 8.2.1 to $g(X,Y)=aX+bY$ and use the linearity of sums (discrete case) or integrals (continuous case).
2. Taking $g(X,Y)=X-Y$, we have $g(X,Y)\geq 0$ with probability one (by assumption). Thus using the representation of $\mathbb{E}(g(X,Y))$ given by Theorem 8.2.1, we get that $\mathbb{E}(X-Y)\geq 0$, which implies the claim.

■

8.3 Exercises

8.3.1 Problems

Exercise 8.1 A random point (X,Y) is distributed uniformly on the square with vertices $(1,1)$, $(1,-1)$, $(-1,1)$, and $(-1,-1)$. That is, the joint pdf is $f(x,y)=\frac{1}{4}$ on the square. Determine the probabilities of the following events:

(a) $X^2+Y^2<1$

(b) $2X-Y>0$

(c) $|X+Y|<1$

■

Exercise 8.2 A pdf is defined by

$$f(x,y) = \begin{cases} C(x+2y) & \text{if } 0 < y < 1 \text{ and } 0 < x < 2; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of C.
(b) Find the marginal distribution of X.
(c) Find the joint cdf of X and Y.
(d) Find the pdf of the random variable $Z = 9/(X+1)^2$.

■

Exercise 8.3
(a) Find $\mathbb{P}(X > \sqrt{Y})$ if X and Y are jointly distributed with pdf

$$f(x,y) = x+y, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1.$$

(b) Find $\mathbb{P}(X^2 < Y < X)$ if X and Y are jointly distributed with pdf

$$f(x,y) = 2x, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 1.$$

■

Exercise 8.4 Assume X and Y are independent and positive random variables, both with the same pdf f. Let $Z = X/Y$. Show that

$$f_Z(z) = \int_0^\infty y \cdot f(zy) \cdot f(y)\, dy.$$

■

8.3.2 Are you up for a challenge?

Exercise 8.5 Let X, Y be two random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} cxy^2 & \text{if } 0 \leqslant x,y \leqslant 1; \\ 0 & \text{otherwise.} \end{cases}$$

Find c, $\mathbb{E}(\max(X,Y))$ and $\mathbb{E}(|X-Y|)$. ■

9. Conditioning and Independence

The goals of this chapters are:

- ▷ to define the **conditional pmf** of X given Y, where X and Y are *discrete* random variables;
- ▷ to define the **conditional pdf** of X given Y, where X and Y are *continuous* random variables;
- ▷ to define the **conditional expectation** and **conditional variance** of a random variable given an event;
- ▷ to define the **independence of random variable** and establish **criteria for independence**.

9.1 Conditional distribution

9.1.1 Conditional distribution in the discrete setting

Definition 9.1.1 Let (X,Y) be a discrete *random* vector with joint pmf $f_{X,Y}$ and marginals f_X and f_Y. The **conditional pmf of X given Y** is the function

$$f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x,\ Y = y)}{\mathbb{P}(Y = y)},$$

defined for all y such that $f_Y(y) > 0$ and all x.

■ **Example 9.1** We roll three dice. Let

$X =$ minimum of 3 results,
$Y =$ maximum of 3 results.

First let us represent $f_{X,Y}$ in a table.

$x\backslash y$	1	2	3	4	5	6	f_X
1	$\frac{1}{6^3}$	$\frac{1}{6^2}$	$\frac{2}{6^2}$	$\frac{3}{6^2}$	$\frac{4}{6^2}$	$\frac{5}{6^2}$	$\frac{91}{6^3}$
2	0	$\frac{1}{6^3}$	$\frac{1}{6^2}$	$\frac{2}{6^2}$	$\frac{3}{6^2}$	$\frac{4}{6^2}$	$\frac{61}{6^3}$
3	0	0	$\frac{1}{6^3}$	$\frac{1}{6^2}$	$\frac{2}{6^2}$	$\frac{3}{6^2}$	$\frac{37}{6^3}$
4	0	0	0	$\frac{1}{6^3}$	$\frac{1}{6^2}$	$\frac{2}{6^2}$	$\frac{19}{6^3}$
5	0	0	0	0	$\frac{1}{6^3}$	$\frac{1}{6^2}$	$\frac{7}{6^3}$
6	0	0	0	0	0	$\frac{1}{6^3}$	$\frac{1}{6^3}$
f_Y	$\frac{1}{6^3}$	$\frac{7}{6^3}$	$\frac{19}{6^3}$	$\frac{37}{6^3}$	$\frac{61}{6^3}$	$\frac{91}{6^3}$	

Now, in order to get $f_{X|Y}(x \mid y)$ for some fixed y, we *fix the column* corresponding to y. For example, for $y = 4$:

$x\backslash y$	4
1	$\frac{3}{6^2}$
2	$\frac{2}{6^2}$
3	$\frac{1}{6^2}$
4	$\frac{1}{6^3}$
5	0
6	0
f_Y	$\frac{37}{6^3}$

so $f_{X|Y}(x \mid y)$ is :

x	
1	$\frac{3/6^2}{37/6^3} = \frac{18}{37}$
2	$\frac{2/6^2}{37/6^3} = \frac{12}{37}$
3	$\frac{1/6^2}{37/6^3} = \frac{6}{37}$
4	$\frac{1/6^3}{37/6^3} = \frac{1}{37}$
5	$\frac{0}{37/6^3} = 0$
6	$\frac{0}{37/6^3} = \frac{0}{37}$

To get $f_{Y|X}(y \mid x)$, we *fix the row corresponding to* x and proceed similarly. ∎

9.1.2 Conditional distribution in the continuous setting

Definition 9.1.2 Let (X,Y) be a continuous random vector with joint density $f_{X,Y}(x,y)$. The **conditional probability density function** of X given Y is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

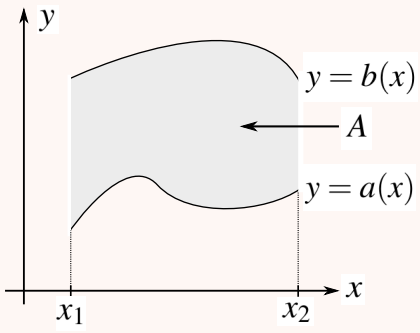
defined for all x and for all y such that $f_Y(y) > 0$.

R As a function of x for fixed y, $f_{X,Y}(x,y)$ is a pdf since

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = 1.$$

! Sometime people mixes things between the discrete and continuous setting. Recall that in the discrete setting, we have $f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X=x,\, Y=y)}{\mathbb{P}(Y=y)}$. In the continuous setting, this is *meaningless*, since both the numerator and denominator are zero!

Note: if $A \subset \mathbb{R}^2$ is a set of the form $\{(x,y) \in \mathbb{R}^2 : a(x) \leq y \leq b(x)\}$



then,

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{x_1}^{x_2} \int_{a(x)}^{b(x)} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x$$
$$= \int_{x_1}^{x_2} f_X(x) \int_{a(x)}^{b(x)} f_{Y|X}(y \mid x)\,\mathrm{d}y\,\mathrm{d}x.$$

R The definition of conditional probability mass/density function extends to random vectors. Let $(X_1,\ldots,X_n)$ be a discrete/continuous random vector. The **conditional probability mass/density function** of $X_1,\ldots,X_m$ given $X_{m+1},\ldots,X_n$ is

$$f_{X_1,\ldots,X_m \mid X_{m+1},\ldots,X_n}(x_1,\ldots,x_m \mid x_{m+1},\ldots,x_n) := \frac{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{f_{X_{m+1},\ldots,X_n}(x_{m+1},\ldots,x_n)},$$

defined for all $x_1,\ldots,x_m$ and for all $x_{m+1},\ldots,x_n$ such that $f_{X_{m+1},\ldots,X_n}(x_{m+1},\ldots,x_n) > 0$.

9.2 Independence of random variables

Definition 9.2.1 Random variables $X_1,\ldots,X_n$ (defined on the same probability space) are **independent** if

$$\mathbb{P}(X_1 \in A_1,\ldots,X_n \in A_n) = \mathbb{P}(X_1 \in A_1)\cdots\mathbb{P}(X_n \in A_n) \qquad \text{for all } A_1,\ldots A_n \subset \mathbb{R}.$$

Note: For $n = 2$, the equality in Definition 9.2.1 could be written

$$\mathbb{P}(X_1 \in A_1 \mid X_2 \in A_2) = \mathbb{P}(X_1 \in A_1) \qquad \text{for all } A_1,A_2 \subset \mathbb{R} \text{ with } \mathbb{P}(A_2) > 0,$$

or equivalently,

$$\mathbb{P}(X_2 \in A_2 \mid X_1 \in A_1) = \mathbb{P}(X_2 \in A_2) \qquad \text{for all } A_1,A_2 \subset \mathbb{R} \text{ with } \mathbb{P}(A_1) > 0.$$

In other words $X_1,\ldots,X_n$ are independent if the events $\{X_1 \in A_1\},\ldots,\{X_n \in A_n\}$ are (mutually) independent for all collections of sets $A_1,\ldots,A_n \subset \mathbb{R}$. Checking this for all choices of sets might be cumbersome. Luckily the following two lemmas give us simple criteria to prove (or disprove) independence of random variables.

Proposition 9.2.1 — Independence criterion for random variables 1. Random variables $X_1, \dots, X_n$ are independent if and only if

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n) \qquad (\clubsuit)$$

(both for discrete and continuous).

Proof. We prove the proposition only in the continuous setting and in dimension $n = 2$. The proof in the discrete setting, is similar, replacing integrals by sums. Extending the proof to higher dimension requires only to write longer expressions (n-th derivative instead of 2-nd derivative, n-fold integral instead of 2-fold integral,...).

Assume that X and Y are independent. Then,

$$F_{X,Y}(x,y) = \mathbb{P}(X \leqslant x,\, Y \leqslant y) = \mathbb{P}(X \leqslant x)\cdot\mathbb{P}(Y \leqslant y) = F_X(x)\cdot F_Y(y),$$

so

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x\,\partial y}(x,y) = \frac{\partial}{\partial x}\left(F_X(x)\cdot f_Y(y)\right) = f_X(x)\cdot f_Y(y).$$

Now for the converse, assume ($\clubsuit$) holds. Then,

$$\mathbb{P}(X\in A,\, Y\in B) = \int_A\int_B f_X(x)\cdot f_Y(y)\,\mathrm{d}y\,\mathrm{d}x = \left(\int_A f_X(x)\,\mathrm{d}x\right)\left(\int_B f_Y(y)\,\mathrm{d}y\right) = \mathbb{P}(X\in A)\cdot\mathbb{P}(Y\in B).$$

■

Note: Note that, for a random vector (X,Y), if $f_Y(y) > 0$ or $f_X(x) > 0$, Equation ($\clubsuit$) can be rewritten as

$$f_{X|Y}(x\,|\,y) = f_X(x) \qquad \text{or} \qquad f_{Y|X}(y\,|\,x) = f_Y(y),$$

which can be interpreted as "being told that $Y = y$ has no effect on the distribution of X" in the first case, and "being told that $X = x$ has no effect on the distribution of Y" in the second.

Question 9.1 Assume the joint pdf of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} 1, & \text{if } 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 2-2x, \\ 0, & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Answer: In order to determine the answer, we can check the definition, finding f_X, f_Y and inspecting if $f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y)$. We compute the marginals

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^{2-2x}\mathrm{d}y = 2-2x, \quad 0<x<1,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \int_0^{1-\frac{y}{2}}\mathrm{d}x = 1-\frac{y}{2}, \quad 0<y<2,$$

and their product

$$f_X(x)\cdot f_Y(y)=\begin{cases}(2-2x)\cdot(1-y/2), & \text{if } 0<x<1,\ 0<y<2,\\ 0, & \text{otherwise.}\end{cases}$$

Since this is not equal to $f_{X,Y}(x,y)$, X and Y are **not** independent. ■

Proposition 9.2.2 — Independence criterion for random variables 2. Assume that there exist non-negative functions g, h such that we can write

$$f_{X,Y}(x,y)=g(x)\cdot h(y).$$

Then, X and Y are independent and

$$f_X(x)=\frac{g(x)}{\int_{-\infty}^{\infty}g(s)\,\mathrm{d}s},\qquad f_Y(y)=\frac{h(y)}{\int_{-\infty}^{\infty}h(t)\,\mathrm{d}t}.$$

Proof. We start noting that

$$1=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x)\cdot h(y)\,\mathrm{d}x\,\mathrm{d}y=\left(\int_{-\infty}^{\infty}g(x)\,\mathrm{d}x\right)\cdot\left(\int_{-\infty}^{\infty}h(y)\,\mathrm{d}y\right).\quad(\spadesuit)$$

Next,

$$f_X(x)=\int_{-\infty}^{\infty}f_{X,Y}(x,t)\,\mathrm{d}t=\int_{-\infty}^{\infty}g(x)\cdot h(t)\,\mathrm{d}t=g(x)\int_{-\infty}^{\infty}h(t)\,\mathrm{d}t,$$

and similarly,

$$f_Y(y)=h(y)\int_{-\infty}^{\infty}g(s)\,\mathrm{d}s.$$

Multiplying, we get

$$f_X(x)f_Y(y)=g(x)h(y)\left(\int_{-\infty}^{\infty}g(s)\,\mathrm{d}s\right)\left(\int_{-\infty}^{\infty}h(t)\,\mathrm{d}t\right)\overset{(\spadesuit)}{=}g(x)h(y)=f_{X,Y}(x,y).$$

■

Question 9.2 Assume the joint pdf of X and Y is

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{384}x^2y^4e^{-y-x/2}, & \text{if } x>0,\ y>0,\\ 0, & \text{otherwise.}\end{cases}$$

Are X and Y independent?

Answer: We can decompose $f_{X,Y}(x,y)=g(x)h(y)$ with

$$g(x)=\begin{cases}\frac{1}{384}x^2e^{-x/2}, & \text{if } x>0,\\ 0, & \text{otherwise,}\end{cases}\qquad\text{and}\qquad h(y)=\begin{cases}y^4e^{-y}, & \text{if } y>0,\\ 0, & \text{otherwise,}\end{cases}$$

so X and Y are independent. ■

9.3 Conditional expectation and conditional variance

Definition 9.3.1 Let X be a random variable and E be an event with $\mathbb{P}(E) > 0$.
The **conditional expectation** X given E is

$$\mathbb{E}[X \mid E] := \frac{\mathbb{E}[X \mathbb{1}_E]}{\mathbb{P}(E)},$$

and the **conditional variance** of X given E is

$$\mathsf{Var}[X \mid E] := \mathbb{E}[X^2 \mid E] - (\mathbb{E}[X \mid E])^2.$$

In the *discrete setting*, one can condition on a second random variable Y taking a specific value y. In that case we write $\mathbb{E}[X \mid Y = y]$ instead of $\mathbb{E}[X \mid \{Y = y\}]$ (we omit the curly brackets). The **conditional expectation of** X **given that** $Y = y$ has then the representation

$$\mathbb{E}[X \mid Y = y] := \begin{cases} \sum_x x f_{X|Y}(x \mid y), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x, & \text{if } X \text{ is continuous} \end{cases}$$

Moreover, for a function $g\colon \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}(g(x) \mid Y = y) = \begin{cases} \sum_x g(x) f_{X|Y}(x \mid y), \\ \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\,\mathrm{d}x. \end{cases}$$

Similarly we write $\mathrm{Var}[X \mid Y = y]$ instead of $\mathrm{Var}[X \mid \{Y = y\}]$. The conditional variance of X given that $Y = y$ can thus be written

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}(X^2 \mid Y = y) - (\mathbb{E}(X \mid Y = y))^2.$$

Question 9.3 Joe and Moe will meet at the bus stop. Assume:

$$\begin{aligned} &\text{Joe's arrival time } = T_1 \sim \mathrm{ContUnif}(0,1), \\ &\text{Moe's arrival time } = T_2 \sim \mathrm{ContUnif}(0,1), \\ &T_1,\ T_2 \text{ are independent.} \end{aligned}$$

Find the expected amount of time that the first to arrive has to wait for the second.

Answer:

$$f_{T_1,T_2}(t_1,t_2) = f_{T_1}(t_1) f_{T_2}(t_2) = \begin{cases} 1, & \text{if } 0 \leqslant t_1, t_2 \leqslant 1, \\ 0, & \text{otherwise.} \end{cases}$$

Waiting time = $g(t_1,t_2) = |t_1 - t_2|$. Then,

$$\mathbb{E}(g(T_1,T_2)) = \int_0^1 \int_0^1 |t_1 - t_2|\,\mathrm{d}t_1\,\mathrm{d}t_2 = 2\int_0^1 \int_0^{t_1} (t_1 - t_2)\,\mathrm{d}t_2\,\mathrm{d}t_1 = \frac{1}{3}.$$

■

Proposition 9.3.1 If X and Y are independent, then for any g, h we have

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)).$$

In particular, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ and $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

Continuous case.

$$\begin{aligned}\mathbb{E}(g(X)h(Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\cdot h(y)\cdot f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int_{-\infty}^{\infty} g(x)\cdot f_X(x)\,\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} h(y)\cdot f_Y(y)\,\mathrm{d}y\right) \\ &= \mathbb{E}(g(X))\mathbb{E}(h(Y)).\end{aligned}$$

The discrete case is treated similarly (sums instead of integrals).

For the final assertion,

$$\begin{aligned}\mathsf{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - ((\mathbb{E}[X])^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + (\mathbb{E}[Y])^2) \\ &= \mathsf{Var}(X) + \mathsf{Var}(Y).\end{aligned}$$

■

9.4 Exercises

9.4.1 Check-up the basics

Exercise 9.1 — Decide whether or not each statement is necessarily true.

