Tabea Rebafka

September 2022

Acknowledgement

These notes are based on the lecture notes by **Lucas Gerin** (Département de Mathématiques Appliquées, École Polytechnique) who taught this course until 2017.

Tabea Rebafka

Contents

Chapter 1

Probability Spaces

Probability theory is the mathematical science of quantifying uncertainty of random events. While we are unable to exactly predict the outcome of flipping a coin, the weather tomorrow, the winner of the next elections or the evolution of the consumer preferences, probability theory aims at determining (at least) probabilities of rain, vote, consumer behaviour and so on, so that we are able to say that some events are more likely than others. That is, probability theory is about quantifying the probability of possible events.
Now mathematical statistics and modern data science built on probabilistic frameworks to analyze and understand data and make predictions and recommandations in all fields of application like medicine, marketing, information science, economy, sociology, communication and many others. To properly understand or even develop new methods and algorithms in data science and machine learning, a solid mathematical background in probability theory is essential. So the goal of this course is to present at least some of the foundations of probability theory.

1.1 Fundamental notions from set theory

First we fix some notation. In mathematics, **sets** of elements are usually denoted with capital letters $A, B, \ldots$. We denote the **complement** of set A by

$$A^c \text{ (or } \bar{A}) = \{x \text{ such that } x \notin A\},$$

the **union** of sets A and B by

$$A \cup B = \{x \text{ such that } x \in A \text{ or } x \in B\},$$

the **intersection** of sets A and B by

$$A \cap B = \{x \text{ such that } x \in A \text{ and } x \in B\},$$

and the **difference** of sets A and B, also called relative complement of A with respect to B, by

$$A \backslash B = A \cap B^c.$$

Let $(A_n)_{n\geq 1}$ be a sequence of sets, then

$$\bigcup_{n\geq 1} A_n = \{x \text{ such that } x \in A_n \text{ for } \textbf{at least} \text{ one } n\}.$$
$$\bigcap_{n\geq 1} A_n = \{x \text{ such that } x \in A_n \text{ for } \textbf{every } n\}.$$

According to **De Morgan's law**, for countable unions or intersections of sets, we have

$$\left(\bigcup_{n\geq 1} A_n\right)^c = \bigcap_{n\geq 1} A_n^c \quad \text{and} \quad \left(\bigcap_{n\geq 1} A_n\right)^c = \bigcup_{n\geq 1} A_n^c.$$

By analogy with **products of sets**

$$A \times B = \{(a, b) \text{ such that } a \in A \text{ and } b \in B\},$$

we write $A^{\mathbb{N}}$ for the **set of infinite sequences** $(a_1, a_2, a_3, \dots)$ with elements a_n in A. That is,

$$A^{\mathbb{N}} = \{(a_1, a_2, a_3, \dots) \text{ such that } a_n \in A \text{ for all } n\}.$$

1.2 Probability spaces

Definition 1.2.1 (Sample space, event)**.** *A sample space represents the set of all possible outcomes of a random experiment. A* **sample space** *is any finite or infinite set Ω. Any subset $A \subset \Omega$ is called an* **event***, including Ω and the empty set $\emptyset$.*

Example.

1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ be the set of possible outcomes of rolling a dice. Then $A = \{2\}$, $B = \emptyset$, $C = \{2, 4, 6\} = \{\text{the result is even}\}$ are events.

2. Consider the outcomes of throwing a fair coin infinitely many times. The corresponding sample space is

   $$\Omega = \{H, T\}^{\mathbb{N}} = \{(\omega_1, \omega_2, \dots) \text{ such that each } \omega_i \in \{H, T\}\},$$

   where H and T stand for the outcomes head and tail, respectively. Two possible events are

   $$A = \{(H, \omega_2, \omega_3, \dots) \text{ such that each } \omega_i \in \{H, T\}\} = \{\text{The first outcome is head}\},$$
   $$B = \{(H, H, H, \dots)\} = \{\text{The coin only turns heads}\}.$$

A probability measure is a function that assigns a probability to every possible event.

Definition 1.2.2. *Let Ω be a sample space. A* **probability measure** *$\mathbb{P}$ on Ω is an application*

$$\mathbb{P} : \{events\} \to [0, 1]$$

such that

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$.
- **(Countable additivity)** *For every sequence of disjoint events* $A_1, A_2, \ldots$

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

A pair $(\Omega, \mathbb{P})$ *is called a* **probability space**.

Example.

- The **uniform measure** on a finite set Ω is defined by

$$\mu(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} \qquad \text{for all } A \subset \Omega.$$

  This is a probability measure.

- The **Dirac measure** (or **Dirac mass**) at a given point $a \in \mathbb{R}$, denoted by δ_a, puts all the **mass** on a. It is defined by

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{otherwise} \end{cases}$$

  Again, this is a probability measure.

- The **Lebesgue measure** on $\mathbb{R}$ is the measure λ that assigns the length to each interval $[a, b] \subset \mathbb{R}$, that is,

$$\lambda\big([a, b]\big) = b - a.$$

  The Lebesgue measure is **not** a probability measure as its values are not restricted to $[0, 1]$. Since the length of $\mathbb{R}$ is infinite, the Lebesgue measure cannot be normalized to construct a probability measure. Nevertheless, the Lebesgue measure is countable additive and it plays a crucial role for continuous distributions.

Properties of probability measures

Plainly from the definition, probability measures have the following properties.

Proposition 1.2.3. *Let* $(\Omega, \mathbb{P})$ *be a probability space.*

(i) **(Monotonicity)** *If* $A \subset B$, *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(ii) *For any event* A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(iii) *For any events* A, B,

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

in particular $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

(iv) **(Union bound)** *More generally, let $(A_n)_{n\geq 1}$ be any sequence of sets (not necessarily disjoint). Then*

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}(A_n).$$

(v) **(Law of total probability)** *Let A be an event and $B_1, B_2, \ldots$ be a sequence of disjoint sets such that $\cup_{n\geq 1} B_n = \Omega$. Then*

$$\mathbb{P}(A) = \sum_{n\geq 1} \mathbb{P}(A \cap B_n).$$

1.3 Independent events and conditioning

Independence

Two events A and B are said to be **independent** if

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

Independence of more than two events is more subtle: for $A_1, A_2, \ldots, A_n$ to be independent, we have to check independence of every sub-family of sets A_i.

Definition 1.3.1 (Independence of events)**.** *Any events $A_1, \ldots, A_n$ are* **(mutually) independent** *if for every $k \leq n$ and every $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ we have*

$$\mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \times \cdots \times \mathbb{P}(A_{i_k}).$$

Example. (A fair coin eventually turns tail).
We turn back to the example of a fair coin flipped infinitely many times. Clearly, all flips are independent. Let us prove rigorously that the coin turns tail at least once.
Denote A_n the event that the n-th result is head. For every n, we have

$$\{\text{the coin never turns tail}\} \quad \subset \quad A_1 \cap A_2 \cap \cdots \cap A_n.$$

Then, by Proposition 1.2.3 (i),

$$\begin{aligned}
\mathbb{P}\left(\text{the coin never turns tail}\right) &\leq \mathbb{P}\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) \\
&= \mathbb{P}(A_1)\mathbb{P}(A_2) \times \cdots \times \mathbb{P}(A_n) \qquad \text{(by independence)}. \\
&= \left(\frac{1}{2}\right)^n.
\end{aligned}$$

Now this is true for every n, and $1/2^n$ tends to 0 as n goes to infinity, implying that

$$\mathbb{P}\left(\text{the coin never turns tail}\right) = 0.$$

Hence,

$$\mathbb{P}\left(\text{the coin turns tail at least once}\right) = 1 - \mathbb{P}\left(\text{the coin never turns tail}\right) = 1.$$

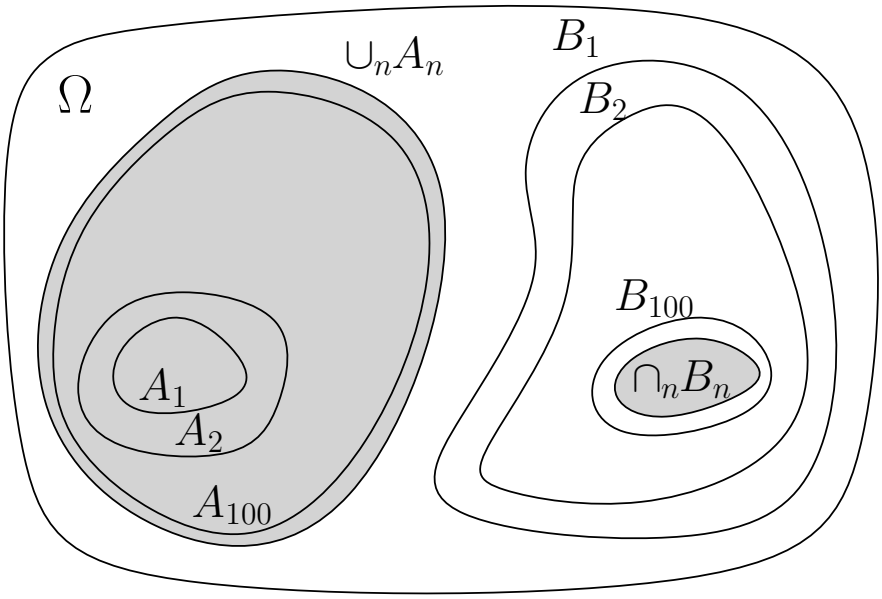


Figure 1.1: Illustration of monotone sequences of events: $(A_n)_{n\geq 1}$ is an increasing, $(B_n)_{n\geq 1}$ a decreasing sequence.