1. If X and Y are independent discrete random variables, then $f_{X|Y}(x \mid y) = f_X(x)$.
2. We repeatedly toss a fair coin. Let U be the number of trials needed to get the first head and V be the number of trials needed to get two successive heads. Then, U and V are independent random variables.
3. For a continuous random vector (X,Y) and real numbers a,b,c,d with $a<b$ and $c<d$,
$$\mathbb{P}((X,Y)\in[a,b]\times[c,d]) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c).$$
4. For a continuous random vector (X,Y) and real numbers a,b,c,d with $a<b$ and $c<d$,
$$\mathbb{P}((X,Y)\in[a,b]\times[c,d]) = \int_a^b\int_c^d f_{X|Y}(x \mid y)\cdot f_{X,Y}(x,y)\,dy\,dx.$$

■

Exercise 9.2 — Decide whether or not each statement is necessarily true.

1. If X and Y are independent with $f_X = f_Y$, then $f_{X,Y}(x,y) = f_X(x)^2$.
2. If X and Y are independent and continuous with $f_X = f_Y$, then $\mathbb{P}(X<Y) = \frac{1}{2}$.
3. If X and Y are independent and discrete with $f_X = f_Y$, then $\mathbb{P}(X<Y) = \frac{1}{2}$.
4. A random variable that is equal to a constant is independent of any other random variable.

■

9.4.2 Problems

Exercise 9.3 The random pair (X,Y) has the distribution

$y\backslash x$	1	2	3
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{6}$	0	$\frac{1}{6}$
4	0	$\frac{1}{3}$	0

Show that X and Y are dependent. Give a probability table for random variables U and V that have the same marginals as X and Y but are independent. ■

Exercise 9.4 John and Mary decide to meet at the train station. The times of their arrivals are independent, each uniformly distributed between noon and 1pm, and whoever arrives first waits for the other. Find the expected amount of time that the first to arrive has to wait. ■

Exercise 9.5 Assume $X_1, \ldots, X_n$ are independent random variables, all with the same cumulative distribution function F.

1. Find the cumulative distribution function of $Z = \max(X_1, \ldots, X_n)$ in terms of F.
2. Find the cumulative distribution function of $Y = \min(X_1, \ldots, X_n)$ in terms of F.

■

Exercise 9.6 In a group of 100 people, assume each has their birthday uniformly distributed over the 365 days of the year and that birthdays are independent.

1. Find the expected number of days of the year in which exactly three people have their birthdays.
2. Find the expected number of sets of 3 people with coinciding birthdays that we can form.

■

Exercise 9.7 Let U, V be independent random variables following a continuous uniform distribution on $(0,1)$.

(a) Let $X = U \cdot V$. Find f_X and $f_{V|X}$.

(b) Find the joint probability density function of W and Z, where $W = U + V$ and $Z = U - V$. Are W and Z independent?

■

Exercise 9.8 Show that, if X and Y are jointly continuous random variables,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \mathbb{E}(X|Y=y) \cdot f_Y(y)\, \mathrm{d}y. \tag{9.1}$$

and, if X and Y are discrete,

$$\mathbb{E}(X) = \sum_{y} \mathbb{E}(X|Y=y) \cdot f_Y(y). \tag{9.2}$$

■

Exercise 9.9 Let $X_1, \ldots, X_n, N$ be independent random variables with the following distributions:

$$N \sim \text{DiscreteUniform}\{1, \ldots, n\}, \qquad X_i \sim \text{Bernoulli}\left(\frac{1}{n-i+1}\right),\ i = 1, \ldots, n.$$

Define $W = \sum_{i=1}^{N} X_i$. Find $\mathbb{E}(W)$. ■

Exercise 9.10 Assume $Y, X_1, X_2, \ldots$ are independent and discrete, Y takes values in $\{0, 1, 2, \ldots\}$

and $X_1, X_2, \ldots$ are identically distributed. Let

$$Z = \begin{cases} 0 & \text{if } Y = 0; \\ X_1 + \cdots + X_Y & \text{if } Y > 0. \end{cases}$$

Show that

$$\mathbb{E}(Z) = \mathbb{E}(Y) \cdot \mathbb{E}(X_1).$$

■

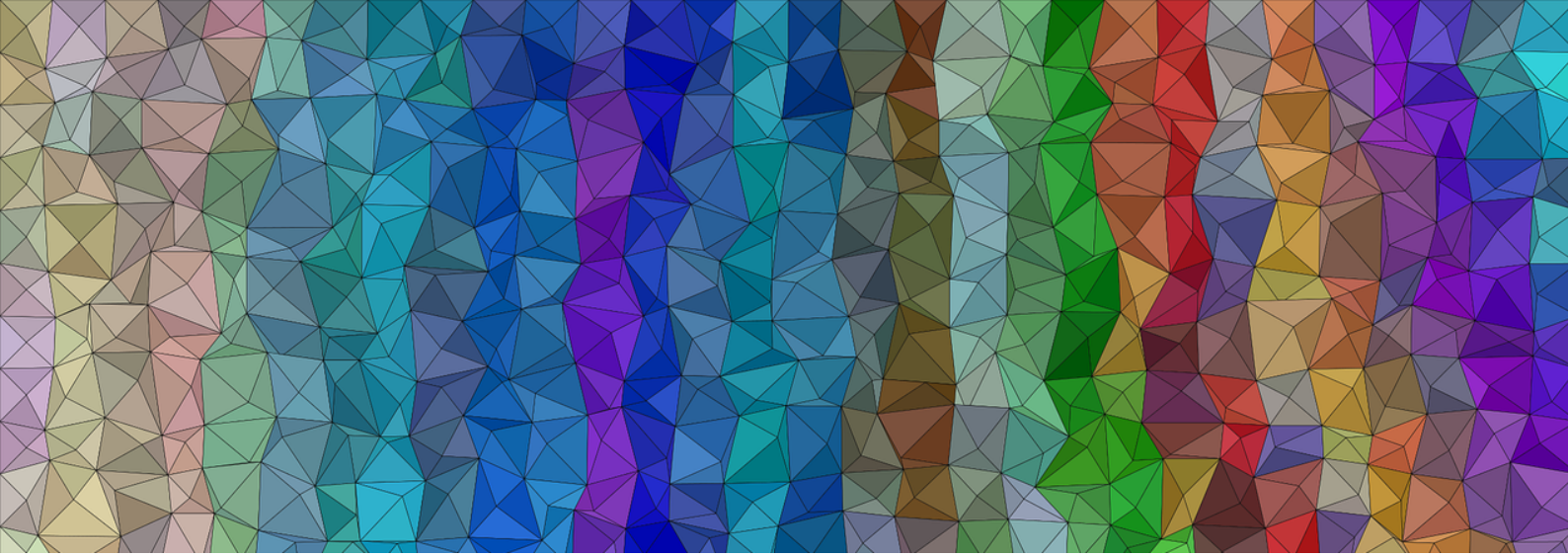
9.4.3 Are you up for a challenge?

Exercise 9.11 Suppose that the joint probability mass function of X and Y is

$$\mathbb{P}(X = i,\, Y = j) = \binom{j}{i} e^{-2\lambda} \frac{\lambda^j}{j!}, \qquad 0 \le i \le j.$$

(a) Find the probability mass function of Y.
(b) Find the probability mass function of X.
(c) Find the probability mass function of $Y - X$.

■



10. Transformation of vectors and correlation

The goals of this chapters are:

- ▷ to introduce tools to find the distribution of the **transformation of a random vector**;
- ▷ to derive the **convolution formula**;
- ▷ to define the **covariance** and **correlation**;
- ▷ to derive **properties of the covariance**.

10.1 Transformations of random vectors

Theorem 10.1.1 Assume X, Y are independent, $U = g_1(X)$ and $V = g_2(Y)$, where $g_1, g_2 \colon \mathbb{R} \to \mathbb{R}$. Then, U and V are independent.

Proof. Let $A, B \subset \mathbb{R}$ be arbitrary. We need to show that the events $E_1 = \{U \in A\}$ and $E_2 = \{V \in B\}$ are independent. We rewrite these event as $E_1 = \{X \in A'\}$ and $E_2 = \{Y \in B'\}$ with $A' = g^{-1}(A)$ and $B' = g^{-1}(B)$. Now, the statement follows directly from the independence of X and Y. ■

Given a random vector (X_1, X_2) and a function $g \colon \mathbb{R}^n \to \mathbb{R}^m$, we will now study the distribution (pmf or pdf) of $g(X_1, X_2)$.

Note: (case $m = n = 2$, but you can extend this comment to any m and n) Since g maps vectors of $\mathbb{R}^2$ into vectors of $\mathbb{R}^2$, we can write

$$g(x_1, x_2) = (g_1(x_1, x_2), g_2((x_1, x_2)),$$

where for each $i \in \{1, 2\}$, $g_i \colon \mathbb{R}^2 \to \mathbb{R}$ is a *coordinate function* of g.
Then, if $(Y_1, Y_2) = g(X_1, X_2)$, we have

$$Y_i = g_i(X_1, X_2), \qquad 1 \leq i \leq 2.$$

Discrete case

Recall that if X is discrete, $g\colon \mathbb{R} \to \mathbb{R}$ and $Y = g(X)$, then

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x)$$

We now have the exact same formula. If (X_1, X_2) is discrete, $g\colon \mathbb{R}^2 \to \mathbb{R}^2$ and $(Y_1, Y_2) = g(X_1, X_2)$, then

$$f_{Y_1,Y_2}(y_1, y_2) = \sum_{\substack{(x_1,x_2)\in \\ g^{-1}(y_1,y_2)}} f_{X_1,X_2}(x_1, x_2).$$

In particular, if g is one-to-one, then

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(g^{-1}(y_1, y_2)).$$

Continuous case

Recall, from Theorem 4.4.2, that if X is continuous, $g\colon \mathbb{R} \to \mathbb{R}$ is differentiable and increasing or decreasing, then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|.$$

We now have:

Theorem 10.1.2 Let (X_1, X_2) be a continuous random vector and let $g\colon D \to R$ one-to-one and differentiable, where

$$D := \{(x_1, x_2) \in \mathbb{R}^2 : f_{X_1,X_2}(x_1, x_2) \neq 0\}, \quad R = g(D) \subset \mathbb{R}^2.$$

If $h = g^{-1}$ and $(Y_1, Y_2) = g(X_1, X_2)$, then

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}(g^{-1}(y_1, y_2)) \cdot |J(y_1, y_2)| & \text{if } (y_1, y_2) \in R, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$J(y_1, y_2) = \det \begin{pmatrix} \partial h_1/\partial y_1 & \partial h_1/\partial y_2 \\ \partial h_2/\partial y_1 & \partial h_2/\partial y_2 \end{pmatrix}.$$

We omit the proof. The coefficient $J(y_1, y_2)$ is called the *Jacobian determinant* of the function h.

R The previous theorem extends to higher-dimensional random vectors: if $(X_1, \ldots, X_n)$ is a continuous random vector, $g(x_1, \ldots, x_n)$ is one-to-one, $h = g^{-1}$ and $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$, then

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}(g^{-1}(x_1, \ldots, x_n)) \cdot |J(y_1, \ldots, y_n)|,$$

where

$$J(y_1, \ldots, y_n) = \det \left(\left(\frac{\partial h_i}{\partial y_j} \right)_{i,j=1,\ldots,n} \right).$$

Question 10.1 Assume $X_1 \sim \text{Exp}(\beta_1)$ and $X_2 \sim \text{Exp}(\beta_2)$ are independent, $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Find f_{Y_1,Y_2}.

Answer: Let $g(x_1,x_2) = (x_1+x_2, x_1-x_2)$. In this type of problem, we should always start determining

- ▷ the domain D where the random vector (X_1,X_2) takes its values, and
- ▷ the range $R = \{g(x_1,x_2) : (x_1,x_2) \in D\}$.

Since X_1 and X_2 are exponentially distributed and independent, we have $f_{X_1,X_2}(x_1,x_2) > 0$ if and only if $f_{X_1}(x_1) > 0$ and $f_{X_2}(x_2) > 0$, that is, if $x_1, x_2 > 0$, so

$$D = \{(x_1,x_2) \in \mathbb{R}^2 : x_1, x_2 > 0\}.$$

We now find R. In this case, a convenient way to find it is to note that the image under g of horizontal lines

$$\{(s, y_0) : s > 0\}$$

are diagonal lines

$$\{(s+y_0, s-y_0) : s > 0\} = \{(y_0, -y_0) + s(1,1) : s > 0\}.$$

By running over all possible values of y_0, we see that

$$R = \{(x,y) : -x < y < x\}.$$

Now that we have found D and R so that $g\colon D \to R$, we find $h = g^{-1}$. It is given by

$$h(y_1,y_2) = (\tfrac{1}{2}(y_1+y_2), \tfrac{1}{2}(y_1-y_2)).$$

Then,

$$|J(y_1,y_2)| = \left|\det \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}\right| = \frac{1}{2}.$$

Finally, for $(y_1,y_2) \in R$,

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= f_{X_1,X_2}(h(y_1,y_2)) \cdot \frac{1}{2} \\ &= f_{X_1}(h_1(y_1,y_2)) \cdot f_{X_2}(h_2(y_1,y_2)) \cdot -\frac{1}{2} \\ &= \frac{1}{2} \cdot \frac{1}{\beta_1} \cdot e^{-h_1(y_1,y_2)\beta_1} \cdot \frac{1}{\beta_2} \cdot e^{-h_2(y_1,y_2)\beta_2} \\ &= \frac{1}{2\beta_1\beta_2} e^{-\frac{(y_1+y_2)\beta_1}{2} - \frac{(y_1-y_2)\beta_2}{2}}. \end{aligned}$$

■

An application: convolutions

If X, Y are random variables then their *convolution* is (the probability distribution of) $X+Y$.

Question 10.2 Let $X_1 \sim \text{Binomial}(n,p)$ and $X_2 \sim \text{Poisson}(\lambda)$ be independent random variables. Find the pmf of $Y := X_1 + X_2$.

Answer: Let $g\colon \mathbb{R}^2 \to \mathbb{R}$, defined by $g(x_1,x_2) = x_1 + x_2$, so that $Y = g(X_1,X_2)$. What values can Y attain? $X_1 \in \{0,\dots,n\}$ and $X_2 \in \{0,1,2,\dots\}$, so $X_1 + X_2 \in \{0,1,2,\dots\}$.

$$f_Y(y) = \sum_{(x_1,x_2)\in g^{-1}(y)} f_{X_1,X_2}(x_1,x_2) = \underbrace{\sum_{(x_1,x_2):x_1+x_2=y}}_{\text{let's take a closer look at this}} f_{X_1}(x_1)\cdot f_{X_2}(x_2).$$

If $y \geq n$, the following are the possible values of x_1, x_2 so that $x_1 + x_2 = y$:

$$\begin{aligned} &x_1 = 0,\ x_2 = y,\\ &x_1 = 1,\ x_2 = y-1,\\ &\dots\\ &x_1 = n,\ x_2 = y-n. \end{aligned}$$

If however $y \in \{0,\dots,n-1\}$, the list of possibilities is:

$$\begin{aligned} &x_1 = 0,\ x_2 = y,\\ &x_1 = 1,\ x_2 = y-1,\\ &\dots\\ &x_1 = y,\ x_2 = 0. \end{aligned}$$

Final answer:

$$f_Y(y) = \begin{cases} \sum_{x_1=0}^{n} \binom{n}{x_1} p^{x_1}(1-p)^{n-x_1} \frac{\lambda^{y-x_1}}{(y-x_1)!}\, e^{-\lambda}, & \text{if } y \in \{n, n+1, \dots\},\\ \sum_{x_1=0}^{y} \binom{n}{x_1} p^{x_1}(1-p)^{n-x_1} \frac{\lambda^{y-x_1}}{(y-x_1)!}\, e^{-\lambda}, & \text{if } y \in \{0,\dots,n\}. \end{cases}$$

■

The example above illustrate that, in general, the sum of two random variables X and Y can have a distribution much more intricate that the distributions of X and Y. However, in some instances sums behave very nicely, as we will see in the next theorem.

Theorem 10.1.3 — Sum of independent Poisson random variables. Let $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ be *Poisson* random variables. If they are *independent*, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Proof.

$$\begin{aligned} f_{X_1+X_2}(k) &= \sum_{x_1=0}^{k} f_{X_1}(x_1)\cdot f_{X_2}(k-x_1)\\ &= \sum_{x_1=0}^{k} \frac{(\lambda_1)^{x_1}}{(x_1)!}\cdot e^{-\lambda_1}\cdot \frac{(\lambda_2)^{k-x_1}}{(k-x_1)!}\cdot e^{-\lambda_2}\\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{x_1=0}^{k} \frac{k!}{(x_1)!(k-x_1)!}(\lambda_1)^{x_1}(\lambda_2)^{k-x_1}\\ &= \frac{(\lambda_1+\lambda_2)^k}{k!}\cdot e^{-(\lambda_1+\lambda_2)} \end{aligned}$$

as required. ■

Theorem 10.1.4 — Convolution formula. Suppose that X and Y are independent random variables.

1. If X and Y are discrete, then $Z = X+Y$ is discrete and its pmf is
$$f_{X+Y}(z) = \sum_x f_X(x) f_Y(z-x).$$
(Note that this sum has countably many non-zero summands because X is discrete.)
2. If X and Y are continuous, then $Z = X+Y$ is continuous and its pdf is
$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x.$$

In concrete cases one can sometimes massage the obtained expression into some nice closed form. One example is when X and Y are Poisson distributed, see Theorem 10.1.3.