Conditioning

Let A and B be two events such that $\mathbb{P}(B) > 0$. We define the **conditional probability of A given B** by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.1}$$

This is the probability of A, given that B occurs.

Remark. The expression $A|B$ alone (without $\mathbb{P}(\dots)$) does **not** make sense. There is no such event!

By iterating formula 1.1 we obtain the following result.

Proposition 1.3.2 (Multiplicative formula for events)**.** *For any events $A_1, \dots, A_n$,*

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \times \cdots \times \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

This result is an important tool in Bayesian statistics.
Using conditional probabilities, the law of total probability (Proposition 1.2.3 (v)) can be written as follows. For any event A and a sequence of disjoint sets $B_1, B_2, \dots$ such that $\cup_{n\geq 1} B_n = \Omega$, we have

$$\mathbb{P}(A) = \sum_{n\geq 1} \mathbb{P}(A|B_n)\mathbb{P}(B_n).$$

1.4 Limits of events

Let $(A_n)_{n\geq 1}$ be a sequence of events. Does it make sense to consider the limit of A_n? And if so, do we have something like $\mathbb{P}(\lim_n A_n) = \lim_n \mathbb{P}(A_n)$? In the case where $(A_n)_{n\geq 1}$ is a **monotone** sequence (see Figure 1.1 for illustration), the following result holds.

Theorem 1.4.1. *Let $(\Omega, \mathbb{P})$ be a probability space.*

1. *Let $(A_n)_{n\geq 1}$ be an increasing sequence of events, i.e. $A_1 \subset A_2 \subset A_3 \subset \dots$ Then*

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

2. *Let $(B_n)_{n\geq 1}$ be a decreasing sequence of events: $B_1 \supset B_2 \supset B_3 \supset \dots$, then*

$$\mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

Chapter 2

Random variables and probability distributions

From now on, we work on a fixed probability space $(\Omega, \mathbb{P})$, where $\mathbb{P}$ is a probability measure. We say that an event A is $\mathbb{P}$-**almost sure** (or just **almost sure** (a.s.) if no ambiguity), if $\mathbb{P}(A) = 1$. Elements of Ω are usually denoted by ω.

2.1 Random variables and their distributions

Definition 2.1.1 (Random variable)**.** *A* **random variable** X *is any real function defined on* Ω*, that is*

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega). \end{array}$$

A simple example of a random variable is the **indicator function**. For a given event A, the indicator function of A is denoted by $\mathbb{1}_A$ and defined by

$$\begin{array}{rccc} \mathbb{1}_A: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases} \end{array}$$

Note that indicator functions are similar to Dirac measures, as $\mathbb{1}_A(\omega) = \delta_\omega(A)$.

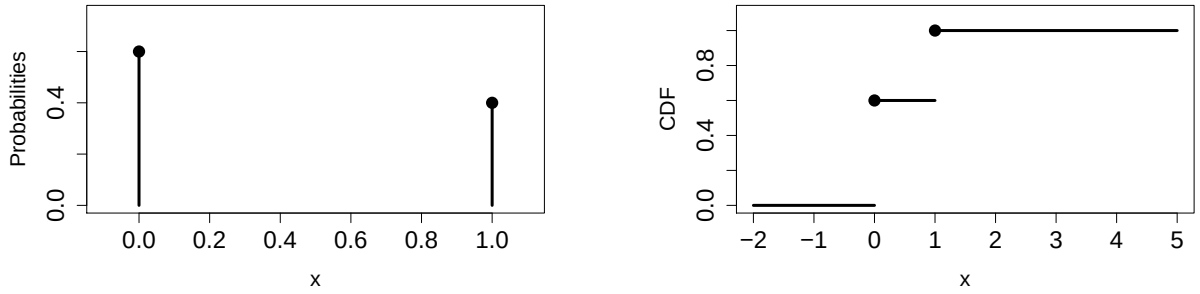
Definition 2.1.2. *The* **probability distribution** *(or* **law***) of a random variable* X*, denoted by* $\mathbb{P}_X$*, is the probability measure on* $\mathbb{R}$ *such that for any event* A

$$\mathbb{P}_X(A) = \mathbb{P}\left(\{\omega \text{ such that } X(\omega) \in A\}\right) = \mathbb{P}(X \in A).$$

We write $X \sim \mathbb{P}_X$ *which reads "*X *has distribution* $\mathbb{P}_X$*".*

Definition 2.1.3. *The* **cumulative distribution function** *(or just distribution function) of* X *is the function* F_X *defined by*

$$\begin{array}{rccc} F_X: & \mathbb{R} & \to & [0,1] \\ & t & \mapsto & \mathbb{P}(X \le t). \end{array}$$



(a) Probability mass function

(b) Cumulative distribution function

Figure 2.1: Bernoulli distribution with parameter $p = 0.4$.

The following result states that any probability distribution is completely described by its cumulative distribution function.

Theorem 2.1.4. *Two random variables X and Y have the same probability distribution, i.e. $\mathbb{P}_X(A) = \mathbb{P}_Y(A)$ for every event A, if and only if $F_X(t) = F_Y(t)$ for every $t \in \mathbb{R}$.*

Any distribution function F_X has the following properties.

(i) F_X is non-decreasing, since for $s \leq t$, $\{X \leq s\} \subset \{X \leq t\}$. Hence, $F_X(s) \leq F_X(t)$.

(ii) F_X has the following limit behavior:

$$\lim_{t \to -\infty} F_X(t) = \mathbb{P}(\emptyset) = 0, \quad \lim_{t \to +\infty} F_X(t) = \mathbb{P}(\Omega) = 1.$$

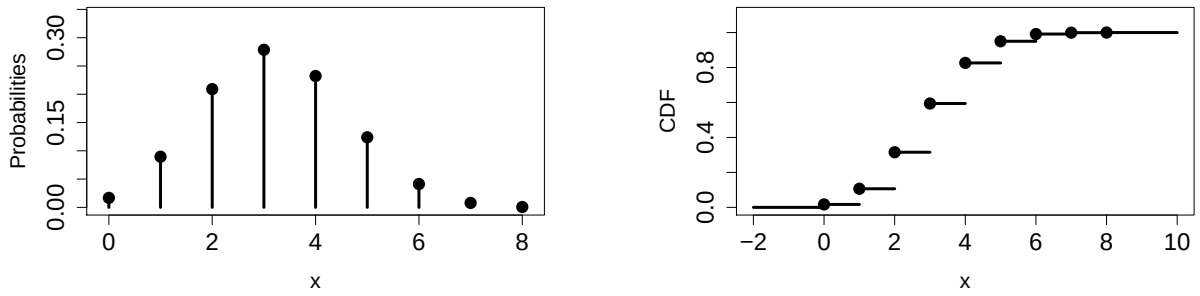
(iii) F_X is right-continuous. Indeed, we see that

$$\begin{aligned} \lim_{n \to +\infty} F_X(t + 1/n) &= \lim_{n \to +\infty} \mathbb{P}(X \leq t + 1/n) \\ &= \mathbb{P}(\cap_{n \geq 1} \{X \leq t + 1/n\}) \qquad \text{(by Theorem 1.4.1)} \\ &= \mathbb{P}(X \leq t) = F_X(t). \end{aligned}$$

Theorem 2.1.5. *Any function F with properties (i), (ii) and (iii) above, is the cumulative distribution function of some random variable.*

Discrete distributions

We say that X has a **discrete distribution** if X takes its values in a finite or countable set $\{x_1, x_2, \dots\}$. Recall that a space Ω is **countable** if Ω is in one-to-one correspondence with $\mathbb{N}$ or a subset of $\mathbb{N}$. For instance, $\mathbb{N}$ and $\mathbb{Z}$ are countable, but $\mathbb{R}$ is not. Discrete distributions are entirely described by their **probability mass function** $p(x) = \mathbb{P}(X = x)$ for $x \in \{x_1, x_2, \dots\}$.