Proof. **Discrete case.**

$$f_{X+Y}(z) = \mathbb{P}(X+Y=z) = \sum_x \mathbb{P}(X=x, Y=z-x) = \sum_x f_X(x) f_Y(z-x).$$

Continuous case. We apply the "transformation theorem" (Theorem 10.1.2) to the pair $(X_1, Z) = (X_1, X_1+X_2) = g(X_1, X_2)$. Using the independence of X_1, X_2 and that theorem, we have

$$f_{X_1,Z}(x_1,z) = f_{X_1}(x_1) f_{X_2}(z-x_1) \left| \det \left(\frac{\partial g_i^{-1}}{\partial y_j} \right)_{i,j} \right|.$$

The derivative matrix of g is:

$$\left(\frac{\partial g_i}{\partial x_j} \right)_{i,j} = \begin{pmatrix} \partial g_1/\partial x_1 & \partial g_1/\partial x_2 \\ \partial g_2/\partial x_1 & \partial g_2/\partial x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

This has determinant one, so the derivative matrix of g^{-1} also has determinant one. Hence $f_{X_1,Z}(x_1,z) = f_{X_1}(x_1) f_{X_2}(z-x_1)$. As we've seen before the marginal pdf f_Z of Z can be obtained from the joint pdf $f_{X_1,Z}$ by "integrating out" X_1:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(z-x_1)\,dx_1,$$

which is the sought expression. ■

By the way, the word *convolution* comes from the Latin *convolvere*, which means "to roll together".

10.2 Covariance and correlation

These are two measures of the strength of the relation between two random variables.

Definition 10.2.1 Let X and Y be random variables. Set $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$.

Covariance between X and Y: $\quad \mathrm{Cov}(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)),$

Correlation between X and Y: $\quad \rho_{X,Y} = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$ if $\sigma_X \sigma_Y > 0$.

If $\mathrm{Cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.

Remark: If $\sigma_X \sigma_Y = 0$ than $\rho_{X,Y}$ is undefined.

Theorem 10.2.1 — Properties of the covariance.

1. $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$,
2. $\mathsf{Cov}(X,X) = \mathrm{Var}(X)$,
3. $\mathsf{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y)$,
4. If X and Y are independent, then they are uncorrelated. **The converse is not true in general.**

Proof.

1. is evident.
2. follows from $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.
3. $$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right) = \mathbb{E}(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)\\ &= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mu_X\mu_Y\\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$
4. If X and Y are independent, then $\mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y) = \mathbb{E}(X)\cdot\mathbb{E}(Y) - \mathbb{E}(X)\cdot\mathbb{E}(Y) = 0$.

■

Lemma 10.2.2 For any random variable X,

$$\mathbb{P}(X=0) = 1 \quad \text{if and only if } \mathbb{E}[X^2] = 0.$$

Proof. If $\mathbb{P}(X=0) < 1$, then $\mathbb{P}(X \neq 0) > 0$. Note:

$$\{|X| > 1\} \subseteq \left\{|X| > \frac{1}{2}\right\} \subseteq \left\{|X| > \frac{1}{3}\right\} \subseteq \cdots$$

and $\bigcup_{n\geqslant 1}\left\{|X| > \frac{1}{n}\right\} = \{X \neq 0\}$, so $\lim_{n\to\infty}\mathbb{P}\left(|X| > \frac{1}{n}\right) = \mathbb{P}(X\neq 0) > 0$, so there exists n such that $\mathbb{P}(|X| > \frac{1}{n}) > 0$. Now,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x \geqslant \int_{\{x:|x|>1/n\}} (1/n)^2 f_X(x)\,\mathrm{d}x \geqslant \frac{1}{n^2}\cdot\mathbb{P}\left(|X| > \frac{1}{n}\right) > 0,$$

and similarly for the discrete case. ■

Theorem 10.2.3 — More Properties of the covariance.

1. $\mathsf{Cov}(aX+bY,Z) = a\mathsf{Cov}(X,Z) + b\mathsf{Cov}(Y,Z)$
2. If either of X or Y is constant, then $\mathsf{Cov}(X,Y) = 0$.
3. $|\mathsf{Cov}(X,Y)| \leqslant \sigma_X\sigma_Y$
4. Assume $\sigma_X\sigma_Y > 0$. Then,

$$\begin{aligned}\mathsf{Cov}(X,Y) = \sigma_X\sigma_Y &\text{ if and only if } X = aY+b \text{ for some } a>0,\ b\in\mathbb{R},\\ \mathsf{Cov}(X,Y) = -\sigma_X\sigma_Y &\text{ if and only if } X = aY+b \text{ for some } a<0,\ b\in\mathbb{R}.\end{aligned}$$

Proof. 1.

$$\begin{aligned}\mathsf{Cov}(aX+bY,Z) &= \mathbb{E}[(aX+bY)Z] - \mathbb{E}[aX+bY]\mathbb{E}[Z]\\ &= a\mathbb{E}[XZ] + b\mathbb{E}[YZ] - a\mathbb{E}[X]\mathbb{E}[Z] - b\mathbb{E}[Y]\mathbb{E}[Z]\\ &= a\mathsf{Cov}(X,Z) + b\mathsf{Cov}(Y,Z).\end{aligned}$$

2. Assume that X is constant (it works similarly if Y is constant). Then
$$\mathsf{Cov}(X,Y) = \mathbb{E}[\underbrace{(X-\mu_X)}_{=0}(Y-\mu_Y)] = 0.$$

3. We first consider the special case where $\sigma_X = 0$. That means that $\mathbb{E}[(X-\mu_x)^2] = 0$. By Lemma 10.2.2, this is equivalent to $\mathbb{P}(X-\mu_X = 0) = 1$. Therefore $\mathsf{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[0\cdot(Y-\mu_Y)] = 0 = \sigma_X\sigma_Y$, and the claim holds.

   Similarly, the claims holds if we assume that $\sigma_Y = 0$.

   Assume now σ_X, $\sigma_Y > 0$. Let $h(t)$ be defined by
$$\begin{aligned} h(t) &= \mathbb{E}\Big[[(X-\mu_X)t + (Y-\mu_Y)]^2\Big] \\ &= \mathbb{E}\Big[t^2(X-\mu_X)^2 + 2t(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2\Big] \\ &= t^2\mathbb{E}\Big[(X-\mu_X)^2\Big] + 2t\mathbb{E}\Big[(X-\mu_X)(Y-\mu_Y)\Big] + \mathbb{E}\Big[(Y-\mu_Y)^2\Big] \\ &= t^2\sigma_X^2 + 2t\mathsf{Cov}(X,Y) + \sigma_Y^2. \end{aligned}$$
   Hence, $h(t)$ is a quadratic function of t. Since $h(t) \geqslant 0$ for all t, h has either no real roots or only one real root, which means the discriminant $(2\mathsf{Cov}(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leqslant 0$, that is, $|\mathsf{Cov}(X,Y)| \leqslant \sigma_X\sigma_Y$.

4. With the above comments on the function h, we see that
$$\begin{aligned} |\mathsf{Cov}(X,Y)| = \sigma_X\sigma_Y &\Leftrightarrow h \text{ has one root } t^*, \\ &\Leftrightarrow \mathbb{E}\Big[[(X-\mu_X)t^* + (Y-\mu_Y)]^2\Big] = 0 \\ &\Leftrightarrow \mathbb{P}\Big[(X-\mu_X)t^* + (Y-\mu_Y) = 0\Big] = 1 \qquad \text{(by Lemma 10.2.2)} \\ &\Leftrightarrow Y = (-t^*)X + \mu_X t^* + \mu_Y \\ &\Leftrightarrow Y = aX + b \text{ with } a = -t^* \text{ and } b = \mu_X t^* + \mu_Y. \end{aligned}$$
   Moreover, by properties of the covariance already proven, we have that if $X = aY + b$ then
$$\mathsf{Cov}(X,Y) = a\mathsf{Cov}(Y,Y) + b\mathsf{Cov}(1,Y) = a\mathsf{Var}(Y)$$
   and therefore has the sign of a.

■

Corollary 10.2.4 $-1 \leqslant |\rho_{X,Y}| \leqslant 1$ and
$$\begin{aligned} \rho_{X,Y} = -1 \quad &\Leftrightarrow \quad Y = aX + b,\ a < 0, \\ \rho_{X,Y} = 1 \quad &\Leftrightarrow \quad Y = aX + b,\ a > 0. \end{aligned}$$

Corollary 10.2.5

1. $\mathsf{Cov}\Big(\sum_{i\leqslant m} a_iX_i,\ \sum_{j\leqslant n} b_jY_j\Big) = \sum_{i\leqslant m}\sum_{j\leqslant n} a_ib_j\mathsf{Cov}(X_i,Y_j)$ (bilinearity)
2. $\mathsf{Var}\Big(\sum_{i\leqslant n} X_i\Big) = \sum_{i\leqslant n}\mathrm{Var}(X_i) + 2\sum_{1\leqslant i<j\leqslant n}\mathsf{Cov}(X_i,X_j)$
3. If $X_1,\ldots,X_n$ are independent, then $\mathrm{Var}\Big(\sum_{i\leqslant n} X_i\Big) = \sum_{i\leqslant n}\mathrm{Var}(X_i)$.

Proof.

1. This follows from item 1 of the previous theorem.
2.

$$\begin{aligned}\operatorname{Var}\Big(\sum_{i\leqslant n} X_i\Big) &= \operatorname{Cov}\Big(\sum_{i\leqslant n} X_i, \sum_{i\leqslant n} X_i\Big)\\ &= \sum_{i\leqslant n}\sum_{j\leqslant n} \operatorname{Cov}(X_i, X_j)\\ &= \sum_{i\leqslant n} \operatorname{Cov}(X_i, X_i) + \sum_{i\leqslant n}\sum_{\substack{j\in\{1,\dots,n\},\\ j\neq i}} \operatorname{Cov}(X_i, X_j)\\ &= \sum_{i\leqslant n} \operatorname{Var}(X_i) + 2\sum_{1\leqslant i<j\leqslant n} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

3. This follows from (4.) in the Theorem 10.2.3.

■

10.3 Exercises

10.3.1 Check-up the basics

Exercise 10.1 — Decide whether or not each statement is necessarily true.

1. Two random variables X and Y, both following a Bernoulli(p) distribution, are independent if and only if $\operatorname{Cov}(X,Y)=0$.
2. A random variable that is equal to a constant is uncorrelated to any other random variable.

■

10.3.2 Problems

Exercise 10.2 Assume X and Y are independent random variables, both following the exponential distribution with parameter 1. Define $U=\min\{X,Y\}$ and $V=\max\{X,Y\}$.

1. Find $f_{U,V}$.
   Note: our usual approach of writing $(U,V)=g(X,Y)$ and using the formula to get $f_{U,V}$ from $f_{X,Y}$ will not work here, because g is not one-to-one. Instead, you can find $F_{U,V}$ and differentiate it.
2. Find $\operatorname{Cov}(U,V)$.

■

Exercise 10.3 We have 5 urns and 12 balls. We place the balls successively into the urns, so that any given ball is equally likely to go into any urn.

(a) Let X and Y be the number of balls that go into urn 1 and 2, respectively. Find $\operatorname{Cov}(X,Y)$.

(b) Find the probability that urns 1 and 2 end up with 2 or more balls each.
Hint. Use the fact that

$$\mathbb{P}(\{X\geq 2\}\cup\{Y\geq 2\})+\mathbb{P}(X=Y=0)+\mathbb{P}(X=Y=1)+\mathbb{P}(X=0,\,Y=1)+\mathbb{P}(X=1,\,Y=0)=1.$$

The probabilities $\mathbb{P}(X=Y=0)$, $\mathbb{P}(X=Y=1)$, $\mathbb{P}(X=0,\,Y=1)$, and $\mathbb{P}(X=1,\,Y=0)$ are not too hard to compute.

(c) Find the variance of the number of urns that end up with 2 or more balls.

■

Exercise 10.4

(a) The joint density of X and Y is given by

$$f_{X,Y}(x,y) = \frac{(y^2 - x^2)}{8} \cdot e^{-y}, \quad 0 < y < \infty, \ -y \leq x \leq y.$$

Show that $\mathbb{E}(X \mid Y = y) = 0$.

(b) The joint density of X and Y is given by

$$f_{X,Y}(x,y) = \frac{e^{-x/y} \cdot e^{-y}}{y}, \quad 0 < x < \infty, \ 0 < y < \infty.$$

Show that $\mathbb{E}(X \mid Y = y) = y$.

(c) The joint density of X and Y is given by

$$f_{X,Y}(x,y) = \frac{e^{-y}}{y}, \quad 0 < x < y < \infty.$$

Compute $\mathbb{E}(X^2 \mid Y = y)$.

■

11. Moment generating function

The goals of this chapters are:

- ▷ to define the **moment generating function** of a random variable;
- ▷ to derive some properties of the **moment generating function**.

11.0 Interlude

Interlude: Using indicator functions
If E is an event, we define the random variable

$$\mathbb{1}_E[s] = \begin{cases} 1, & \text{if } s \in E, \\ 0, & \text{if } s \notin E. \end{cases}$$

This is a Bernoulli random variable (since it only attains the values 0 and 1) with parameter

$$p = \mathbb{P}(\mathbb{1}_E = 1) = \mathbb{P}(E).$$

It is called the **indicator function** of E. It is extremely useful in computing expectations of more complicated random variables.
Note also that

$$\mathbb{E}[\mathbb{1}_E] = 1 \cdot \mathbb{P}(\mathbb{1}_E = 1) + 0 \cdot \mathbb{P}(\mathbb{1}_E = 0) = \mathbb{P}(\mathbb{1}_E = 1) = \mathbb{P}(E).$$

Question 11.1 d ducks fly over h hunters. Each hunter chooses a duck uniformly at random and shoots, hitting (and killing) it with probability p. All the hunters' choices and shots are independent. Let X denote the number of surviving ducks. Find $\mathbb{E}[X]$.

Answer: For $i = 1, 2, \ldots, d$, let $X_i = \mathbb{1}_{\{\text{Duck } i \text{ survives}\}}$. Then, $X = \sum_{i \leqslant d} X_i$, so

$$\begin{aligned}
\mathbb{E}[X] = \sum_{i \leqslant d} \mathbb{E}[X_i] &= \sum_{i \leqslant d} \Big(1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) \Big) = \sum_{i \leqslant d} \mathbb{P}(X_i = 1) = d\mathbb{P}(X_1 = 1) \\
&= d\mathbb{P}\Big(\text{Duck 1 survives} \Big) \\
&= d\mathbb{P}\Big(\bigcap_{j \leqslant h} \{\text{Hunter } j \text{ does not hit duck 1}\} \Big) \\
&= d \prod_{j \leqslant h} \mathbb{P}\Big(\text{Hunter } j \text{ does not hit duck 1} \Big) \qquad \text{(by independence)} \\
&= d\mathbb{P}\Big(\text{Hunter 1 does not hit duck 1} \Big)^h \\
&= d\Big(\underbrace{1 - \tfrac{1}{d}}_{\text{does not choose}} + \underbrace{\tfrac{1}{d}(1-p)}_{\text{chooses, misses}} \Big)^h = d(1 - p/d)^h.
\end{aligned}$$

■

Question 11.2 Let $X_1, \ldots, X_n$ be independent Bernoulli random variables and $X = \sum_{i \leqslant n} X_i$. Compute the expectation and variance of X.

Answer: If we see each X_i as a representation of the outcome of a Bernoulli trial with

$$X_i = 1 \longrightarrow \text{success}, \qquad X_i = 0 \longrightarrow \text{failure},$$

we have that $X = \#\text{successes}$, so $X \sim \text{Binomial}(n, p)$. Hence,

$$\mathbb{E}[X] = \sum_{i \leqslant n} \mathbb{E}[X_i] = \sum_{i \leqslant n} p = np,$$

$$\text{Var}(X) \overset{(\star)}{=} \sum_{i \leqslant n} \text{Var}(X_i) = \sum_{i \leqslant n} p(1-p) = np(1-p),$$

where in the equality marked $(\star)$ we have used independence. We thus have a new (and very simple) proof of the formulas for the expectation and variance of $\text{Bin}(n, p)$. ■

11.1 Moment generating function of random variables

Definition 11.1.1 — Moment generating function. The **moment generating function** of a random variable X is the function

$$M_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum e^{tx} f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, \mathrm{d}x, & \text{if } X \text{ is absolutely continuous,} \end{cases}$$

provided that the sum/integral converges for all t in an interval of the form $(-h, h)$, $h > 0$.

The following theorem explains the name "moment generating function".

Theorem 11.1.1 $\mathbb{E}(X^n) = \dfrac{d^n}{dt^n} M_X(t) \Big|_{t=0}$.

Sketch, continuous case only. Assuming that we are allowed[1] to "differentiate under the integral sign":

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx}\cdot f_X(x)\,dx = \int_{-\infty}^{\infty}\frac{d}{dt}\left(e^{tx}\cdot f_X(x)\right)dx = \int_{-\infty}^{\infty} x\cdot e^{tx}\cdot f_X(x)\,dx = \mathbb{E}(Xe^{tX}),$$

so when $t=0$ this is equal to $\mathbb{E}(X)$.

Similarly,

$$\frac{d^n}{dt^n}M_X(t) = \mathbb{E}(X^n e^{tX}),$$

so

$$\frac{d^n}{dt^n}M_X(0) = \mathbb{E}(X^n).$$

The discrete case is similar, replacing integrals by sums. ■

Question 11.3 Let $X\sim \mathrm{Poi}(\lambda)$. Compute the mgf of X.

Answer:

$$M_X(t) = \sum_{x\geqslant 0} e^{tx}\cdot\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x\geqslant 0}\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

■

Question 11.4 Let $X\sim \mathrm{Bin}(n,p)$. Compute the mgf of X.

Answer:

$$M_X(t) = \sum_{x\leqslant n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x\leqslant n}\binom{n}{x}(e^t p)^x(1-p)^{n-x} = (1-p+pe^t)^n,$$

where the last equality follows from the Binomial formula $(a+b)^n = \sum_{i\leqslant n}\binom{n}{i}a^i b^{n-i}$. ■

Question 11.5 Let $X\sim\mathcal{N}(0,1)$. Compute the mgf of X.

Answer: By definition,

$$M_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\cdot e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx)}\,\mathrm{d}x.$$

We can rewrite the exponent of the integrand, observing that

$$x^2-2tx = x^2-2tx+t^2-t^2 = (x-t)^2-t^2,$$

and thus get

$$M_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}((x-t)^2-t^2)}\,\mathrm{d}x = e^{t^2/2}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\,\mathrm{d}x = e^{t^2/2}.$$

[1] It is not always allowed to switch the order of integration and differentiation. The conditions of when it is or is not allowed will be covered in later courses, such as *Probability and Measure*.

■

Proposition 11.1.2 $M_{aX+b}(t) = e^{bt} M_X(at)$.

Proof.

$$M_{aX+b}(t) = \mathbb{E}(e^{(aX+b)t}) = e^{bt}\mathbb{E}(e^{atX}) = e^{bt} M_X(at).$$

■

Question 11.6 Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Compute the mgf of X.

Answer: We have $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. So

$$X = \sigma Z + \mu \quad \Longrightarrow M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} \cdot M_Z(\sigma t) = e^{\mu t} \cdot e^{\frac{\sigma^2 t^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

■

Question 11.7 Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Compute $M_X'(0)$ and $M_X''(0)$.

Answer: We have

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}},$$
$$M_X'(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}(\mu + \sigma^2 t),$$
$$M_X''(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}(\mu + \sigma^2 t)^2 + e^{\mu t + \frac{\sigma^2 t^2}{2}} \cdot \sigma^2.$$

Thus

$$M_X'(0) = \mu,$$
$$M_X''(0) = \mu^2 + \sigma^2.$$

■

Theorem 11.1.3 If X, Y are such that $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X = F_Y$ (that is, X and Y have the same distribution).

We omit the proof of this theorem.

Proposition 11.1.4 If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t) M_Y(t), \quad \text{for all } t \geq 0.$$

Proof.

$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX} e^{tY}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}) = M_X(t) M_Y(t).$$

■

Question 11.8 Let $X \sim \text{Bin}(n,p)$. Compute the mgf of X using Proposition 11.1.4.

Answer: Since X is Binomialy distributed (with parameters n and p) it can be interpreted as the sum of n i.i.d. Bernoulli random variables $X_1,\dots,X_n$ (of parameter p). Thus, iterating Proposition 11.1.4, we get

$$\begin{aligned} M_X(t) &= M_{X_1+\cdots+X_n}(t) = M_{X_1+\cdots+X_{n-1}}(t)\times M_{X_n}(t) = \cdots = M_{X_1}(t)\times\cdots\times M_{X_n}(t) = M_{X_1}(t)^n \\ &= (\mathbb{E}[e^{tX_1}])^n = (pe^{t\times 1}+(1-p)e^{t\times 0})^n = (e^t+1-p)^n. \end{aligned}$$

Naturally we find the same answer as for Question 11.4. ■

Question 11.9 Let $X \sim \mathcal{N}(\mu_1,\sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2,\sigma_2^2)$ and assume they are independent. Find the distribution of $X+Y$.

Answer: Recall that, if $X \sim \mathcal{N}(\mu,\sigma^2)$, then

$$M_X(t) = e^{\mu_1 t+\sigma_1^2 t^2/2}, \qquad \text{and} \qquad M_Y(t) = e^{\mu_2 t+\sigma_2^2 t^2/2}.$$

Thus,

$$M_{X+Y}(t) = M_X(t)\cdot M_Y(t) = e^{(\mu_1+\mu_2)t+(\sigma_1^2+\sigma_2^2)t^2/2}.$$

This is the mgf of a $\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ random variable. Hence, by Theorem 11.1.3, we have

$$X+Y \sim \mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2).$$

■

The last example can be slightly generalized, as follows. It will be quite useful in the next chapter.

Proposition 11.1.5 Let $X \sim \mathcal{N}(\mu_1,\sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2,\sigma_2^2)$ independent normally distributed random variables. Let $a,b,c \in \mathbb{R}$. Then $aX+bY+c \sim \mathcal{N}(a\mu_1+b\mu_2+c, a^2\sigma_1^2+b^2\sigma_2^2)$.

Proof. The independence of X and Y implies that aX and $bY+c$ are independent. Thus

$$\begin{aligned} M_{aX+bY+c}(t) &= M_{aX}(t)M_{bY+c} && \text{(by Proposition 11.1.4)} \\ &= M_X(at)e^{ct}M_Y(bt) && \text{(by Proposition 11.1.2)} \\ &= e^{\mu_1 at+\frac{\sigma_1^2a^2t^2}{2}}e^{ct}e^{\mu_2 bt+\frac{\sigma_2^2b^2t^2}{2}} && \text{(see Question 11.7)} \\ &= e^{(\mu_1 a+\mu_2 b+c)t+\frac{(a^2\sigma_1^2+b^2\sigma_2^2)t^2}{2}} \\ &= M_Z(t) \quad \text{with } Z \sim \mathcal{N}(a\mu_1+b\mu_2+c, a^2\sigma_1^2+b^2\sigma_2^2) && \text{(see Question 11.7 again).} \end{aligned}$$

Therefore, Theorem 11.1.3 yields the proof. ■

11.2 Moment generating functions of random vectors

Definition 11.2.1 — Joint mgf. The **joint moment generating function** of a random vector $(X_1,\dots,X_n)$ is the function $M_{X_1,\dots,X_n}(t_1,\dots,t_n) = \mathbb{E}(e^{t_1X_1+\cdots+t_nX_n})$.

R If $(X_1,\dots,X_n)$ is *discrete*,

$$M_{X_1,\dots,X_n}(t_1,\dots,t_n) = \sum_{(x_1,\dots,x_n)} e^{t_1x_1+\cdots t_nx_n}\cdot f_{X_1,\dots,X_n}(x_1,\dots,x_n),$$

provided that the sum converges in an interval of the form $(-h,h)^n$, $h>0$.

R If $(X_1,\dots,X_n)$ is *continuous*,

$$M_{X_1,\dots,X_n}(t_1,\dots,t_n) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t_1x_1+\cdots+t_nx_n}\cdot f_{X_1,\dots,X_n}(x_1,\dots,x_n)\ \mathrm{d}x_1\cdots\mathrm{d}x_n,,$$

provided that the integral converges in an interval of the form $(-h,h)^n$, $h>0$.

11.3 Exercises

11.3.1 Check-up the basics

Exercise 11.1 — Decide whether or not each statement is necessarily true.

1. If X is a discrete random variable such that, for some constant $M>0$, we have $\mathbb{P}(|X|\leq M)=1$, then $M_X(t)$ is defined for all $t\in\mathbb{R}$.
2. If X is a random variable such that $M_X(t)$ is defined for all t, then $M_X(t)^2 = M_{X^2}(t)$ for all t.
3. If $X\sim\text{Geom}(p)$ for $p\in(0,1)$, then $M_X(t)=\frac{pe^t}{1-(1-p)e^t}$ for $t<-\log(1-p)$.

■

11.3.2 Problems

Exercise 11.2 Show that, if $X\sim\mathcal{N}(\mu,\sigma^2)$, then $M_X(t)=e^{\mu t+\sigma^2t^2/2}$. ■

12. Law of large numbers

The goals of this chapters are:

- ▷ to define various concepts from Statistics: **random sample**, **parameter**, **statistic**, **unbiased/consistent estimator**, **sample mean/variance**;
- ▷ to establish some properties related to these concepts;
- ▷ to define **convergence in probability** of a sequence of random variables;
- ▷ to derive the **Markov**'s and **Chebyschev**'s inequalities;
- ▷ to derive the **weak law of large number**.

12.1 Basic concepts of random samples

We will start discussing the difference of perspective that exists between Probability Theory and Statistics. Let us see an example:

Question 12.1 — Probability problem. A coin gives heads with probability $\frac{1}{3}$ and tails with $\frac{2}{3}$. If we toss it five times, what is the probability that we get: Ⓗ Ⓣ Ⓣ Ⓗ Ⓗ?

Answer: $\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2$. ■

Question 12.2 — Statistics problem. A coin gives heads with some unknown probability p (and tails with probability $1-p$). We toss it five times and get: Ⓗ Ⓣ Ⓣ Ⓗ Ⓗ. What can we say about p?

In Question 12.1, we know the parameter p and want to find how likely a certain outcome is. In the second problem, we start with the observation of the outcome and want information about the parameter. Naturally, in the second problem, the only thing we can say *for sure* is that $p \neq 0$ and $p \neq 1$. Still, suppose we ask a more precise question, such as:
What is the value of p that makes the outcome Ⓗ Ⓣ Ⓣ Ⓗ Ⓗ **most likely**?

Answer: Let

$$g(p) = p^3(1-p)^2, \quad 0 < p < 1.$$

p maximizes $g(p)$ if and only if p maximizes $\log(g(p)) = 3\log(p) + 2\log(1-p)$.

$$\frac{\mathrm{d}}{\mathrm{d}p}(\log(g(p))) = \frac{3}{p} - \frac{2}{1-p} = 0 \quad \Longleftrightarrow \quad p = \frac{3}{5}.$$

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}(\log(g(p))) = -\frac{3}{p^2} - \frac{2}{(1-p)^2} < 0,$$

so $\frac{3}{5}$ is indeed a maximizer. ∎

Regarding the last question of the above example, something more general can be shown in the same way: if $X_1, \ldots, X_n \sim$ Bernoulli(p) and independent, the value of p that maximizes $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid p)$ is

$$\hat{p} = \frac{\#\text{ heads in } x_1, \ldots, x_n}{n}.$$

We call $\hat{p}$ an **estimator** for the unknown parameter p. Since $\hat{p}$ is the value of p that makes the observed outcome most likely, it is called the **maximum likelihood** estimator.
With these examples in mind, let us now give some formal definitions.

Definition 12.1.1 — Random sample. A **Random sample** of size n is simply a sequence $X_1, \ldots, X_n$ of independent random variables, all with the same pdf or pmf $f(x)$. We thus have:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i \leqslant n} f(x_i).$$

We refer to f as the pmf/pdf of the population. This terminology is due to the following way of thinking:

- ▷ We have an infinite population of entities (individuals, objects, ...).
- ▷ There is a numerical attribute associated to each member of the population (example: person height).
- ▷ f describes how the attribute is distributed among the population.
- ▷ We "select" n individuals from the population and record their attributes to obtain $X_1, \ldots, X_n$.

Definition 12.1.2 — Parameter. A **parameter** is a constant that defines the population pmf/pdf $f(x)$

For example, if f corresponds to the Exponential(β) distribution, then β is the parameter.

Definition 12.1.3 — Statistic. A **statistic** is a function of a random sample.

$$Y = T(X_1, \ldots, X_n), \qquad T : \mathbb{R}^n \to \mathbb{R}.$$

We often use a statistic to estimate a parameter. In the coin example seen earlier,

$$\begin{aligned} f &\longrightarrow \text{pmf of Bernoulli}(p) \\ X_1, \ldots, X_n &\longrightarrow \text{random samples, Bernoulli random variables} \\ Y = \frac{X_1 + \cdots + X_n}{n} &\longrightarrow \text{statistic used to predict the parameter } p \text{ ("estimator")} \end{aligned}$$

Definition 12.1.4 — Unbiased estimator. A statistic Y is a **unbiased estimator** for the parameter θ if $\mathbb{E}(Y) = \theta$.

Pay attention: parameters are numbers (or sometimes vectors); statistics are random variables!

We now define the two most common statistics of random samples.

Definition 12.1.5 — Sample mean/variance.

Sample mean: $$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n},$$

Sample variance: $$S_n^2 := \frac{1}{n-1}\sum_{i\leqslant n}(X_i - \overline{X}_n)^2.$$

We will see that $\overline{X}_n$ is an estimator for μ (the population mean $= \mathbb{E}(X_i)$) and S_n^2 is an estimator for σ^2 (the population variance $= \mathrm{Var}(X_i)$).

Lemma 12.1.1

$$S_n^2 = \frac{1}{n-1}\sum_{i\leqslant n} X_i^2 - \frac{n}{n-1}\overline{X}_n^2.$$

Proof. We need to prove: $(n-1)S_n^2 = \sum_{i\leqslant n} X_i^2 - n\overline{X}_n^2$. Indeed,

$$\begin{aligned}(n-1)S_n^2 &= \sum_{i\leqslant n}(X_i - \overline{X}_n)^2 \\ &= \sum_{i\leqslant n}(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2) \\ &= \sum_{i\leqslant n} X_i^2 - 2\underbrace{\sum_{i\leqslant n} X_i\overline{X}_n}_{n\overline{X}_n} + n\overline{X}_n^2 \\ &= \sum_{i\leqslant n} X_i^2 - n\overline{X}_n^2.\end{aligned}$$

■

Theorem 12.1.2 Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2. Then,

(a) $\mathbb{E}(\overline{X}_n) = \mu$,
(b) $\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X_1)}{n}$,
(c) $\mathbb{E}(S_n^2) = \sigma^2$.

Proof. (a) $\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i\leqslant n}\mathbb{E}(X_i) = \frac{1}{n}n\mu = \mu$.

(b) $\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = \frac{\mathrm{Var}(X_1)}{n}$.

(c) First note that

$$\mathbb{E}(S_n^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i\leqslant n} X_i^2 - \frac{n}{n-1}\overline{X}_n^2\right) = \frac{1}{n-1}\sum_{i\leqslant n}\mathbb{E}(X_i^2) - \frac{n}{n-1}\mathbb{E}(\overline{X}_n^2).$$

The first expectation (of the right hand side in the last display) is easily handled:

$$\mathbb{E}(X_i^2) = \mathrm{Var}(X_i) + \mathbb{E}(X_i)^2 = \sigma^2 + \mu^2.$$

For the second expectation, we first note that

$$\overline{X}_n^2 = \frac{1}{n^2}\sum_{i\neq j} X_i X_j + \frac{1}{n^2}\sum_i X_i^2,$$

and that the expectation of the summands above are $\mathbb{E}X_iX_j = \mathbb{E}X_i\mathbb{E}X_j = \mu^2$ (by independence of X_i and X_j when $i \neq j$) and $\mathbb{E}X_i^2 = \sigma^2 + \mu^2$; and therefore

$$\mathbb{E}(\overline{X}_n^2) = \frac{\#\{(i,j) \mid i,j \in \{1,\ldots,n\} \text{ and } i \neq j\}}{n^2}\mu^2 + \frac{n}{n^2}(\sigma^2+\mu^2) = \mu^2 + \frac{\sigma^2}{n}.$$

It follows that

$$\mathbb{E}(S_n^2) = \frac{n}{n-1}(\sigma^2+\mu^2) - \frac{n}{n-1}\left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n\sigma^2}{n-1} + \frac{n\mu^2}{n-1} - \frac{n\mu^2}{n-1} - \frac{\sigma^2}{n-1} = \sigma^2.$$

■

Question 12.3 Let $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ be independent. Find the distribution of $\overline{X}_n$.

Answer: We have already seen that

$$M_{X_i}(t) = e^{\mu t + \sigma^2 t^2/2}.$$

On the other hand

$$M_{\overline{X}_n}(t) = \mathbb{E}e^{t\frac{X_1+\cdots+X_n}{n}} = \mathbb{E}e^{\frac{t}{n}(X_1+\cdots+X_n)} = M_{X_1+\cdots+X_n}(t/n) = M_{X_1}(t/n)\times\cdots\times M_{X_n}(t/n)$$
$$= M_{X_1}(t/n)^n,$$

Therefore, we get

$$M_{\overline{X}_n}(t) = \left(e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}}\right)^n = e^{\mu t + \frac{(\sigma^2/n)t^2}{2}},$$

so $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. ■

Question 12.4 Let $X_1,\ldots,X_n$ are independent, all distributed as $\Gamma(\alpha,\beta)$. Find the distribution of $\overline{X}_n$.

Answer: If $X \sim \Gamma(\alpha,\beta)$, then

$$M_X(t) = \mathbb{E}e^{tX} = \int_0^\infty e^{tx}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}\,\mathrm{d}x = \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-t)^\alpha}\underbrace{\int_0^\infty e^{-y}y^{\alpha-1}\,\mathrm{d}y}_{\Gamma(\alpha)} = \left(\frac{\beta}{\beta-t}\right)^\alpha$$
$$= \left(\frac{1}{1-t/\beta}\right)^\alpha$$

Hence

$$M_{\overline{X}_n}(t) = M_{X_1}(t/n)^n = \left(\frac{1}{1-\frac{t}{\beta n}}\right)^{\alpha n},$$

so $\overline{X}_n \sim \Gamma(\alpha n, \beta n)$. ■

12.2 Convergence concepts

Definition 12.2.1 A sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a constant $c \in \mathbb{R}$ if:

$$\text{for all } \varepsilon > 0, \qquad \mathbb{P}(|X_n - c| > \varepsilon) \xrightarrow{n\to\infty} 0$$

which can be read as: "as n gets larger, it becomes very unlikely that X_n is far from c". We write: $X_n \xrightarrow[\mathbb{P}]{n\to\infty} c$.

Definition 12.2.2 Let $X_1, \ldots, X_n$ be a random sample from a pmf/pdf $f(x)$ with parameter θ. Assume Y_n is a statistic associated to $X_1, \ldots, X_n$. We say Y_n is a **consistent estimator** of θ if Y_n converges to θ in probability as $n \to \infty$.