(a) Probability mass function

(b) Cumulative distribution function

Figure 2.2: Binomial distribution with parameters $n = 8$ and $p = 0.4$.

Important discrete distributions

- The **Bernoulli distribution** $B(p)$ with parameter $p \in [0, 1]$ models the success and failure of an experiment. Its probability mass function is defined as

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

  See Figure 2.1 for its probability mass function and its cumulative distribution function.

- The **binomial distribution** $B(n, p)$ with parameters $n \geq 1$ and $p \in [0, 1]$ models the number of successes in n independent Bernoulli trials. Its probability mass function is given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

  See Figure 2.2.

- The **geometric distribution** with success probability $p \in [0, 1]$ models the first success in a series of Bernoulli trials. Its probability mass function is given by

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \ldots$$

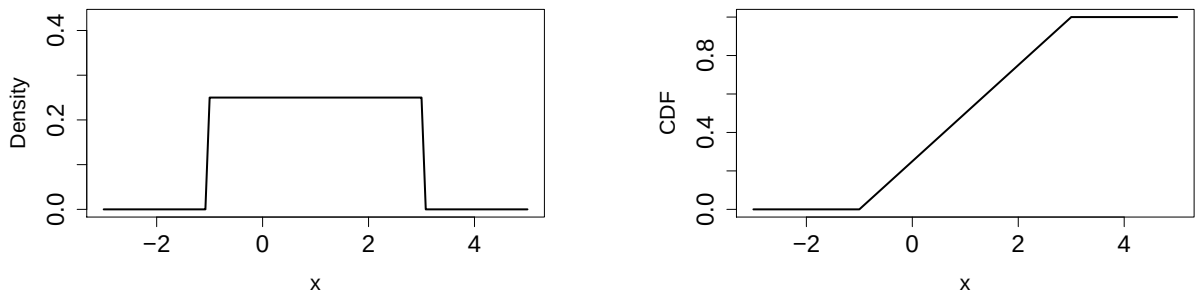
- The **Poisson distribution** with parameter $\lambda > 0$ is defined by

$$\mathbb{P}(X = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

- The **discrete uniform distribution** on a finite set of values $\{x_1, \ldots, x_m\}$ is defined by

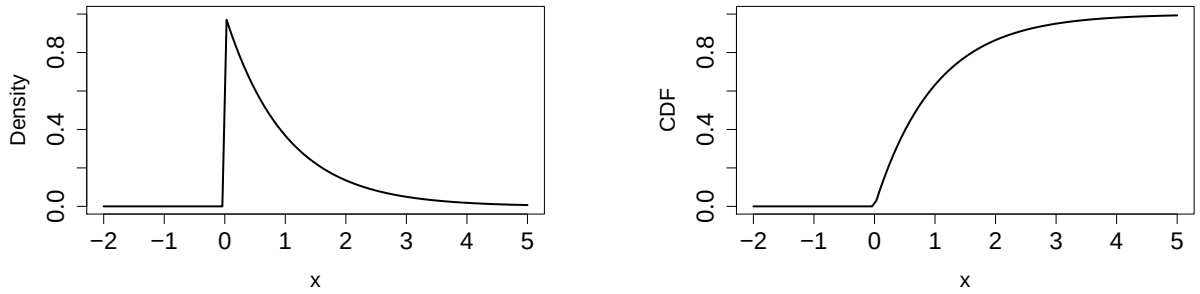
$$\mathbb{P}(X = x_k) = \frac{1}{m} \quad \text{for } k = 1, \ldots, m.$$

The cumulative distribution function of any discrete distribution is a step function. Furthermore, the jumps indicate the values taken by the random variable and the height of the jump indicates the associated probability, see Figures 2.1 and 2.2.



(a) Density function (b) Cumulative distribution function

Figure 2.3: Uniform distribution on $[-1, 3]$.



(a) Density function (b) Cumulative distribution function

Figure 2.4: Exponential distribution with parameter $\lambda = 1$.

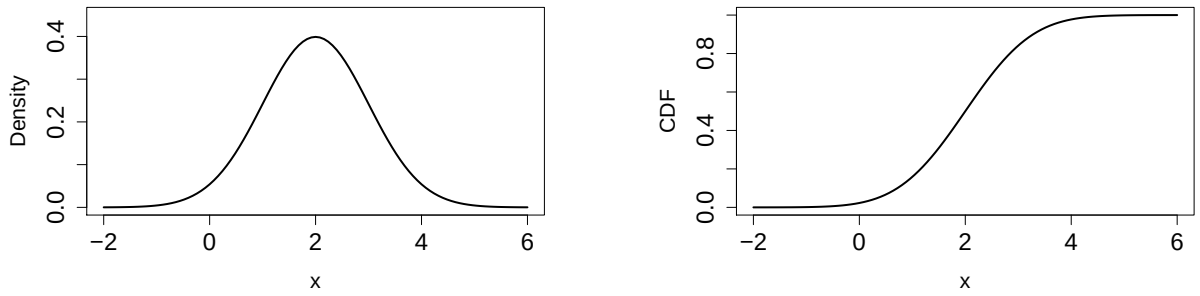
Continuous distributions

Definition 2.1.6 (Continuous random variable)**.** *We say that X has* **continuous distribution** *if X takes its values in $\mathbb{R}$ (or in an interval of $\mathbb{R}$) and if there is a non-negative function f such that for any event A*

$$\mathbb{P}(X \in A) = \int_A f(x)\mathrm{d}x. \tag{2.1}$$

The function f is called the **density** *of X.*

Any density function f is non-negative and integrates to one, *i.e.* $\int_{\mathbb{R}} f(x)\mathrm{d}x = \mathbb{P}(X \in \mathbb{R}) = 1$. The density entirely describes the distribution of the random variable. However, density functions are not unique. Any density function whose values are modified in a countable number of points defines exactly the same distribution. This is due to fact that the probability



(a) Density function

(b) Cumulative distribution function

Figure 2.5: Normal distribution with parameters $\mu = 2$ and $\sigma^2 = 1$.

of observing any given point x is zero for a continuous distribution. That is,

$$\mathbb{P}(X = x) = \int_{\{x\}} f(x)\mathrm{d}x = 0, \qquad \forall x \in \mathbb{R},$$

as the Lebesgue integral over a single point, $\{x\}$, is always zero.

Important continuous distributions

- The **uniform distribution** $U[a, b]$ on the interval $[a, b]$ has density

  $$f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x).$$

  See Figure 2.3 for its density and its cumulative distribution function.

- The **exponential distribution** $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$ has density

  $$f(x) = \lambda \exp(-\lambda x)\mathbb{1}_{x\geq 0}.$$

  See Figure 2.4.

- The **normal distribution** (or gaussian distribution) $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$ has density

  $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  See Figure 2.5.

The cumulative distribution function of any continuous distribution is continuous.

There exist distributions that are neither discrete nor continuous. For instance, consider the random variable $X = \min\{1, Y\}$ with $Y \sim \mathcal{E}(1)$. Clearly, X takes its value in the entire interval $[0, 1]$, but its distribution is not continuous since the probability $\mathbb{P}(X = 0) = \mathbb{P}(Y > 1) > 0$ is strictly positive. The distribution of X is called a *censored distribution*.

Cumulative distribution functions and densities

If X is continuous with density f, then by (2.1) we have, for all $t \in \mathbb{R}$,

$$F_X(t) = \mathbb{P}(X \in (-\infty, t]\ ) = \int_{-\infty}^{t} f(x)\mathrm{d}x. \tag{2.2}$$

By differentiation we have $F'_X(t) = f(t)$ for almost all t. Note that, for instance, the uniform distribution on the interval $[a, b]$ has a density, but its distribution function is not differentiable at a and b, see Figure 2.3.

Example. Let X follow the uniform distribution in $[0, 1]$. What is the density (if any) of X^2 ? To answer this question we compute the cumulative distribution function F_{X^2} of X^2. The map $x \mapsto \sqrt{x}$ is increasing and maps $[0, 1]$ onto itself. So for each $t \in (0, 1)$,

$$\left\{X^2 \leq t\right\} = \left\{X \leq \sqrt{t}\right\}.$$

Therefore, for $t \in (0, 1)$,

$$F_{X^2}(t) = \mathbb{P}(X^2 \leq t) = \mathbb{P}(X \leq \sqrt{t}) = \int_0^{\sqrt{t}} \mathrm{d}x = \sqrt{t}.$$

By differentiating the distribution function F_{X^2}, we obtain the density f_{X^2} of X^2. That is,

$$f_{X^2}(t) = F'_{X^2}(t) = \frac{\partial}{\partial t}\sqrt{t} = \frac{1}{2\sqrt{t}}\mathbb{1}_{[0,1]}(t).$$

Simulation of continuous random variables

All programming languages provide a generator of pseudo-random variables to sample from the uniform distribution $U[0, 1]$. In Matlab and Scilab, for instance, this is obtained by the function `rand()`, in Python by `numpy.random.uniform(size=1)` and in R by `runif()`. Those generators of the uniform distribution are the starting point for the simulation of realisations of any other distribution. In other words, realisations of a given distribution F are obtained by transformations of realisations of the uniform distribution. The most common method to do so is the **inverse transform sampling**, which is described in the following algorithm:

Algorithm
U=`rand()`
`return` $F^{-1}(U)$.