Theorem 12.2.1 — Weak Law of Large Numbers. Let X_1, $X_2, \ldots$ be independent and identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, $\overline{X}_n$ is a consistent estimator for μ, that is,

$$\text{for all } \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(|\overline{X}_n - \mu| > \varepsilon\right) = 0.$$

Before we prove the weak law of large number we need to state the following two very useful bounds.

Theorem 12.2.2 — Markov inequality. Let $a > 0$ and Y be any non-negative random variable. Then,

$$\mathbb{P}(Y \geqslant a) \leqslant \frac{1}{a}\mathbb{E}[Y].$$

Proof. The proof relies on the very simple inequality

$$\mathbb{1}\{Y \geq a\} \leq \frac{Y}{a},$$

which one can check immediately because the left hand side equals 0 if $Y < a$, and the right hand side is at least 1 if $Y \geq a$. It implies

$$\mathbb{P}(Y \geqslant a) = \mathbb{E}\left[\mathbb{1}\{Y \geqslant a\}\right] \leqslant \mathbb{E}\left[\frac{Y}{a}\right] = \frac{1}{a}\mathbb{E}[Y].$$

■

Theorem 12.2.3 — Chebyschev's inequality. If X is any random variable and $a > 0$ then

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}.$$

Proof. The random variable $Z := (X - \mathbb{E}X)^2$ is nonnegative. Hence we can apply Markov's inequality:

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) = \mathbb{P}((X - \mathbb{E}X)^2 \geq a^2) \overset{\text{Markov}}{\leq} \frac{\mathbb{E}Z^2}{a^2} = \frac{\mathsf{Var}(X)}{a^2},$$

where we use the definition of $\mathsf{Var}(X)$ for the last equation. ■

The proof of the weak law of large number is now a one-liner.

Proof of the weak law of large number. Applying the observations (a) and (b) of Theorem 12.1.2 on the mean and variance of $\overline{X}_n$ and Chebyschev, we have:

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) = \mathbb{P}(|\overline{X}_n - \mathbb{E}\overline{X}_n| > \varepsilon) \overset{\text{Chebyschev}}{\leq} \frac{\mathsf{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\mathsf{Var}(X_1)}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0.$$

■

12.3 Exercises

12.3.1 Problems

Exercise 12.1

a) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with $X_n \sim \mathsf{Ber}(1/n)$. Show that $X_n \xrightarrow[\mathbb{P}]{n\to\infty} 0$. *In words: Show that $(X_n)_{n\in\mathbb{N}}$ converges in probability to 0.*

b) Let $X_n \sim \mathsf{Bin}(n,p)$ for $n \in \mathbb{N}$ and $p \in (0,1)$. Show that $\frac{1}{n}X_n \xrightarrow[\mathbb{P}]{n\to\infty} p$.

■

Exercise 12.2 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with expectation $-\infty < \mu < \infty$ and variance $0 < \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$ and for $\alpha \in \mathbb{R}$ define a function h_n given by

$$h_n(\alpha) := \mathbb{P}(S_n \leqslant \alpha n).$$

Find $h = \lim_{n\to\infty} h_n$. ■

Exercise 12.3 Color blindness appears in 1% of the people in a certain population. How large must a sample be if the probability of this sample containing a color-blind person is to be 0.95 or more? (Assume that the population is large enough to be considered infinite, so that sampling can be considered to be with replacement). ■

Exercise 12.4 Let $X_1, \ldots, X_n$ be iid with pdf $f_X(x)$, and let $\overline{X}$ denote the sample mean. Show that

$$f_{\overline{X}}(x) = n f_{X_1+\cdots+X_n}(nx).$$

■

Exercise 12.5 Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(0,1)$ random variables. Let

$$Y_1 = \left| \frac{1}{n} \sum_{i=1}^{n} X_i \right|, \qquad Y_2 = \frac{1}{n} \sum_{i=1}^{n} |X_i|.$$

Calculate $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_2)$, and establish an inequality between them. ■

Exercise 12.6 Let $X_1, X_2, \ldots$ be independent and identically distributed random variables. As usual, let

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Prove the following recursion relations:

$$\overline{X}_{n+1} = \frac{X_{n+1} + n\overline{X}_n}{n+1},$$

$$nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1} \right) (X_{n+1} - \overline{X}_n)^2.$$

■

Exercise 12.7 Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent random variables, each being uniformly distributed on the interval $[1,2]$. Show that the sequence

$$\left(\left(\prod_{i=1}^{n} X_i \right)^{1/n} \right)_{n \in \mathbb{N}}$$

converges in probability and find the limit.
Hint: Take the logarithm + Weak Law of Large Numbers. ■

Exercise 12.8 Assume U follows the continuous uniform distribution on $(0,1)$. For each $k \in \{1,2,3,\ldots\}$, let X_k denote the kth digit in the decimal expansion of U. For instance, if $U = 0.9241\ldots$, we have $X_1 = 9$, $X_2 = 2$, $X_3 = 4$, $X_4 = 1$,... Note that, for every $k \in \{1,2,3,\ldots\}$ and every $i \in \{0,1,\ldots,9\}$, we have

$$\begin{aligned} f_{X_k}(i) = \mathbb{P}(X_k = i) &= \sum_{i_1=0}^{9} \sum_{i_2=0}^{9} \cdots \sum_{i_{k-1}=0}^{9} \mathbb{P}(X_1 = i_1, \ldots, X_{k-1} = i_{k-1}, X_k = i) \\ &= \sum_{i_1=0}^{9} \sum_{i_2=0}^{9} \cdots \sum_{i_{k-1}=0}^{9} \mathbb{P}(0.i_1 i_2 \cdots i_{k-1} i \le U < 0.i_1 i_2 \cdots i_{k-1}(i+1)) \\ &= \sum_{i_1=0}^{9} \sum_{i_2=0}^{9} \cdots \sum_{i_{k-1}=0}^{9} \frac{1}{10^k} = 10^{k-1} \cdot \frac{1}{10^k} = \frac{1}{10}. \end{aligned}$$

Hence,

$$f_{X_k}(i) = \frac{1}{10}, \qquad i \in \{0,1,\dots,9\}. \tag{♣}$$

Additionally, for every $i_1,\ i_2,\dots,\ i_k$ we have

$$\begin{aligned} f_{X_1,\dots,X_k}(i_1,\dots,i_k) &= \mathbb{P}(X_1 = i_1,\dots,X_k = i_k) \\ &= \mathbb{P}\left(0.i_1 i_2 \cdots i_{k-1} i_k \le U < 0.i_1 i_2 \cdots i_{k-1}(i_k+1)\right) = \frac{1}{10^k}. \end{aligned} \tag{♠}$$

By putting together (♣) and (♠), we see that

$$f_{X_1,\dots,X_k}(x_1,\dots,x_k) = f_{X_1}(x_1)\cdot f_{X_2}(x_2)\cdots f_{X_k}(x_k),$$

that is, *the random variables $X_1, X_2, \dots$ are independent, all following a discrete uniform distribution in the set $\{0,1,\dots,9\}$.*

(a) For $i \in \{0,1,\dots,9\}$, let $Y_n^{(i)}$ denote the number of times the digit i appears among the first n digits of the decimal expansion of U. Determine the distribution of $Y_n^{(i)}$.

(b) Prove the **Chebyshev inequality**. That is, for any random variable X,

$$\mathbb{P}(|X - \mathbb{E}(X)| > x) \le \frac{\mathrm{Var}(X)}{x^2}, \qquad x > 0.$$

Use this and part (a) to find an upper bound for $\mathbb{P}\left(Y_n^{(i)} < \left(\frac{1}{10} - \varepsilon\right) n\right)$, for $\varepsilon > 0$.

(c) Prove:

$$\text{for all } \varepsilon > 0,\ \lim_{n\to\infty} \mathbb{P}\left(\begin{array}{l}\text{each of the digits from 0 to 9 appears among}\\ \text{the } n \text{ first digits of } U \text{ more than } (\frac{1}{10} - \varepsilon)\cdot n\\ \text{times}\end{array}\right) = 1.$$

■

Trivia

Sir Francis Galton (1822–1911) was a cousin of Charles Darwin. He was a "gentleman of leisure" (he was so rich that he did not have to work) who liked to entertain himself by playing around with science. He wrote an enormous amount of articles and books (340 in total) and made serious contributions in statistics, psychology, sociology, biology and forensic science. He for instance designed a method for classifying fingerprints that is still in use today. More frivolous contributions include a paper in the journal "nature" (nowadays this is one of the most prestigious scientific journals) on how to cut a birthday cake and a "beauty map" of Britain. For the beauty map, he went around lots of places in Britain while secretly rating the attractiveness of the ladies he saw around him by carving marks into a stick, and then later he made a map of Britain showing these "measurements". If you think that is bad, wait until you read this: he was also a proponent of "eugenics" (a term he invented), that is "improving" the human race by controlling who gets to procreate (for instance by sterilizing or even killing "undesirables" – although it has to be said that Galton himself was only focused on encouraging the "desirables" to have more children.).

In 1894 he published the design of a device that is now called the "Galton board" (also "bean box" and "quincunx"). It is an experimental setup where balls are dropped though a small hole onto nails that are arranged in several horizontal rows, in such a way that that balls can pass through the gaps between the nails and when they do so they will hit the nails of the next row, and pass through a gap, and hit a nail on the next row and so on until the last row of nails. After the balls pass through the last row, they fall into "partitions". (See Figure 12.2.)

If you drop lots of balls then the heights of the stacks of balls in the "partitions" (almost) always form a similar, "bell shaped" pattern, with the highest stack in the middle.

What is going on here?

Let us assume the construction of the machine is such that if a ball passes though a hole in row i, it will always fall onto the nail of row $i+1$ directly underneath the hole, and the ball and nail are perfectly round and the ball will hit the nail exactly in its middle. Then it makes sense that the ball will go through the hole to the left of the nail it has just hit with probability $1/2$ and trough the hole right of the nail with probability $1/2$. So if there are n rows, the path of any one ball is described by the $\mathrm{Bin}(n, 1/2)$-distribution. This binomial distribution describes how many times the ball moved

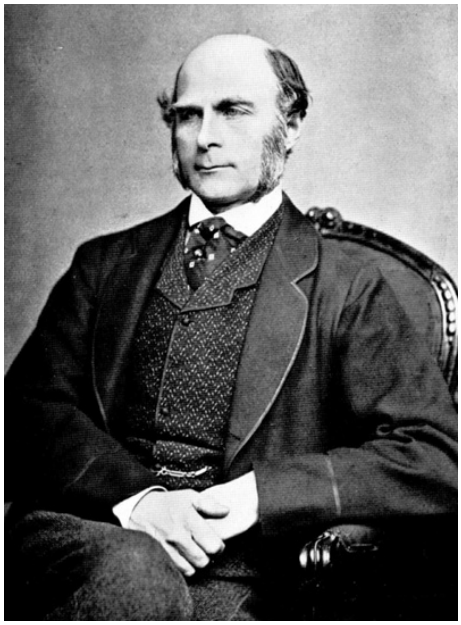


Figure 12.1: Sir Francis Galton, ca. 1850

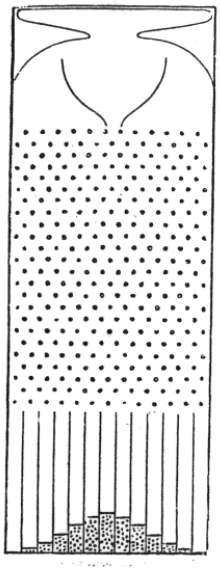


Figure 12.2: Left: The first ever Galton board, constructed under supervision of Galton himself, in the library of University College London, Right: a diagram from Galton's article describing the device.

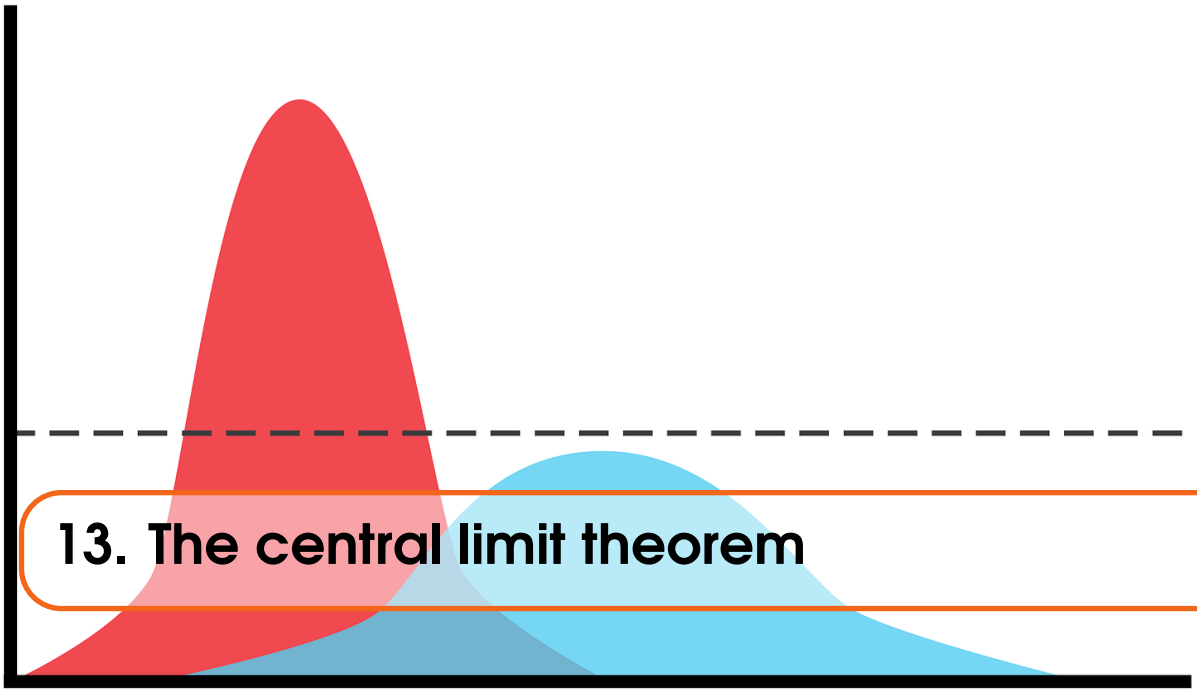
right. The expected number of balls (height) of the middle "partition" should therefore be

$$\mathbb{P}(\text{Bin}(n,1/2) = n/2) \times \#\text{balls}.$$

(Assuming n is even so $n/2$ is an integer.) Similarly the height of the partition i to the right (or $-i$ to the left) of the middle should be $\approx \mathbb{P}(\text{Bin}(n,1/2) = n/2+i) \times \#\text{balls}$. The *expected* height of this partition equals this expression and, by the law of large numbers, if we drop a large number of balls then the number of balls in the middle partition would be close to the expected number, with a

probability that tends to one as the number of balls dropped into the device goes to infinity. Similarly the height of the partition i to the right of the middle should be $\approx \mathbb{P}(\text{Bin}(n,1/2) = n/2+i) \times$ #balls. Conclusion: the heights of the partitions ought to be proportional to the pmf of the binomial distributions.

On some newer versions of the Galton board (if I don't forget I'll bring one to the lecture) have the pdf of a normal with appropriate parameters μ, σ^2 drawn on them. And the heights of the partitions typically closely match the heights of this curve ....



13. The central limit theorem

The goals of this chapters are:

- ▷ to define **convergence in distribution** of a sequence of random variable;
- ▷ to derive the **central limit theorem**.

13.1 Convergence in distribution

Definition 13.1.1 A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for every } x \in \mathbb{R} \text{ at which } F_X \text{ is continuous.}$$

We denote this by: $X_n \xrightarrow[(\mathrm{d})]{n\to\infty} X$.

Note: If X is a *continuous* random variable, then F_X is continuous everywhere, so

$$X_n \xrightarrow[(\mathrm{d})]{n\to\infty} X \quad \text{if and only if} \quad F_{X_n}(x) \xrightarrow{n\to\infty} F_X(x) \text{ for every } x \in \mathbb{R}.$$

Question 13.1 Let $a, b \in \mathbb{R}$, $a < b$. Let X_n be discrete, uniformly distributed on

$$\left\{a + \frac{b-a}{n},\ a + 2\frac{b-a}{n}, \ldots, a + n\frac{b-a}{n} = b\right\},$$

and let $X \sim \mathrm{ContUnif}(a,b)$. Show that $X_n \xrightarrow[(\mathrm{d})]{n\to\infty} X$.