More precisely, suppose that the cumulative distribution function F, from which we want to generate realisations, is an invertible function with inverse F^{-1}. Then the random variable $Y = F^{-1}(U)$ where $U \sim U[0, 1]$ has distribution F. Indeed, since F is increasing, we have

$$\begin{aligned} F_Y(t) &= \mathbb{P}(Y \leq t) = \mathbb{P}(F^{-1}(U) \leq t) \\ &= \mathbb{P}(U \leq F(t)) \\ &= \int_0^{F(t)} \mathrm{d}x \qquad (\text{since } U \sim U[0, 1]) \\ &= F(t). \end{aligned}$$

This justifies the algorithm above.

If F is not invertible, the same result holds with the **pseudo-inverse function** F^{-1} of F defined by

$$F^{-1}(u) = \inf\{x \text{ such that } F(x) \geq u\} \quad \text{for } u \in [0,1],$$

and the same algorithm can be applied to generate pseudo-random variables from distribution F.

Example. (Simulation from the exponential distribution) Consider the exponential distribution $\mathcal{E}(1)$. The distribution function is $F(t) = \int_0^t \mathrm{e}^{-u} du = 1 - \mathrm{e}^{-t}$ for $t > 0$. To find F^{-1}, we solve, for $t > 0$,

$$F(t) = x \iff 1 - \mathrm{e}^{-t} = x \iff 1 - x = \mathrm{e}^{-t} \iff t = -\log(1-x).$$

That is, $F^{-1}(x) = -\log(1-x)$. Thus the command `-log(1-rand())` returns a realisation of the exponential distribution $\mathcal{E}(1)$.

2.2 Expectation and moments

Expectation

Let X be a random variable defined on a probability space $(\Omega, \mathbb{P})$. The expectation is the average value taken by X, that is, its a weighted mean of the outcomes of X, where the weights are probabilities. The mean is a centrality measure of the distribution of X and indicates the order of magnitude of the values taken by X.

Definition 2.2.1 (Expectation of a random variable)**.**

- *Let X have a discrete distribution taking its values in $\{x_1, x_2, \dots\}$. If $\sum_{k\geq 1} |x_k| \mathbb{P}(X = x_k) < \infty$, the* **expectation** *or* **mean** *$\mathbb{E}[X]$ of X exists and is given by*

$$\mathbb{E}[X] = \sum_{k\geq 1} x_k \mathbb{P}(X = x_k).$$

- *Let X have a continuous distribution with density f. If $\int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x < \infty$, the expectation $\mathbb{E}[X]$ of X exists and is given by*

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

The expectation is not defined for all distributions. For instance, the *Cauchy distribution* is not integrable. Its density is given by

$$f(x) = \frac{1}{\pi(1+x^2)},$$

and in fact, one can show that $\int_{\mathbb{R}} |x| f(x) \mathrm{d}x = +\infty$.

Theorem 2.2.2 (Expectation of a random variable)**.** *The expectation* $\mathbb{E}[X]$ *of* X *(if it exists) is a real number with the following properties.*

- *If* $X = \mathbb{1}_A$*, then*
$$\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A).$$
- **(Linearity)** *For any real numbers* a, b *and any integrable random variables* X, Y*,*
$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$
- **(Monotonicity)** *If* $X \leq Y$ *almost surely, i.e.* $\mathbb{P}(X \leq Y) = 1$ *or* $X(\omega) \leq Y(\omega)$ *for almost every* ω*, and if* $\mathbb{E}[|Y|] < \infty$*, then* $\mathbb{E}[|X|] < \infty$ *and*
$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

Theorem 2.2.3. *Let* X *be a random variable and* $\phi : \mathbb{R} \to \mathbb{R}$ *be a function. Set* $Y = g(X)$*.*

- *If* X *is discrete and* $\mathbb{E}[Y]$ *exists, then*
$$\mathbb{E}[Y] = \mathbb{E}[\phi(X)] = \sum_{k \geq 1} \phi(x_k)\mathbb{P}(X = x_k).$$
- *If* X *is continuous with density* f *and* $\mathbb{E}[Y]$ *exists, then*
$$\mathbb{E}[Y] = \mathbb{E}[\phi(X)] = \int_{\mathbb{R}} \phi(x) f(x) \mathrm{d}x.$$

Proposition 2.2.4. *Let* X *be an almost surely non-negative random variable, i.e.* $\mathbb{P}(X \geq 0) = 1$*. If* $\mathbb{E}[X] = 0$*, then* $X = 0$ *almost surely.*

Example. Let X be a random variable following the uniform distribution on $[0, 1]$. Then, for every n,
$$\mathbb{E}[X^n] = \int x^n \mathbb{1}_{[0,1]}(x) \mathrm{d}x = \int_{[0,1]} x^n \mathrm{d}x = \frac{x^{n+1}}{n+1}\bigg|_{x=0}^{x=1} = \frac{1}{n+1}.$$

Exercise. Let X be a random variable with exponential distribution $\mathcal{E}(1)$. Show by induction that $\mathbb{E}[X^n] = n!$ for all n.

Variance and other higher-order moments

The variance of a random variable is a measure of the dispersion or variability of a random variable around its mean.

Definition 2.2.5 (Variance of a random variable)**.** *Let* X *be a random variable such that* $\mathbb{E}[X^2] < +\infty$*. The* **variance** *of* X *is defined by*
$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$
The **standard deviation** *is defined as* $\sigma = \sqrt{\mathrm{Var}(X)}$*.*

The variance has the following properties.

Proposition 2.2.6. *Let X, Y be random variables such that $\mathbb{E}[X^2] < +\infty$ and $\mathbb{E}[Y^2] < +\infty$. Then*

(i) $0 \leq \mathrm{Var}(X) < \infty$

(ii) $\mathrm{Var}(X) = 0 \iff \mathbb{P}(X = c) = 1$ *for some constant c.*

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

(iv) For any constants a, b, $\mathrm{Var}(aX + b) = \mathrm{Var}(aX) = a^2\mathrm{Var}(X)$.

(v) If X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. To show (i), note that $(X - \mathbb{E}[X])^2 \geq 0$ almost surely. By monotonicity of the expectation, $\mathbb{E}[(X - \mathbb{E}[X])^2] \geq 0$. To show $\mathrm{Var}(X) < \infty$, notice that $|X| \leq X^2 + 1$ almost surely. Again by monotonicity,

$$\mathbb{E}\left[|X|\right] \leq \mathbb{E}\left[1 + X^2\right] = 1 + \mathbb{E}\left[X^2\right] < +\infty,$$

since $\mathbb{E}[X^2] < +\infty$. Thus, $\mathrm{Var}(X) < \infty$.
To show (iii), by expanding the term inside the expectation, we obtain

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\right] \\
&= \mathbb{E}[X^2] + \mathbb{E}\left[\mathbb{E}[X]^2\right] - 2\mathbb{E}\left[X\mathbb{E}[X]\right] \qquad \text{(by linearity)} \\
&= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

For (iv) we compute

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

Finally, to show (vi), first consider the case when $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. We find

$$\begin{aligned}
\mathrm{Var}(X + Y) &= \mathbb{E}[(X + Y - \underbrace{\mathbb{E}[X + Y]}_{=0})^2] = \mathbb{E}[X^2 + Y^2 + 2XY] \\
&= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[X][Y] \qquad \text{(by independence)} \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}$$

For the general case, use (v), that is, adding a constant does not have any impact on the variance.

□

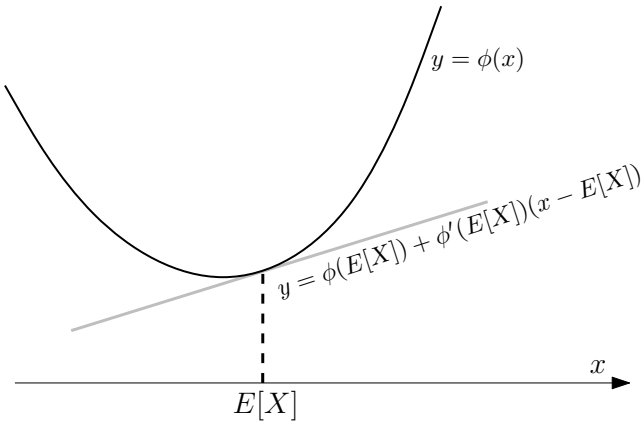


Figure 2.6: Illustration of convexity for the proof of Jensen's inequality.