Answer:

▷ if $x \le a$, $F_{X_n}(x) = 0 \xrightarrow{n\to\infty} 0 = F_X(x)$;

▷ if $x \ge b$, $F_{X_n}(x) = 1 \xrightarrow{n\to\infty} 1 = F_X(x)$;

▷ if $x \in (a,b)$,

$$F_{X_n}(x) = \sum_{\substack{i\in\{1,2,\dots\}:\\ a+i\frac{b-a}{n}\le x}} \frac{1}{n} = \sum_{i=1}^{\lfloor n\frac{x-a}{b-a}\rfloor} \frac{1}{n} = \frac{1}{n}\left\lfloor n\frac{x-a}{b-a}\right\rfloor$$
$$= \frac{1}{n}\left[n\frac{x-a}{b-a} + \underbrace{\left\lfloor n\frac{x-a}{b-a}\right\rfloor - n\frac{x-a}{b-a}}_{\in[-1,0]}\right] \xrightarrow{n\to\infty} \frac{x-a}{b-a} = F_X(x).$$

■

Lemma 13.1.1 If X is continuous and $X_n \xrightarrow[\text{(d)}]{n\to\infty} X$, then

$$\mathbb{P}(X_n = x) \xrightarrow{n\to\infty} 0, \quad \text{for all } x \in \mathbb{R}.$$

Proof. Fix $\varepsilon > 0$. Since F_X is continuous at x, there exists $\delta > 0$ such that

$$F_X(x+\delta) - F_X(x-\delta) < \varepsilon/3.$$

Also, since $F_{X_n}(x+\delta) \xrightarrow{n\to\infty} F_X(x+\delta)$ and $F_{X_n}(x-\delta) \xrightarrow{n\to\infty} F_X(x+\delta)$, there exists n_0 such that, for all $n \ge n_0$,

$$|F_{X_n}(x-\delta) - F_X(x-\delta)| < \varepsilon/3, \qquad |F_{X_n}(x+\delta) - F_X(x+\delta)| < \varepsilon/3.$$

Hence, if $n \ge n_0$,

$$\begin{aligned}\mathbb{P}(X_n = x) &\le \mathbb{P}(x-\delta < X_n \le x+\delta)\\ &= F_{X_n}(x+\delta) - F_{X_n}(x-\delta)\\ &\le \left(F_X(x+\delta) + \frac{\varepsilon}{3}\right) - \left(F_X(x-\delta) - \frac{\varepsilon}{3}\right)\\ &< \varepsilon,\end{aligned}$$

completing the proof. ■

Proposition 13.1.2 If X is continuous and $X_n \xrightarrow[\text{(d)}]{n\to\infty} X$, then for every interval $I \subset \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(X_n \in I) = \mathbb{P}(X \in I).$$

Proof. Let us prove this for the case $I = (a,b)$; all other types of interval follow because of Lemma 13.1.1.

$$\begin{aligned}\mathbb{P}(X_n \in (a,b)) &= \mathbb{P}(X_n \in (a,b]) - \mathbb{P}(X_n = b)\\ &= F_{X_n}(b) - F_{X_n}(a) - \mathbb{P}(X_n = b)\\ &\xrightarrow{n\to\infty} F_X(b) - F_X(a) - \underbrace{\mathbb{P}(X = b)}_{=0} = \mathbb{P}(X \in (a,b))\end{aligned}$$

■

13.2 Central limit theorem

Theorem 13.2.1 — Central Limit Theorem. Let $X_1, X_2, \dots$ be independent and identically distributed with mean μ and variance σ^2 (both finite). Then,

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[\text{(d)}]{n\to\infty} Z, \quad \text{where } Z \sim \mathcal{N}(0,1).$$

We will see the proof of the central limit theorem a bit later in this chapter.

R We can summarize the above convergence by:

$$n \text{ large} \implies \sqrt{n}\cdot\frac{\overline{X}_n - \mu}{\sigma} \overset{\text{(d)}}{\approx} Z$$

"$\approx$" means: "is approximately distributed as". This can be rearranged as:

$$\sum_{i=1}^{n} X_i \overset{\text{(d)}}{\approx} \mu\cdot n + \sigma\sqrt{n}Z.$$

We can see μn as a "first-order term" and $\sigma\sqrt{n}Z$ as a "second-order term". This approximation means that $\sum_{i=1}^n X_i$ has near-Gaussian fluctuations of order $\sigma\sqrt{n}$ from its mean μn.
It is also useful to note that:

$$\sqrt{n}\cdot\frac{\overline{X}_n - \mu}{\sigma} = \frac{\sum_{i=1}^n X_i - \mu n}{\sigma\sqrt{n}}.$$

Question 13.2 A lamp burns after an amount of time (in weeks) that is exponentially distributed with rate parameter 0.2. When it burns, it is immediately replaced. Estimate the probability that it takes more than 245 weeks for the 40th lamp to be replaced.

Answer: T_i = Lifetime of lamp i, $i = 1, 2, \dots$. $T_i \sim \text{Exp}(0.2)$, so $\mu = 5$ and $\sigma^2 = 5^2$.

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{40} T_i > 245\right) &= \mathbb{P}\left(\frac{\sum_{i=1}^{40} T_i - 40\cdot 5}{5\sqrt{40}} > \frac{245 - 40\cdot 5}{5\sqrt{40}}\right) \\
&\approx \mathbb{P}(Z > 1.42) \\
&= 1 - F_Z(1.42) \\
&\approx 1 - 0.9222 \qquad \text{(see Table 13.3)} \\
&= 0.0778.
\end{aligned}$$

■

Moment-generating functions are useful for determining if a sequence converges in distribution.

Theorem 13.2.2 Assume that $X_1, X_2, \dots$ and X are such that

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t) \text{ for all } t \text{ in a neighborhood of } 0.$$

Then, $X_n \xrightarrow[\text{(d)}]{n\to\infty} X$.

We admit the last theorem, without proof.

We now turn to the proof of the Central Limit Theorem.

Proof of Theorem 13.2.1. We will only prove it under the extra assumption that $M_{X_i}(t)$ is well defined for all t in an interval of the form $(-\delta,\delta)$, $\delta > 0$.
So assume $X_1, X_2, \ldots$ have mean μ and variance σ^2. Let $Y_i = \frac{X_i-\mu}{\sigma}$, so that $\mathbb{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = 1$.
Also note that $M_{Y_i}(t)$ is well defined in a neighborhood of 0 since:

$$M_{Y_i}(t) = M_{\frac{X_i}{\sigma}-\frac{\mu}{\sigma}}(t) = e^{-\mu/\sigma} \cdot M_{X_i}(t/\sigma).$$

Also,

$$M'_{Y_i}(0) = \mathbb{E}(Y_i) = 0 \qquad \text{and} \qquad M''_{Y_i}(0) = \mathbb{E}(Y_i^2) = \mathrm{Var}(Y_i) = 1.$$

Now,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \frac{X_i-\mu}{\sigma} = \frac{\sum_{i=1}^{n} X_i - \mu \cdot n}{\sigma} = n \cdot \frac{\overline{X}_n - \mu}{\sigma},$$

so

$$\sqrt{n} \cdot \frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

so

$$M_{\sqrt{n}\frac{\overline{X}_n-\mu}{\sigma}}(t) = M_{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}(t) = M_{\sum_{i=1}^{n} Y_i}(t/\sqrt{n}) = M_{Y_i}(t/\sqrt{n})^n = e^{n \cdot \log M_{Y_i}(t/\sqrt{n})}.$$

Let $h(s) = \log M_{Y_i}(s)$. Taylor expansion about $s = 0$ gives:

$$h(s) = h(0) + h'(0) \cdot s + \frac{h''(0)}{2} \cdot s^2 + R(s),$$

with $\lim_{s\to 0} \frac{R(s)}{s^2} = 0$.

$$h(0) = \log M_{Y_i}(0) = \log \mathbb{E}(e^{0 \cdot Y_i}) = \log(1) = 0,$$

$$h'(0) = \frac{M'_{Y_i}(0)}{M_{Y_i}(0)} = \frac{\mathbb{E}(Y_i)}{1} = 0,$$

$$h''(0) = \frac{M''_{Y_i}(0) \cdot M_{Y_i}(0) - (M'_{Y_i}(0))^2}{(M_{Y_i}(0))^2} = \frac{1 \cdot 1 - 0^2}{1^2} = 1.$$

We thus have

$$\log M_{Y_i}(s) = h(s) = \frac{s^2}{2} + R(s)$$
$$\Longrightarrow \log M_{Y_i}(t/\sqrt{n}) = h(t/\sqrt{n}) = \frac{t^2}{2n} + R(t/\sqrt{n}).$$

Hence,

$$e^{n \cdot \log M_{Y_i}(t/\sqrt{n})} = e^{\frac{t^2}{2} + n \cdot R\left(\frac{t}{\sqrt{n}}\right)}.$$

We have $\lim_{n\to\infty} \frac{R(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0$, so

$$\lim_{n\to\infty} n \cdot \frac{R(t/\sqrt{n})}{t^2} = 0,$$

so

$$\lim_{n\to\infty} n \cdot R\left(\frac{t}{\sqrt{n}}\right) = 0.$$

Hence,

$$\lim_{n\to\infty} M_{\sqrt{n}\cdot\frac{\overline{X}_n-\mu}{\sigma}}(t) = \underbrace{e^{t^2/2}}_{\text{mgf of } \mathcal{N}(0,1)}$$

■

Normal approximation to binomial

When n is large and p is not too close to 0 or 1, we have the approximation

$$\text{Bin}(n,p) \approx \mathcal{N}(np, np(1-p)). \qquad (\clubsuit)$$

R Recall that np is the expectation of Bin(n, p) and $np(1-p)$ is the variance of Bin(n,p).

R Bin(n, p) is discrete and $\mathcal{N}(np, np(1-p))$ is continuous. So how does the approximation work? We use it as follows:

$$X \sim \text{Bin}(n,p), \quad Y \sim \mathcal{N}(np, np(1-p)), \quad k \in \{0,1,\dots,n\}$$

$$\implies \mathbb{P}(X=k) \approx \mathbb{P}\left(k-\frac{1}{2} \le Y \le k+\frac{1}{2}\right) = \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} f_Y(y)\,\mathrm{d}y.$$

and similarly

$$\mathbb{P}(X \le b) \approx \int_{-\infty}^{b+\frac{1}{2}} f_Y(y)\,\mathrm{d}y = F_Y(b+\tfrac{1}{2}), \qquad \mathbb{P}(X \ge a) \approx \int_{a-\frac{1}{2}}^{\infty} f_Y(y)\,\mathrm{d}y = 1-F_Y(a-\tfrac{1}{2}).$$

R We will use this approximation when

$$n \ge 15 \qquad np \ge 5, \qquad n(1-p) \ge 5.$$

Question 13.3 A student takes an exam with 50 questions. She gets the answer to each question correctly with probability $\frac{2}{3}$, independently of all others. What is the probability that her grade is below 60%?

Answer: 60% corresponds to 30 correct answers.

▷ **Exact answer:** Let X be the number of correct answers given by the student; then, $X \sim \text{Bin}(50, \frac{2}{3})$, so

$$\mathbb{P}(X<30) = \sum_{k=0}^{30} \binom{50}{k} \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{50-k}.$$

▷ **Solution with approximation:** We have $np = \frac{100}{3}$ and $np(1-p) = \frac{100}{9}$, so X is approximated by $Y \sim \mathcal{N}\left(\frac{100}{3}, \frac{100}{9}\right)$. We have

$$\mathbb{P}(Y \le 29.5) = \mathbb{P}\left(Z \le \frac{29.5-100/3}{\sqrt{100/9}}\right) = \mathbb{P}(Z<-1.15) = F_Z(-1.15) \approx 12.51\%.$$

For the last approximation we used the Table 13.3 as follow:

$$F_Z(-1.15) = 1 - F_Z(1.15) \approx 1 - 0.87493 = 0.12507 \approx 12.51\%.$$

■

■ **Example 13.1** Let's reconsider the (idealized) Galton board with n rows of nails and, for simplicity, let us assume there are exactly $n+1$ partitions. We label these partitions $0,\dots,n$ (left to right). Let us suppose further we drop N balls (N very large). As argued before the number of times an individual ball moves right is described by a binomial distribution with parameters $n, 1/2$. So a ball will fall for instance in the leftmost partition (labeled 0) if the were 0 right-moves, if the middle partition if there were exactly $n/2$ right-moves and in the rightmost partition if there were exactly n right-moves. We assume the balls do not influence each other (this can be achieved if we have the patience to only let a ball pass through the hole once the previous ball has stopped moving). Let $X_1,\dots,X_N$ be i.i.d. $\sim$ Bin(n,1/2) be the number of right-moves of the N balls.

For $1 \le j \le N, 0 \le i \le n$ denote:

$$Y_{i,j} := 1_{\{X_j=i\}} = \begin{cases} 1 & \text{if ball } j \text{ lands in partition } i, \\ 0 & \text{otherwise.} \end{cases}.$$

The the heights $H_1,\dots H_n$ of the partitions are given by:

$$H_i = \sum_{j=1}^{N} Y_{i,j} = |\{j : \text{ ball } j \text{ lands in the } i\text{-th partition }\}|.$$

Now we fix some $0 \le i \le n$, then $Y_{i,1},\dots,Y_{i,N}$ are i.i.d $\sim \text{Ber}(p_i)$ with $p_i = \mathbb{P}(X_1 = i) = \cdots = \mathbb{P}(X_N = i)$. By the law of large numbers, for any $\varepsilon > 0$:

$$\mathbb{P}\left((p-\varepsilon)N \le H_i \le (p+\varepsilon)N\right) \overset{N\to\infty}{\longrightarrow} 1.$$

So the partition heights will be proportional to $p_i = \mathbb{P}(\text{Bin}(n,1/2) = i)$. (Not only is the expectation $\mathbb{E}H_i = pN$, but the actual number H_i is overwhelmingly likely to be close to pN.)

How do we explain that the heights seem to follow a normal curve?

This is just the normal approximation to the binomial / the CLT: for n large we have $\text{Bin}(n,1/2) \overset{d}{\approx} \mathcal{N}(n/2, n/4)$. Therefore, for a large number n of rows:

$$p_0 + \cdots + p_i \approx \mathbb{P}(\mathcal{N}(n/2,n/4) \le i)$$

and so we expect that p_i is close to the pdf of the $\mathcal{N}(n/2,n/4)$ distribution evaluated at $x = i$. ■

13.3 Standard normal distribution table

z	0	0,01	**0,02**	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,5279	0,53188	0,53586
0,1	0,53983	0,5438	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,6293	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,6591	0,66276	0,6664	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,7054	0,70884	0,71226	0,71566	0,71904	0,7224
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,7549
0,7	0,75804	0,76115	0,76424	0,7673	0,77035	0,77337	0,77637	0,77935	0,7823	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,8665	0,86864	0,87076	0,87286	0,87493	0,87698	0,879	0,881	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,9032	0,9049	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	**0,9222**	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,9452	0,9463	0,94738	0,94845	0,9495	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,9608	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,9732	0,97381	0,97441	0,975	0,97558	0,97615	0,9767
2	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,9803	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,983	0,98341	0,98382	0,98422	0,98461	0,985	0,98537	0,98574
2,2	0,9861	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,9884	0,9887	0,98899
2,3	0,98928	0,98956	0,98983	0,9901	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,9918	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,9943	0,99446	0,99461	0,99477	0,99492	0,99506	0,9952
2,6	0,99534	0,99547	0,9956	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,9972	0,99728	0,99736
2,8	0,99744	0,99752	0,9976	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,999
3,1	0,99903	0,99906	0,9991	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,9994	0,99942	0,99944	0,99946	0,99948	0,9995
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,9996	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,9997	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,9998	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,9999	0,9999	0,9999	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

Table 13.1: CDF of a standard normal distribution.
Example: the value in the row "1,4" and column "0,02" gives the approximation $F_Z(1.42) \simeq 0.9222$.

13.4 Exercises

13.4.1 Problems

Exercise 13.1 Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables, each with pdf $f(x) = 2|x|^{-5}\mathbb{1}_{\mathbb{R}\setminus[-1,1]}(x)$. Find a sequence $(a_n)_{n\in\mathbb{N}}$ such that

$$\frac{\sum_{i=1}^{n} X_i}{a_n} \xrightarrow[d]{n\to\infty} Z \qquad \text{where} \qquad Z \sim \mathcal{N}(0,1).$$

Reminder: $\xrightarrow[d]{n\to\infty}$ *refers to convergence in distribution.* ■

Exercise 13.2 You decide to go to the casino to play Roulette. 17 is your favorite number and therefore, every round you put some of your money solely on 17. From your probability course you know that the amount of games you have to play until you win once follows a geometric distribution with parameter $1/37$. Estimate the probability that after $19 \cdot \mathbb{E}[X]$ games, where $X \sim \mathsf{Geo}(1/37)$, you already won 37 times or even more than that.
Remark: You won't need a calculator for this. The numbers are chosen in a way to guarantee that a lot cancels out in the respective fractions. ■

Trivia

George Pólya (1887 – 1985) was a Hungarian mathematician who made contributions to various areas of math including probability theory. One of his books is called *how to solve it*. Many first year students have questions like “I understand the proof but how on earth did you come up with it?”. The aforementioned book tries to answer that question and teaches you various “heuristic” methods to find or guess the solution to math problems. It is very suitable for you in the current stage of your math career and I would strongly recommend to read it.

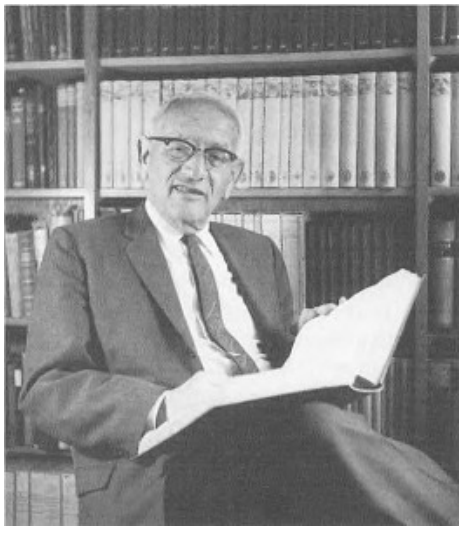


Figure 13.1: George Pólya

In the 1920s, when Pólya was a student in Budapest, he rented a room in a house managed by a landlady. One day Pólya went out for a walk at the same time that one of his housemates was having a walk with his fiancée. During the walk Pólya kept running into them on the streets of Budapest. This caused him some embarrassment as he was worried the couple might think he was

spying on them.

After he returned to the house, he invented a math problem so that he could justify running into them so often. Let us consider the two-dimensional integer lattice $\mathbb{Z}^2$ (which represents a simplified map of Budapest). Two particles (representing Pólya, respectively the couple) both start at the origin and keep moving to a randomly chosen neighbour of the point they're currently (here a neighbour is a point at distance one).