The quantity $\mathbb{E}[X^2]$ is called the second moment of X, and $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ the second central moment. These definitions can be generalized.

Definition 2.2.7 (Moments of a random variable)**.** *Let X be a random variable such that $\mathbb{E}[|X|^k] < +\infty$ for some $k \in \mathbb{N}$. Then*

- *the k-**th moment** of X exists and is defined by $\mathbb{E}[X^k]$.*
- *the k-**th central moment** of X exists and is defined as*

$$\mathbb{E}\left[(X - \mathbb{E}[X])^k\right].$$

Moments play an important role in statistics. We have already seen that the second central moment is the variance and a measure of the variability of a random variable. The (standardized) third central moment is called skewness and is an indicator of asymmetry of the distribution. The (standardized) fourth central moment is called kurtosis and describes the tails of a distribution. More precisely, the kurtosis indicates whether a distribution is heavy tailed or light tailed.

2.3 Inequalities

This section provides some useful inequalities.

Theorem 2.3.1 (Jensen's inequality)**.** *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function and X be an integrable random variable. Then*

$$\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\phi(X)\right].$$

For instance we have $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ and $\exp\{\mathbb{E}[X]\} \leq \mathbb{E}[\mathrm{e}^X]$. Moreover, the theorem implies that for any concave function ϕ we have $\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$, provided that $\mathbb{E}[|X|] < \infty$.

Proof. For simplicity we assume that ϕ is differentiable. By convexity of ϕ, the tangent line at some point a is below the curve of ϕ (see Figure 2.6). More precisely, for all real x we have

$$\phi(x) \geq \phi(a) + \phi'(a)(x - a).$$

In particular, this holds for $a = \mathbb{E}[X]$ and for any real number $x = X(\omega)$ with $\omega \in \Omega$, so that

$$\phi(X) \geq \phi(\mathbb{E}[X]) + \phi'(\mathbb{E}[X])(X - \mathbb{E}[X]) \ a.s.$$

By taking the expectation,

$$\begin{aligned}\mathbb{E}[\phi(X)] &\geq \mathbb{E}\Big[\phi(\mathbb{E}[X]) + \phi'(\mathbb{E}[X])(X - \mathbb{E}[X])\Big] \\ &\geq \mathbb{E}[\phi(\mathbb{E}[X])] + \phi'(\mathbb{E}[X])\mathbb{E}\left[X - \mathbb{E}[X]\right] \\ &= \phi(\mathbb{E}[X]) + 0,\end{aligned}$$

since $\phi(\mathbb{E}[X])$ is a constant, and $\mathbb{E}\left[X - \mathbb{E}[X]\right] = \mathbb{E}[X] - \mathbb{E}[X] = 0$.

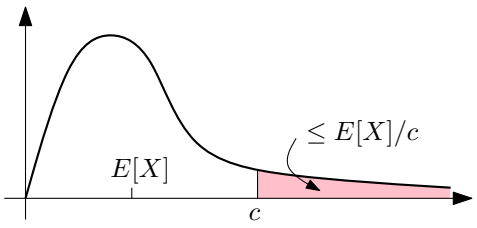
□

We now state two important inequalities that bound the probability that X deviates from its mean.

Theorem 2.3.2. *Let X be an integrable random variable.*

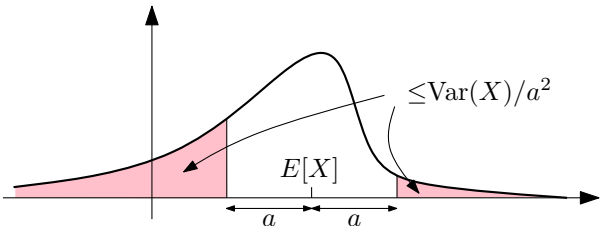
- **Markov's inequality.** *If X is non-negative, i.e. $X \geq 0$ almost surely, then for any constant $c > 0$,*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}.$$



- **Chebyshev's inequality.** *If $\mathbb{E}[X^2] < +\infty$, then for any constant $a > 0$,*

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) \leq \frac{\mathrm{Var}(X)}{a^2}.$$



Proof. To show Markov's inequality, note that

$$\begin{aligned}1 \geq \mathbb{1}_{X \geq c} \ a.s. &\implies X \geq X\mathbb{1}_{X \geq c} \ a.s. \qquad \text{(since } X \geq 0 \ a.s.\text{)} \\ &\implies \mathbb{E}[X] \geq \mathbb{E}\left[X\mathbb{1}_{X \geq c}\right] \geq c\mathbb{E}\left[\mathbb{1}_{X \geq c}\right].\end{aligned}$$

To conclude use that $\mathbb{E}[\mathbb{1}_{X\geq c}] = \mathbb{P}(X \geq c)$.
For Chebyshev's inequality, first note that

$$\{|X - \mathbb{E}[X]| \geq a\} = \left\{(X - \mathbb{E}[X])^2 \geq a^2\right\},$$

so that

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) = \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq a^2\right).$$

Now, $(X - \mathbb{E}[X])^2$ is a non-negative random variable. Applying Markov's inequality yields

$$\mathbb{P}\left((X - \mathbb{E}[X])^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

This completes the proof. □

2.4 How to find the distribution of X?

1st method: Find the distribution function by an explicit computation

We have already seen this method in the example on p.16.

2nd method: Find the density with a change of variable

The strategy relies on the following result.

Theorem 2.4.1 (Characterization with bounded and continuous ϕ)**.** *Let X and Y be two random variables.*

- *If $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(Y)]$ for* **every** *bounded and continuous function ϕ, then X and Y have the same distribution.*
- *In particular, if one can find a function f such that, for* **every** *bounded and continuous function ϕ,*

  $$\mathbb{E}[\phi(X)] = \int \phi(x) f(x) \mathrm{d}x,$$

  then f is the density of X.

Example. Consider a random variable X with uniform distribution on $[0, 1]$. We illustrate the second method by showing (again) that X^2 has density $\frac{1}{2}\sqrt{x}\mathbb{1}_{[0,1]}(x)$. Let ϕ be any bounded and continuous function. Then we compute

$$\mathbb{E}[\phi(X^2)] = \int \phi(x^2) \underbrace{\mathbb{1}_{[0,1]}(x)}_{\text{density of } X} \mathrm{d}x = \int_0^1 \phi(x^2)\mathrm{d}x.$$

We make the change of variables $t = x^2$, $x = \sqrt{t}$, $\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1}{2\sqrt{t}}$. This gives

$$\mathbb{E}[\phi(X^2)] = \int_{x=0}^{x=1} \phi(x^2)\mathrm{d}x = \int_{t=0}^{t=1} \phi(t) \underbrace{\frac{1}{2\sqrt{t}}}_{\text{density of } X^2} \mathrm{d}t.$$

As this holds for any bounded and continuous ϕ, according to the theorem, X^2 has density $\frac{1}{2\sqrt{t}}\mathbb{1}_{[0,1]}(t)$.

3rd method: Find the characteristic function

We now introduce an important tool: the characteristic function, also called Fourier transform.

Definition 2.4.2. *The* **characteristic function** Φ_X *of a random variable* X *is defined as*

$$\begin{array}{rccc} \Phi_X(t): & \mathbb{R} & \to & \mathbb{C} \\ & t & \mapsto & \mathbb{E}[\mathrm{e}^{itX}]. \end{array}$$

Here i is the complex number satisfying $i^2 = -1$. All you need to know about the exponential of a complex number is that the power rule $\mathrm{e}^{z+z'} = \mathrm{e}^z\mathrm{e}^{z'}$ also holds for complex numbers and that for any real number r we have $|\mathrm{e}^{ir}| = 1$.
The characteristic function is well defined for any probability distribution. As it names indicates, it completely describes the associated probability distribution.

Theorem 2.4.3. *X and Y have the same distribution if and only if $\Phi_X(t) = \Phi_Y(t)$ for all t.*

Example. Let X have the exponential distribution with parameter 1. For the characteristic function we obtain

$$\begin{aligned} \Phi_X(t) = \mathbb{E}[\mathrm{e}^{itX}] &= \int_0^{+\infty} \mathrm{e}^{itx}\mathrm{e}^{-x}\mathrm{d}x \\ &= \int_0^{+\infty} \mathrm{e}^{x(it-1)}\mathrm{d}x = \left.\frac{\mathrm{e}^{x(it-1)}}{it-1}\right|_{x=0}^{x=+\infty} \\ &= \frac{1}{it-1}\left(\lim_{x\to+\infty} \mathrm{e}^{x(it-1)} - 1\right) \\ &= \frac{1}{it-1}\left(\lim_{x\to+\infty} \underbrace{\mathrm{e}^{xit}}_{\text{of modulus 1}} \times \underbrace{\mathrm{e}^{-x}}_{\to 0} - 1\right) = \frac{1}{1-it}. \end{aligned}$$

Plainly from the definition, the characteristic function has the following properties. First, note that $\Phi_X(0) = \mathbb{E}[\mathrm{e}^0] = \mathbb{E}[1] = 1$, and for every t, $|\Phi_X(t)| = |\mathbb{E}[\mathrm{e}^{itX}]| \leq \mathbb{E}[|\mathrm{e}^{itX}|] = 1$. That is, the characteristic function is well-defined on $\mathbb{R}$ for any random variable X. Furthermore, if $\mathbb{E}[|X|] < +\infty$, then by swapping expectation and derivative according to Theorem 2.6.4, we find that

$$\Phi'_X(t) = \frac{\partial}{\partial t}\mathbb{E}\left[\mathrm{e}^{itX}\right] = \mathbb{E}\left[\frac{\partial}{\partial t}\mathrm{e}^{itX}\right] = \mathbb{E}[iX\mathrm{e}^{itX}].$$

In particular, $\Phi'_X(0) = i\mathbb{E}[X]$.

For a discrete random variable X with values in $\mathbb{N}$, we usually use the **generating function** G_X defined by

$$G_X(z) = \mathbb{E}[z^X] = \sum_{k\geq 0} \mathbb{P}(X = k)z^k.$$

Like for the characteristic function, two random variables X and Y have the same distribution if $G_X(z) = G_Y(z)$ for all z.

2.5 L^p-spaces

Definition 2.5.1 (L^p-space)**.** *Let $p \geq 1$ be a real number. The space $L^p(\Omega, \mathbb{P})$ (or just L^p-**space**, if there is no ambiguity) is defined as the set of random variables X such that $\mathbb{E}[|X|^p] < +\infty$. Furthermore, we define a norm on the L^p-space called the the L^p-**norm** by*

$$\|X\|_p = \mathbb{E}\left[|X|^p\right]^{1/p}.$$

In particular, L^1 is the set of random variables such that $\mathbb{E}[|X|] < +\infty$. If $X \in L^1$, we say that X is **integrable**.

Note that in the definition, p is any real number in $[1, +\infty)$, but in practice we usually consider only integer values of p, i.e. $L^1, L^2, \ldots$

Example.

- Let X be a bounded random variable, that is, there exists a constant c such that $|X| \leq c$ almost surely. Then $\mathbb{E}[|X|^p] \leq c^p < +\infty$. Thus, all bounded random variables belong to the L^p-space for any $p \geq 1$.

- Here is an example which proves that L^1 and L^2 are not identical sets, i.e. $L^1 \neq L^2$. Let X have density $\frac{2}{(x+1)^3}$ on $[0, +\infty)$. First, we check that this is a density. Indeed,

$$\int_0^\infty \frac{2}{(x+1)^3}\mathrm{d}x = -\frac{1}{(x+1)^2}\Big|_{x=0}^{x=+\infty} = 1.$$

  Then, it is easy to see that $X \in L^1$, since

$$\mathbb{E}[|X|] = \int_0^\infty x\frac{2}{(x+1)^3}\mathrm{d}x = 1 < +\infty,$$

  while

$$\mathbb{E}[|X|^2] = \int_0^\infty x^2\frac{2}{(x+1)^3}\mathrm{d}x = +\infty,$$

  that is, $X \notin L^2$.

- We have seen that when X follows the exponential distribution $\mathcal{E}(1)$, then $\mathbb{E}[X^p] = p! < +\infty$ for all p. Hence, X belongs to all L^p-spaces.

The L^p-space is a **vector space**, which means that for any random variables $X, Y \in L^p$ and any constant $a \in \mathbb{R}$,

$$aX \in L^p, \quad \text{and} \quad X + Y \in L^p.$$

In the definition above we stated that $\|\cdot\|_p$ is a **norm** on L^p. Indeed, this can be verified by showing that the following properties hold: for any $X, Y \in L^p$ and $a \in \mathbb{R}$, we have

- $\|aX\|_p = |a| \, \|X\|_p$,
- $\|X\|_p = 0$ if and only if $X = 0$ almost surely,
- **Triangle inequality:** $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.

An important property of L^p-spaces is that they are nested, that is, they are all included one another.

Theorem 2.5.2. *For $q > p$, $L^q \subset L^p$. More precisely,*

$$\cdots \subset L^p \subset L^{p-1} \subset L^{p-2} \subset \cdots \subset L^2 \subset L^1.$$

The theorem says that if $\mathbb{E}[|X|^q] < \infty$, then $\mathbb{E}[|X|^p] < \infty$ for all $p < q$.

Proof. The theorem can be proven by Jensen's inequality. The trick is to write $|X|^q = (|X|^p)^{q/p}$. As $q/p > 1$, the map $x \mapsto x^{q/p}$ is convex on $\mathbb{R}_+$. Hence, by Jensen's inequality,

$$\mathbb{E}\left[|X|^q\right] = \mathbb{E}\left[(|X|^p)^{q/p}\right] \geq \left(\mathbb{E}\left[|X|^p\right]\right)^{q/p}.$$

Taking both sides to the power $1/p$ yields

$$\mathbb{E}\left[|X|^q\right]^{1/q} \geq \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}.$$

That is, $\|X\|_q \geq \|X\|_p$. It follows that, if $\|X\|_q$ is finite, so is $\|X\|_p$. □

Definition 2.5.3 (L^p-convergence)**.** *Let $(X_n)_{n\geq 0}$ be a sequence of random variables. One says that X_n* **converges to X in L^p** *if*

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

We write $X_n \xrightarrow{L^p} X$.

Obviously, this amounts to say that $\|X_n - X\|_p$ tends to zero. Now, if $X_n \xrightarrow{L^q} X$ for some q, then $X_n \xrightarrow{L^p} X$ for every $p < q$, since we have

$$\|X_n - X\|_p \leq \underbrace{\|X_n - X\|_q}_{\to 0}.$$

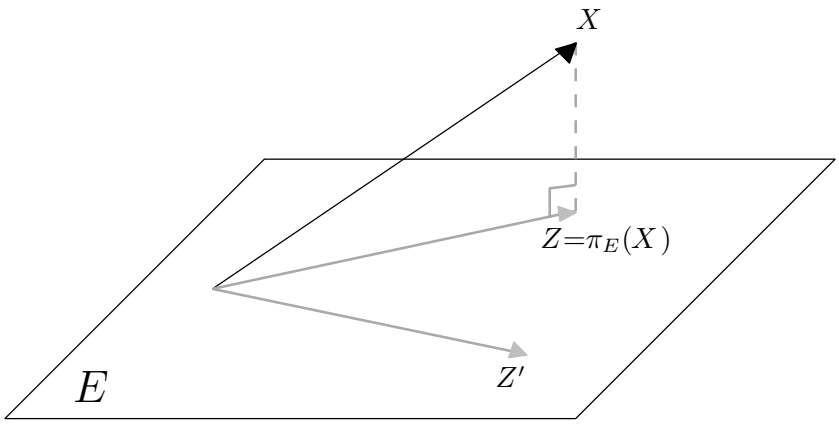


Figure 2.7: Illustration of the orthogonal projection $\pi_E(X)$ of X onto E, verifying that $(X - \pi_E(X))$ is orthogonal to any element $Z' \in E$.

The special case of the L^2-space

Let X and Y be in $L^2(\Omega, \mathbb{P})$. Then the product XY is integrable, due to the following inequality.

Theorem 2.5.4 (Cauchy-Schwarz's inequality)**.** *Let $X, Y \in L^2$. Then*

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2}\mathbb{E}[Y^2]^{1/2}, \tag{2.3}$$

with equality if and only if there are real constants a, b such that $a \neq 0$ or $b \neq 0$ and $aX + bY = 0$ almost surely.

Since the right-hand side in (2.3) is finite, $XY \in L^1$.

When we set $Y = 1$ *a.s.*, we get $\mathbb{E}[|X|] \leq \mathbb{E}[X^2]^{1/2}$. Therefore, $X \in L^2 \Rightarrow X \in L^1$.

A **scalar product** (or **inner product**) on the L^2-space is given by

$$\langle X, Y \rangle = \mathbb{E}[XY].$$

This means that the following properties hold: for all $X, X', Y \in L^2$ and constants $a, b \in \mathbb{R}$,

- **Symmetry:** $\langle X, Y \rangle = \langle Y, X \rangle$.
- **Linearity:** $\langle aX + bX', Y \rangle = a\langle X, Y \rangle + b\langle X', Y \rangle$.
- **Positive-definiteness:** $\langle X, X \rangle \geq 0$. Moreover, $\langle X, X \rangle = 0$ if and only if $X = 0$ a.s.