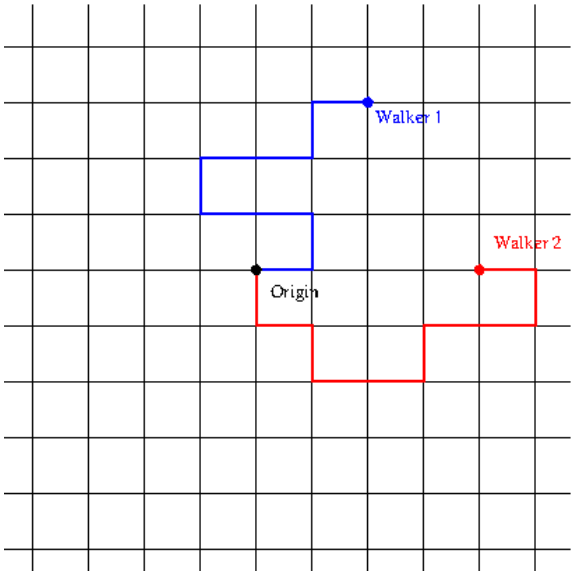


Figure 13.2: Two random walkers on the integer grid.

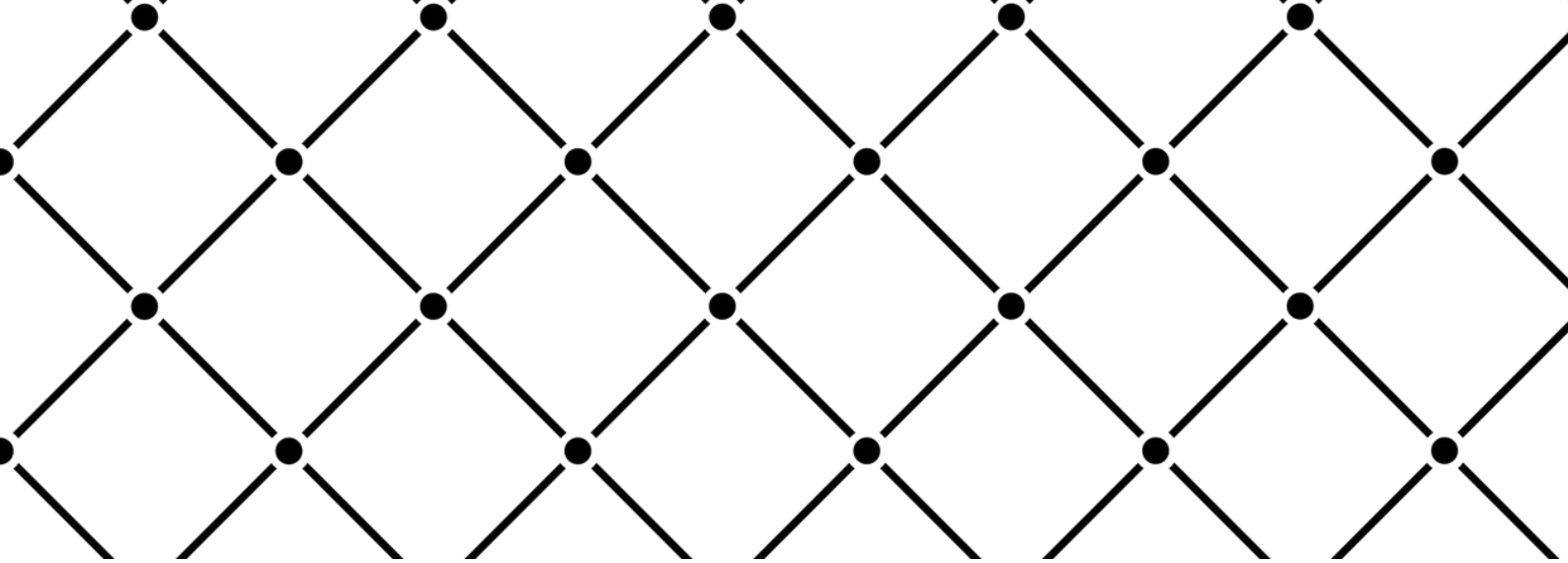
He now wished to determine what the probability is of the two random walkers ever meeting.

To distinguish the particles let us say one of them is red, the other blue. Let R_t, B_t denote the position of these particles at time $t = 0, 1, 2, \ldots$. Suppose the red particle moves on even timesteps and the blue particle on odd timesteps. If we now consider the difference vector $D_t := R_t - B_t$ then we see that $D_t = \underline{0}$ iff. the particles are on the same point of $\mathbb{Z}^2$. Moreover, D_t follows a "symmetric random walk" – meaning that at every timestep D_t moves to one of the four points of $\mathbb{Z}^2$ closest to the current location. For obvious reasons the symmetric random walk is sometimes also called the "drunkard's walk".

We have transformed the problem of two particles meeting into the problem of whether (or: with what probability) a single particle performing a random walk will re-visit the origin.

We can generalise the question to arbitrary dimension d: a particle moves on $\mathbb{Z}^d$, starting at the origin and always moving to one of the $2d$ closest vertices of the d-dimensional integer grid (chosen uniformly at random). What is the probability that the particle ever visits the origin again? In 1920s, Pólya proved the answer equals 1 if $d = 1$ or $d = 2$ but < 1 if $d \geq 3$. This can be paraphrased as "a drunkard will eventually find his way home in a (2-dimensional) city or if he's moving along a (1-dimensional) street, but a drunken astronaut may get lost in space forever." This also vindicates Pólya: if we assume the map of Budapest is like the integer grid and he and the couple behave like random particles then they are bound to meet.

The proof of Pólya's theorem is too complicated for this course, but it can be a motivation for you to take the course "stochastic processes" in which Pólya's theorem will be treated in full. It has a cool proof that makes use of the theory of electricity (from physics).



14. Random walks

14.1 Random walks and gambler's ruin

14.1.1 Random walks on the integers

The **symmetric random walk** on the integers $\mathbb{Z}$ is essentially the same as the betting game setup we've discussed a few times already. Let X_t denote the step the particle makes at time t. Then $X_1, X_2, \ldots$ are i.i.d. with

$$X_t = \begin{cases} +1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

And, the position at time t is then

$$S_t := \sum_{s=1}^{t} X_s.$$

Note that $\mathbb{E}S_t = t \cdot \mathbb{E}X_1 = 0$. By the LLN we moreover have

$$\mathbb{P}(-\varepsilon t < S_t < \varepsilon t) \xrightarrow{t\to\infty} 1 \tag{14.1}$$

(for every fixed $\varepsilon > 0$). So the random walk typically stays relatively close to its starting point.

The CLT allows us to quantify this more precisely. We have $\mathsf{Var}(S_t) = t \cdot \mathsf{Var}(X_1) = t$ and hence, by the CLT, $S_t \overset{d}{\approx} \mathcal{N}(0,t)$. This allows us to derive for instance that

$$\mathbb{P}(-\sqrt{t} < S_t < \sqrt{t}) \xrightarrow{t\to\infty} \mathbb{P}(-1 < Z < 1) = \Phi(1) - \Phi(-1),$$

where $Z \sim \mathcal{N}(0,1), \Phi = F_Z$ as usual. The number $\Phi(1) - \Phi(-1)$ is a constant and neither zero or one, so S_t has a decent chance of being less than $\sqrt{t}$ in absolute value, and also a decent chance of being more than $\sqrt{t}$ in absolute value.

It is also natural to consider **non-symmetric random walks**, where the probability of moving right is p and the probability of moving left is $1-p$. Let us defined $X_1, X_2, \ldots$ and $S_1, S_2, \ldots$ as

before, mutatis mutandis (Latin: "changing what needs to be changed"). As you may remember, in this case $\mathbb{E}X_i = 2p-1, \mathrm{Var}X_i = 4p(1-p)$. Now the LLN says:

$$\mathbb{P}((2p-1-\varepsilon)t < S_t < (2p-1+\varepsilon)t) \overset{t\to\infty}{\longrightarrow} 1. \tag{14.2}$$

(for every fixed $\varepsilon > 0$). So the random walker is in a certain sense either moving towards $+\infty$ (when $p > \frac{1}{2}$) or $-\infty$ (when $p < \frac{1}{2}$), with spead $|2p-1|$. Applying the CLT in a manner similar to what we did for the symmetric case ($p = \frac{1}{2}$) that the random walker has a decent chance of being within $\sqrt{t}$ of $(2p-1)t$ and a decent chance of being further away than $\sqrt{t}$ of $(2p-1)t$.

It is only natural to expect that: "if $p = \frac{1}{2}$ then the walker will return to the origin with probability one and if $p \neq \frac{1}{2}$ then there is a nonzero probability the walker will disappear to $-\infty$ or $+\infty$ without ever return to the origin." (The first statement is just the $d = 1$ case of Pólya's theorem.)

Now note that while (14.2) implies that $\mathbb{P}(S_t = 0) \to 0$ if $p \neq \frac{1}{2}$, this does not exclude the possibility that the random walker might return to the origin with probability one. If we consider the infinite sequence $S_1, S_2, \ldots$ then it might be the case that there is always an "exceptional time" t when $S_t = 0$ (even though the marginal probability of $S_t = 0$ for any predetermined time t is close to zero.)

Similarly, (14.1) does not tell us that we will necessarily always find some t with $S_t = 0$.

14.1.2 Gambler's ruin

A gambler has a starting capital of i€. He repeatedly plays a game where he bets 1€ each time. With probability p he wins and he receives his money back plus an additional 1€, and with probability $1-p$ he loses his 1€.

If his capital reaches 0€ he is of course no longer able to play. And, if he manages to gather a capital of N€, he will go home contently. (And pays his university fees of N€ the next day.)

What is the chance the gambler will go home contently?

To answer this question, we define the events

$$H := \{\text{the gambler goes Home Happy}\},$$
$$W := \{\text{the gambler Wins the first game}\},$$

and set

$$p_i := \mathbb{P}_i(H) \quad (i = 0, \ldots, N).$$

Here the subscript i denotes that the gambler starts with a capital of i€. We see immediately that

$$p_0 = 0, \quad p_N = 1.$$

We further remark that, for $0 < i < N$,

$$\mathbb{P}_i(H|W) = \mathbb{P}_{i+1}(H), \quad \mathbb{P}_i(H|W^c) = \mathbb{P}_{i-1}(H).$$

(If he wins the first game then the situation is as if he's not played any game yet but has a starting capital of $i+1$€ and similarly if he loses, the situation is as if he's not played any game but has a starting capital of $i-1$€.)
Also note that, for $0 < i < N$,

$$\mathbb{P}_i(H) = \mathbb{P}_i(H|W)\mathbb{P}_i(W) + \mathbb{P}_i(H|W^c)\mathbb{P}_i(W^c).$$

We've therefore derived the following relation for the p_i-s:

$$p_i = (1-p)\cdot p_{i-1} + p\cdot p_{i+1}, \quad \text{for } i = 1,\ldots,N-1. \tag{14.3}$$

Rewriting gives:

$$(1-p)\cdot p_i - (1-p)\cdot p_{i-1} = p\cdot p_{i+1} - p\cdot p_i,$$

In other words:

$$p_{i+1} - p_i = \left(\frac{1-p}{p}\right)(p_i - p_{i-1}). \tag{14.4}$$

Repeatedly applying (14.4) gives:

$$\begin{aligned} p_{i+1} - p_i &= \left(\tfrac{1-p}{p}\right)(p_i - p_{i-1}) \\ &= \left(\tfrac{1-p}{p}\right)^2 (p_{i-1} - p_{i-2}) \\ &= \left(\tfrac{1-p}{p}\right)^3 (p_{i-2} - p_{i-3}) \\ &\vdots \\ &= \left(\tfrac{1-p}{p}\right)^i (p_1 - p_0) \\ &= \left(\tfrac{1-p}{p}\right)^i p_1 \end{aligned}$$

(in the last step we use $p_0 = 0$.)

On the other hand:

$$\begin{aligned} p_{i+1} &= p_{i+1} - p_0 \\ &= (p_{i+1} - p_i) + (p_i - p_{i-1}) + \cdots + (p_1 - p_0) \\ &= \left(\tfrac{1-p}{p}\right)^i p_1 + \left(\tfrac{1-p}{p}\right)^{i-1} p_1 + \cdots + \left(\tfrac{1-p}{p}\right) p_1 + p_1. \end{aligned}$$

In the special case that $p = \frac{1}{2}$ we have $(1-p)/p = 1$. So then:

$$p_{i+1} = (i+1)p_1.$$

Because we need to have that $1 = p_N = Np_1$ it follows that $p_1 = 1/N$ and hence also:

$$p_i = \frac{i}{N} \quad \text{for } i = 0,\ldots,N.$$

(if $p = \frac{1}{2}$.)

But what if $p \neq \frac{1}{2}$?

In this case the formula for summing the first i terms of a geometric series shows that:

$$p_{i+1} = \frac{1 - \left(\frac{1-p}{p}\right)^{i+1}}{1 - \left(\frac{1-p}{p}\right)} \cdot p_1.$$

(Notice that we cannot apply this formula when $p = 1/2$ because the denominator would be zero then.) From

$$1 = p_N = \frac{1 - \left(\frac{1-p}{p}\right)^{N}}{1 - \left(\frac{1-p}{p}\right)} \cdot p_1,$$

we find that

$$p_1 = \left(1 - \left(\frac{1-p}{p}\right)\right) / \left(1 - \left(\frac{1-p}{p}\right)^N\right).$$

So for general $0 \leq i \leq N$ we have:

$$p_i = \frac{1 - \left(\frac{1-p}{p}\right)^i}{1 - \left(\frac{1-p}{p}\right)^N}.$$

(Provided $p \neq \frac{1}{2}$.)

14.1.3 Tying together the gamblers ruin and random walks.

We are going to use the observations on the gambler's to prove:

Theorem 14.1.1 Consider the random walk with probability p of moving right. Then

$$\mathbb{P}(\text{The random walk will revisit the origin}) = \mathbb{P}(\exists t > 0 : S_t = 0) = \begin{cases} 1 & \text{if } p = \frac{1}{2}, \\ 2(1-p) & \text{if } p > \frac{1}{2}, \\ 2p & \text{if } p < \frac{1}{2}. \end{cases}$$

The proof will rely on the following observation:

Lemma 14.1.2 In the gambler's ruin, the probability that the gambler keeps playing indefinitely, without ever reaching 0€ or N€, equals zero.

Proof. We leave the proof as an exercise. ■

Proof of the theorem in the case $p = \frac{1}{2}$. We first consider the situation when the walker starts on the point 1 rather than 0. To emphasize this alternative situation, we use the notation $\mathbb{P}_1(.)$. Let us write

$$E := \{\text{The walker will visit } 0\},$$
$$E_N := \{\text{The walker will reach } 0 \text{ before it reaches } N\}.$$

Then we have that (when the walker starts on 1)

$$E = \bigcup_N E_N, \quad E_1 \subseteq E_2 \subseteq E_3 \subseteq \ldots$$

To clarify the first equality a bit more: note that if the walkers visits 0 at some finite time t then there will be some N that the walker has not yet visited (any $N > t$ will do). By the lemma and the observations on the gambler's ruin we know:

$$\mathbb{P}_1(E_N) = \mathbb{P}(\text{gambler with starting capital 1 reaches 0 before } N) = 1 - \frac{1}{N}.$$

Thus,

$$\mathbb{P}_1(E) = \lim_{N\to\infty} \mathbb{P}(E_N) = 1.$$

Similarly, in the case the walker starts on -1 we have $\mathbb{P}_{-1}(E) = 1$. And then also

$$\mathbb{P}(E) = \mathbb{P}(X_1 = 1)\mathbb{P}(E|X_1 = 1) + \mathbb{P}(X_1 = -1)\mathbb{P}(E|X_1 = -1) = \frac{1}{2}\mathbb{P}_{-1}(E) + \frac{1}{2}\mathbb{P}_1(E) = 1.$$

This concludes the proof when $p = \frac{1}{2}$. ■

Proof of the theorem in the case when $p \neq \frac{1}{2}$. By symmetry (swapping left and right, and p and $1-p$), we only need to prove this for the case when $p > \frac{1}{2}$. We let E, E_N be as above. Assuming first the walker starts on $+1$, we see that

$$\mathbb{P}_1(E_N) = \mathbb{P}(\text{gambler with starting capital 1 reaches 0 before } N) = 1 - \frac{1 - \left(\frac{1-p}{p}\right)}{1 - \left(\frac{1-p}{p}\right)^N},$$

and

$$\mathbb{P}_1(E) = \lim_{N\to\infty} \mathbb{P}_1(E_N) = \frac{1-p}{p}.$$

Now suppose the walker starts on -1. Defining E_{-N} analogously to E_N we have (when we start on -1):

$$E = \bigcup_N E_{-N}, \quad E_{-1} \subseteq E_{-2} \subseteq E_{-3} \subseteq \ldots,$$

giving

$$\mathbb{P}_{-1}(E) = \lim_{N\to\infty} \mathbb{P}_{-1}(E_{-N}).$$

Also note that $\mathbb{P}_{-1}(E_{-N})$ is the probability a gambler with starting fortune $N-1$€ reaches N€ before 0€, with is probability of winning at each round is p. So,

$$\mathbb{P}_{-1}(E_{-N}) = \frac{1 - \left(\frac{1-p}{p}\right)^{N-1}}{1 - \left(\frac{1-p}{p}\right)^N},$$

and

$$\mathbb{P}_{-1}(E) = \lim_{N\to\infty} \mathbb{P}_{-1}(E_{-N}) = 1.$$

Putting everything together:

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(X_1 = 1)\mathbb{P}(E|X_1 = 1) + \mathbb{P}(X_1 = -1)\mathbb{P}(E|X_1 = -1)\\ &= p\mathbb{P}_1(E) + (1-p)\mathbb{P}_{-1}(E)\\ &= 2(1-p).\end{aligned}$$