By analogy with the usual scalar product in geometry, we say that X and Y are **orthogonal** if $\langle X, Y \rangle = 0$ and we write $X \perp Y$. Moreover, we can consider orthogonal projections in L^2.

Let E be a linear subspace of L^2 and $X \in L^2$. By definition, the orthogonal projection of X onto E, denoted by $\pi_E(X)$, is the unique random variable $\pi_E(X) \in E$ such that $(X - \pi_E(X)) \perp Z'$ for every element $Z' \in E$, see Figure 2.7.

Furthermore, the orthogonal projection $\pi_E(X)$ of X onto E can be characterized by a minimization problem.

Theorem 2.5.5 (Orthogonal projection as a minimization problem)**.** *Let $X \in L^2$ and E be a linear subspace of L^2. Then*

$$Z = \pi_E(X) \quad \Longleftrightarrow \quad \begin{cases} Z \in E \\ \mathbb{E}[(X-Z)^2] = \min_{Z' \in E} \mathbb{E}[(X-Z')^2] \end{cases}$$

Example. Let $X \in L^2$. What is the orthogonal projection $\pi_E(X)$ of X onto the linear subspace E defined as the set of all constant random variables? To put it differently, if we have to characterize a probability distribution by a single real number, which is the most appropriate, the most informative value to summarize the distribution function of X? Mathematically, here E is defined as

$$E = \{X = a \text{ almost surely}, a \in \mathbb{R}\}.$$

1st method: using orthogonality. Clearly, $\pi_E(X) \in E$. That is, $\pi_E(X)$ can be written as a constant, say $\pi_E(X) = a$ *a.s.* Then

$$X - \pi_E(X) = X - a \perp \mathbf{1},$$

where $\mathbf{1} \in E$ is the random variable equal to one almost surely. This means that

$$0 = \langle X - a, \mathbf{1} \rangle = \mathbb{E}[(X-a) \times 1] = \mathbb{E}[X] - a,$$

which is equivalent to $a = \mathbb{E}[X]$. Finally,

$$\pi_E(X) = \mathbb{E}[X].$$

2nd method: using minimization. According to Theorem 2.5.5, $a = \pi_E(X)$ is the solution of

$$\mathbb{E}[(X-a)^2] = \min_{b \in \mathbb{R}} \mathbb{E}[(X-b)^2].$$

Let us minimize the function $f(b) = \mathbb{E}[(X-b)^2]$. We have

$$\begin{aligned} f(b) &= \mathbb{E}[X^2] + \mathbb{E}[b^2] + \mathbb{E}[-2bX] = \mathbb{E}[X^2] + b^2 - 2b\mathbb{E}[X] \\ f'(b) &= 0 + 2b - 2\mathbb{E}[X] \\ f''(b) &= 2. \end{aligned}$$

As $f''(b) > 0$ for all $b \in \mathbb{R}$, f is convex. Moreover, $f'(b) = 0$ is equivalent to $b = \mathbb{E}[X]$. Hence, f is minimal at $b = \mathbb{E}[X]$. Therefore,

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{b \in \mathbb{R}} \mathbb{E}[(X-b)^2].$$

We have proved (again) that $\pi_E(X) = \mathbb{E}[X]$.

2.6 Swapping $\mathbb{E}$ and limit

When can we swap expectations and limits? When is $\lim_{n\to\infty} \mathbb{E}[X_n]$ equal to $\mathbb{E}[X]$, where X is the almost sure limit of $(X_n)_{n\geq 1}$?

Theorem 2.6.1 (Monotone convergence theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of* **non-negative** *random variables. Assume that*

- *$(X_n)_{n\geq 1}$ is non-decreasing, i.e. $(X_n(\omega))_{n\geq 1}$ is non-decreasing for almost every ω,*
- *$(X_n)_{n\geq 1}$ converges almost surely to some limit X, i.e.*

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

Then

$$\mathbb{E}[X] = \lim_{n\to+\infty} \mathbb{E}[X_n].$$

An important application of monotone convergence is the following result.

Proposition 2.6.2 (Swapping $\sum$ and $\mathbb{E}$)**.** *Let $(X_k)_{k\geq 1}$ be a sequence of non-negative random variables. Then*

$$\mathbb{E}\left[\sum_{k=0}^{\infty} X_k\right] = \sum_{k=0}^{\infty} \mathbb{E}[X_k],$$

where both sides can be infinite.

For the proof note that $\left(\sum_{k=0}^{n} X_k\right)_{n\geq 1}$ is a non-decreasing sequence converging to $\sum_{k=0}^{\infty} X_k$, so that Theorem 2.6.1 applies.

To deal with arbitrary sequences, a different assumption is required: domination.

Theorem 2.6.3 (Dominated-convergence theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that*

- *$(X_n)_{n\geq 1}$ converges to X almost surely,*
- *all random variables X_n are* **dominated** *by some integrable random variable Y, i.e. for all $n \geq 1$ and for all $\omega \in \Omega$, $|X_n(\omega)| \leq |Y(\omega)|$ and $\mathbb{E}[|Y|] < +\infty$.*

Then

$$\mathbb{E}[X] = \lim_{n\to+\infty} \mathbb{E}[X_n].$$

In fact, under these assumptions, we have the even stronger result

$$\lim_{n\to+\infty} \mathbb{E}[|X_n - X|] = 0.$$

Finally, we state a useful result on swapping derivatives and expectations, which is obtained by the dominated convergence theorem.

Theorem 2.6.4 (Differentiating inside expectations)**.** *Let I be an interval and a function*

$$\begin{array}{rccc} f: & I \times \mathbb{R} & \to & \mathbb{R} \\ & (t, X) & \mapsto & f(t, X). \end{array}$$

Assume that for every $t \in I$, the random variable $X \mapsto f(t, X)$ is integrable and that there is a function g such that $\left|\frac{\partial}{\partial t} f(t, X)\right| \leq g(X)$ with $\mathbb{E}[g(X)] < +\infty$. Then

$$\frac{\partial}{\partial t}\mathbb{E}[f(t, X)] = \mathbb{E}\left[\frac{\partial}{\partial t} f(t, X)\right].$$

Chapter 3

Random vectors

3.1 Definition

A **random vector** is the collection of a finite number of random variables $X_1, \ldots, X_d$ in a vector $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$. The **joint distribution** of $\mathbf{X} = (X_1, \ldots, X_d)$, denoted by $\mathbb{P}_{\mathbf{X}}$ or $\mathbb{P}_{(X_1,\ldots,X_d)}$, is the measure on $\mathbb{R}^d$ defined by

$$\mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}\left((X_1, \ldots, X_d) \in A\right) \quad \text{for } A \subset \mathbb{R}^d.$$

The distribution of an element X_k of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is called the **marginal distribution** of X_k. The **cumulative distribution function** $F_{\mathbf{X}} : \mathbb{R}^d \to [0, 1]$ of $\mathbf{X}$ is defined by

$$F_{\mathbf{X}}(t_1, \ldots, t_d) = \mathbb{P}(X_1 \leq t_1, \ldots, X_d \leq t_d), \quad (t_1, \ldots, t_d)^T \in \mathbb{R}^d.$$

The characteristic function is also defined for random vectors and has the same properties as in the univariate case, in particular Theorem 2.4.3 continues to hold. The **characteristic function** $\Phi_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{C}$ of a random vector $\mathbf{X}$ is defined as

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[\mathrm{e}^{i\mathbf{t}^T\mathbf{X}}\right] = \mathbb{E}\left[\exp\left\{i\sum_{k=1}^{d} t_k X_k\right\}\right], \quad \mathbf{t} = (t_1, \ldots, t_d)^T \in \mathbb{R}^d.$$

The **mean vector** of $\mathbf{X}$ is the column vector of the expectations of the elements of $\mathbf{X}$ given by

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}.$$

The goal of the present chapter is to introduce tools to study the distribution of random vectors $\mathbf{X} \in \mathbb{R}^d$.

3.2 Joint and marginal densities

We first present some technical results that are necessary to properly define and handle multiple integrals that occur with densities of random vectors.

Fubini theorems

To lighten notations in the following, we write formulas only for the case $d = 2$, but all results are valid in the general d-dimensional case.

Theorem 3.2.1 (First Fubini theorem)**.** *Let f be a* **non-negative** *function*

$$\begin{array}{rccc} f: & \mathbb{R}\times\mathbb{R} & \to & \mathbb{R}_+ \\ & (x,y) & \mapsto & f(x,y). \end{array}$$

Then

$$\int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} f(x,y)\mathrm{d}x \right) \mathrm{d}y = \int_{x\in\mathbb{R}} \left(\int_{y\in\mathbb{R}} f(x,y)\mathrm{d}y \right) \mathrm{d}x,$$

where both sides might be equal to $+\infty$. Thus, the multiple integral $\iint_{\mathbb{R}^2} f(x,y)\mathrm{d}x\mathrm{d}y$. is defined without ambiguity.