■

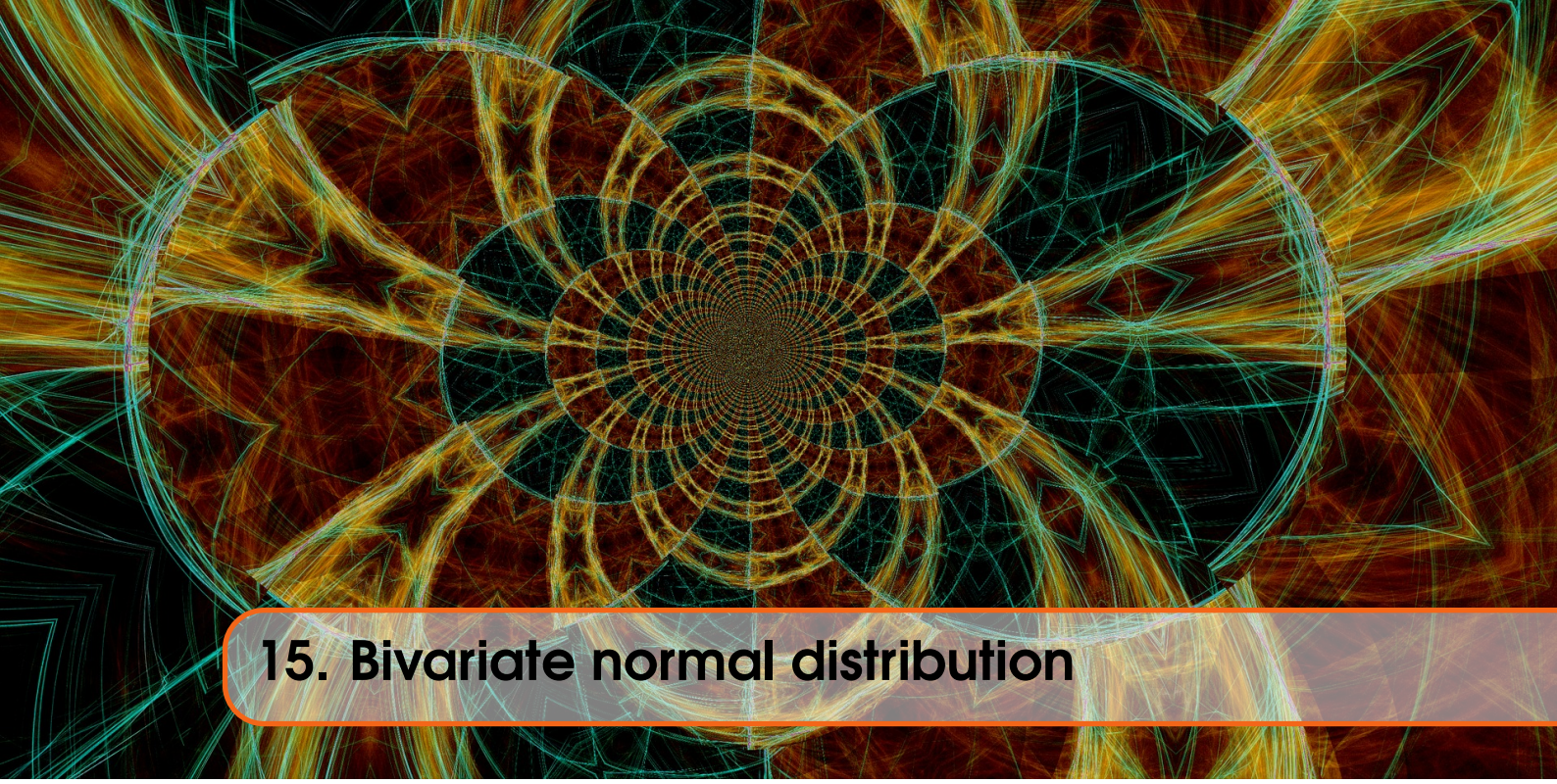
Finally we remark that this last theorem also implies

Corollary 14.1.3

$$\mathbb{P}(\text{ the walker visits the origin infinitely many times }) = \begin{cases} 1 & \text{if } p = \frac{1}{2},\\ 0 & \text{otherwise.}\end{cases}.$$

Proof. This is a nice challenge for you to find how this is indeed a corollary of Theorem 14.1.1. ■

So had Pólya and the couple been walking along a street they would be meeting infinitely many times.



15. Bivariate normal distribution

The goals of this chapters are:

- ▷ to define the **bivariate normal distribution**;
- ▷ to identify **linear combinations of (bivariate) normal random variables** as (bivariate) normal random variable with appropriate parameters.

15.1 The bivariate normal distribution

Recall that $X \sim \mathcal{N}(\mu, \sigma^2)$ has pdf $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and mgf $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$.

Definition 15.1.1 — Bivariate normal distribution. A random vector (X,Y) is said to follow a **bivariate normal distribution** with parameters $\mu_X \in \mathbb{R}$, $\mu_Y \in \mathbb{R}$, $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$ and $\rho \in (-1,1)$ if

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}\right)\right\}.$$

We write: $(X,Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$

Note that if $\rho = 0$ and (X,Y) is as above, then its joint pdf has the form

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \times \exp\left\{-\frac{1}{2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sigma_X} \times \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}}_{f_X(x)} \times \underbrace{\frac{1}{\sqrt{2\pi}\sigma_Y} \times \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}}_{f_Y(y)}. \end{aligned}$$

We recognize the pdf's of univariate normally distributed random variables. Therefore, in that case, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and X and Y are independent.

What happens in the other cases ($\rho \neq 0$)? Proposition 15.1.2 will give an answer to this question. But, first we need the following lemma.

Lemma 15.1.1 — Linear combination. If $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}(0,1)$ are independent and

$$U = \sigma_1 Z_1 + \mu_1, \qquad V = \rho\sigma_2 Z_1 + \sqrt{1-\rho^2}\cdot\sigma_2 Z_2 + \mu_2,$$

then $(U,V) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Proof. Let

$$g(z_1, z_2) = (\sigma_1 z_1 + \mu_1,\ \rho\sigma_2 z_1 + \sqrt{1-\rho^2}\cdot\sigma_2 z_2 + \mu_2),$$

so that $(U,V) = g(Z_1, Z_2)$. We have

$$g(z_1, z_2) = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{1-\rho^2}\cdot\sigma_2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

Let us find $\boxed{h = g^{-1}}$. For this, recall that a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is invertible if and only if $\det(A) = ad - bc \neq 0$, and in that case,

$$A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Using this, we find

$$\begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{1-\rho^2}\cdot\sigma_2 \end{pmatrix}^{-1} = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \begin{pmatrix} \sqrt{1-\rho^2}\cdot\sigma_2 & 0 \\ -\rho\sigma_2 & \sigma_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ -\frac{\rho}{\sigma_1\cdot\sqrt{1-\rho^2}} & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix}.$$

Hence,

$$\begin{aligned} h(u,v) &= \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ -\frac{\rho}{\sigma_1\cdot\sqrt{1-\rho^2}} & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix} \begin{pmatrix} u-\mu_1 \\ v-\mu_2 \end{pmatrix} \\ &= \left(\frac{1}{\sigma_1}(u-\mu_1),\ -\frac{\rho}{\sigma_1\sqrt{1-\rho^2}}(u-\mu_1) + \frac{1}{\sigma_2\sqrt{1-\rho^2}}(v-\mu_2) \right), \end{aligned}$$

so

$$J(u,v) = \det \begin{pmatrix} \partial h_1/\partial u & \partial h_1/\partial v \\ \partial h_2/\partial u & \partial h_2 \partial v \end{pmatrix} = \det \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ -\frac{\rho}{\sigma_1\cdot\sqrt{1-\rho^2}} & \frac{1}{\sigma_2\sqrt{1-\rho^2}} \end{pmatrix} = \frac{1}{\sigma_1\sigma_2\sqrt{1-\sigma^2}}.$$

Then,

$$\begin{aligned} f_{U,V}(u,v) &= f_{Z_1,Z_2}(h(u,v)) \cdot |J(u,v)| \\ &= \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot f_{Z_1}(h_1(u,v)) \cdot f_{Z_2}(h_2(u,v)) \\ &= \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} \cdot (h_1(u,v)^2 + h_2(u,v)^2) \right\}. \end{aligned} \tag{♣}$$

Note that

$$h_1(u,v)^2+h_2(u,v)^2=\frac{(u-\mu_1)^2}{\sigma_1^2}+\frac{\rho^2(u-\mu_1)^2}{\sigma_1^2(1-\rho^2)}+\frac{(v-\mu_2)^2}{\sigma_2^2(1-\rho^2)}-2\rho\frac{(u-\mu_1)(v-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)}.$$

Plugging this back in (♣), we see that $f_{U,V}$ has the required form. ■

Proposition 15.1.2 If $(X,Y)\sim\mathcal{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$, then

$$X\sim\mathcal{N}(\mu_X,\sigma_X^2),\qquad Y\sim\mathcal{N}(\mu_Y,\sigma_Y^2),\qquad \rho_{X,Y}=\rho.$$

Proof. By Lemma 15.1.1, if $Z_1,Z_2\sim\mathcal{N}(0,1)$ are independent and

$$X=\sigma_X Z_1+\mu_X,\qquad Y=\rho\sigma_Y Z_1+\sqrt{1-\rho^2}\sigma_Y Z_2+\mu_Y,$$

then $(X,Y)\sim\mathcal{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$. By Proposition 11.1.5,

$$X\sim\mathcal{N}(\mu_X,\sigma_X^2),\qquad Y\sim\mathcal{N}\left(\mu_Y,\ (\rho\sigma_Y)^2+(\sqrt{1-\rho^2}\sigma_Y)^2\right)=\mathcal{N}(\mu_Y,\sigma_Y^2).$$

Moreover, as $\mathsf{Cov}(Z_1,Z_1)=\mathsf{Var}(Z_1)=1$ and $\mathsf{Cov}(Z_1,Z_2)=0$,

$$\begin{aligned}\mathsf{Cov}(X,Y)&=\mathsf{Cov}(\sigma_X Z_1+\mu_X,\ \rho\sigma_Y Z_1+\sqrt{1-\rho^2}\sigma_Y Z_2+\mu_Y)\\&=\sigma_X\rho\sigma_Y\mathsf{Cov}(Z_1,Z_1)+\sigma_X\sqrt{1-\rho^2}\sigma_Y\mathsf{Cov}(Z_1,Z_2)=\sigma_X\sigma_Y\rho.\end{aligned}$$

Hence, $\rho_{X,Y}=\mathsf{Cov}(X,Y)/(\sigma_X\sigma_Y)=\rho$. ■

Similarly, we can also prove

Corollary 15.1.3 If $(X,Y)\sim\mathcal{N}(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$, then

$$aX+bY\sim\mathcal{N}(a\mu_X+b\mu_Y,\ a^2\sigma_X^2+b^2\sigma_Y^2+2ab\rho\sigma_X\sigma_Y).$$

Proof. As before, we let $Z_1,Z_2\sim\mathcal{N}(0,1)$ be independent and let

$$X=\sigma_X Z_1+\mu_X,\qquad Y=\rho\sigma_Y Z_1+\sqrt{1-\rho^2}\sigma_Y Z_2+\mu_Y.$$

Then, by Proposition 11.1.5,

$$\begin{aligned}&aX+bY\\&=(a\sigma_X+b\rho\sigma_Y)Z_1+b\sqrt{1-\rho^2}\sigma_Y Z_2+a\mu_X+b\mu_Y\\&\sim\mathcal{N}((a\sigma_X+b\rho\sigma_Y)\cdot 0+b\sqrt{1-\rho^2}\sigma_Y\cdot 0+a\mu_X+b\mu_Y,(a\sigma_X+b\rho\sigma_Y)^2\cdot 1+(b\sqrt{1-\rho^2}\sigma_Y)^2\cdot 1)\\&\sim\mathcal{N}(a\mu_X+b\mu_Y,\ a^2\sigma_X^2+b^2\sigma_Y^2+2ab\rho\sigma_X\sigma_Y).\end{aligned}$$

■

15.2 Exercises

15.2.1 Problems

Exercise 15.1 Assume (X,Y) is a random vector following a bivariate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2 and $\rho \in (-1,1)$. Let

$$\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Find the correlation between Z and W as a function of the parameters of (X,Y). ■

Exercise 15.2 Let (X,Y) follow a bivariate normal distribution with parameters $\mu_1,\mu_2,\in \mathbb{R}$, $\sigma_1^2,\sigma_2^2 > 0$, $\rho \in (-1,1)$. Show that

$$(U,V) := ((X-\mu_1)/\sigma_1, (Y-\mu_2)/\sigma_2)$$

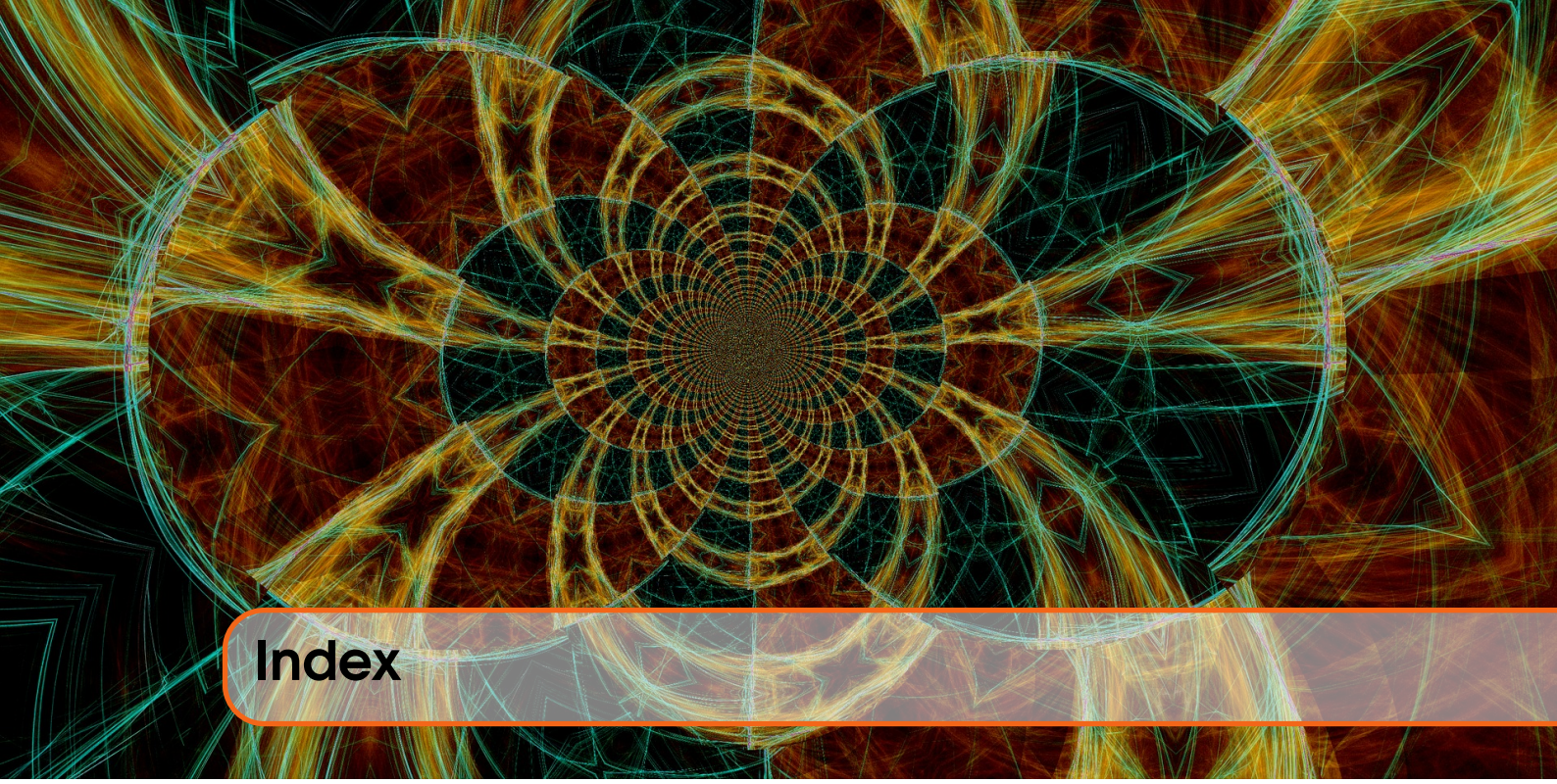
follows a bivariate standard normal distribution. ■

Exercise 15.3 Let $X = (X_1,X_2)$ be a random vector. The *covariance matrix* $\Sigma \in \mathbb{R}^{2\times 2}$ of X is defined by $\Sigma_{ij} = \mathrm{Cov}(X_i,X_j)$, where $1 \le i,j \le 2$. An alternative definition of the bivariate normal distribution can be achieved via the covariance matrix. $X = (X_1,X_2)$ follows a bivariate normal distribution with parameters $\mu = (\mu_1,\mu_2)$ and Σ if

$$f_X(x) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right), \quad x \in \mathbb{R}^2.$$

i) Show that this definition coincides with the definition from the lecture.
ii) Let $A \in \mathbb{R}^{2\times 2}$ be matrix with full rank (i.e., it is invertible) and with $\det A \neq 0$ and let $b \in \mathbb{R}^2$. Show that $Y := AX + b$ also follows a bivariate normal distribution with parameters $\tilde{\mu} = A\mu + b$ and $A\Sigma A^T$.

■



Index

I

J

L

M

N

O

P

R

S

U

V

W