Theorem 3.2.2 (Second Fubini theorem)**.** *Let f be a any real-valued function*

$$\begin{array}{rccc} f: & \mathbb{R}\times\mathbb{R} & \to & \mathbb{R} \\ & (x,y) & \mapsto & f(x,y). \end{array}$$

Assume that $\iint_{\mathbb{R}^2} |f|\mathrm{d}x\mathrm{d}y$ is finite. Then

$$\int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} f(x,y)\mathrm{d}x \right) \mathrm{d}y = \int_{x\in\mathbb{R}} \left(\int_{y\in\mathbb{R}} f(x,y)\mathrm{d}y \right) \mathrm{d}x,$$

where both sides are necessarily finite. Thus, the multiple integral $\iint_{\mathbb{R}^2} f(x,y)\mathrm{d}x\mathrm{d}y$ is well defined and finite.

Joint and marginal densities

Definition 3.2.3 (Joint density)**.** *We say that (X,Y) has joint density $f : \mathbb{R}^2 \to \mathbb{R}_+$ if for any event $A \subset \mathbb{R}^2$*

$$\mathbb{P}\left((X,Y)\in A\right) = \iint_A f(x,y)\mathrm{d}x\mathrm{d}y.$$

Then, for every integrable function ϕ,

$$\mathbb{E}\left[\phi(X,Y)\right] = \iint_{\mathbb{R}^2} \phi(x,y)f(x,y)\mathrm{d}x\mathrm{d}y. \tag{3.1}$$

This quantity is well-defined either if $\phi \geq 0$ or if $\iint_{\mathbb{R}^2} |\phi(x,y)|f(x,y)\mathrm{d}x\mathrm{d}y < +\infty$ according to the Fubini theorems.

The joint density of (X,Y) is a function defined on $\mathbb{R}^2$, see Figure 3.1 for an example.

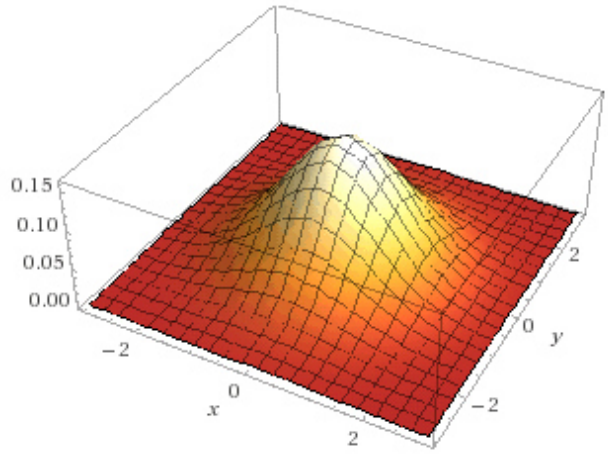


Figure 3.1: Joint density of (X, Y), where X and Y are independent having standard Gaussian distribution $\mathcal{N}(0, 1)$.

Example. Consider the function $f(x, y) = (x + y)\mathbb{1}_{[0,1]\times[0,1]}(x, y)$. Let us show that f is a density. Clearly, f is non-negative. Now, by the first Fubini theorem,

$$\begin{aligned}\iint_{\mathbb{R}^2} f(x,y)\mathrm{d}x\mathrm{d}y &= \iint_{[0,1]\times[0,1]} (x+y)\mathrm{d}x\mathrm{d}y = \int_{x\in[0,1]} \left(\int_{y\in[0,1]} (x+y)\mathrm{d}y\right)\mathrm{d}x \\ &= \int_{x\in[0,1]} \left(\int_{y\in[0,1]} x\mathrm{d}y + \int_{y\in[0,1]} y\mathrm{d}y\right)\mathrm{d}x \\ &= \int_{x\in[0,1]} (x+1/2)\mathrm{d}x \\ &= \frac{1}{2} + \frac{1}{2} = 1.\end{aligned}$$

Thus, f is a density function.

Proposition 3.2.4 (Marginal densities)**.** *If (X, Y) has density $(x, y) \mapsto f(x, y)$ then X has density $x \mapsto f_X(x) := \int_{y\in\mathbb{R}} f(x, y)\mathrm{d}y$ and f_X is called the* **marginal density** *of X. Likewise, Y has density $y \mapsto f_Y(y) := \int_{x\in\mathbb{R}} f(x, y)\mathrm{d}x$.*

Proof. For any bounded and continuous function ϕ of a single variable x, we apply formula (3.1) to obtain

$$\mathbb{E}\left[\phi(X)\right] = \iint \phi(x)f(x, y)\mathrm{d}x\mathrm{d}y.$$

As the function $|\phi|$ is bounded by some $c > 0$, we have

$$\iint |\phi(x)f(x, y)|\mathrm{d}x\mathrm{d}y \leq \iint cf(x, y)\mathrm{d}x\mathrm{d}y = c\iint f(x, y)\mathrm{d}x\mathrm{d}y = c \times 1 < +\infty.$$

Hence, according to the second Fubini theorem, we can first integrate with respect to y and write

$$\mathbb{E}\left[\phi(X)\right] = \int_x \phi(x) \underbrace{\left(\int_y f(x,y)\mathrm{d}y\right)}_{=f_X(x)} \mathrm{d}x.$$

Applying Theorem 2.4.1, this completes the proof. □

Example (continued). Let the random vector (X,Y) have density $f(x+y) = (x+y)\mathbb{1}_{[0,1]\times[0,1]}(x,y)$. By the proposition, the marginal density f_X of X is given by

$$f_X = \int_{y\in[0,1]} (x+y)\mathrm{d}y = x + \frac{1}{2}, \qquad \text{if } x \in [0,1],$$

and $f_X(x) = 0$ otherwise.

It occurs that random variables X and Y both have densities, while the random vector (X,Y) has not. For example, consider the case where $X \sim \mathcal{E}(1)$ and $Y = 2X$. Then the pair $(X,Y) = (X,2X)$ takes its values on the line $D = \{y = 2x\}$ with probability one, that is

$$\mathbb{P}((X,Y) \in D) = 1.$$

However, if (X,Y) had a density f, then the probability $\mathbb{P}((X,Y) \in D)$ would be zero, as D has Lebesgue measure zero. More precisely,

$$\begin{aligned}
\mathbb{P}((X,Y) \in D) = \mathbb{E}[\mathbb{1}_{Y=2X}] &= \iint \mathbb{1}_{y=2x} f(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_x \left(\int_y \mathbb{1}_{y=2x} f(x,y)\mathrm{d}y\right)\mathrm{d}x \\
&= \int_x \left(\int_{y=2x}^{2x} f(x,y)\mathrm{d}y\right)\mathrm{d}x \\
&= \int_x 0 \times \mathrm{d}y = 0.
\end{aligned}$$

3.3 Independence of random variables

Definition 3.3.1 (Independence of random variables)**.** *Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space $(\Omega, \mathbb{P})$. We say that $X_1, \ldots, X_n$ are (mutually)* **independent** *if for all events $B_1, \ldots, B_n$,*

$$\mathbb{P}\left(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i).$$

Moreover, we say that the random variables $X_1, X_2, \dots$ are **independent and identically distributed (i.i.d.)**, if, for every n, $X_1, \dots, X_n$ are independent and the random variables X_i have all the same distribution.

If X and Y are discrete random variables taking their values in $\{x_1, x_2, \dots\}$ and $\{y_1, y_2, \dots\}$, respectively, then X and Y are independent if and only if

$$\mathbb{P}(\{X = x_i\} \cap \{Y = y_j\}) = \mathbb{P}(X = x_i)\mathbb{P}(Y = y_j) \quad \text{for all } \ i, j.$$

The following proposition states some useful consequences of independence.

Proposition 3.3.2. *Let $X_1, X_2, \dots, X_n$ be independent and $\phi_1, \dots, \phi_n$ be some functions.*

(i) Then $\phi_1(X_1), \phi_2(X_2), \dots, \phi_n(X_n)$ are independent.

(ii) If all random variables $\phi_k(X_k)$ are integrable, then

$$\mathbb{E}\left[\phi_1(X_1)\phi_2(X_2)\dots\phi_n(X_n)\right] = \mathbb{E}[\phi_1(X_1)]\mathbb{E}[\phi_2(X_2)]\dots\mathbb{E}[\phi_n(X_n)].$$

Independence of random variables can also be described by properties of the cumulative distribution function and the characteristic function.

Theorem 3.3.3. *For any random variables $X_1, \dots, X_n$ the following assertions are equivalent:*

(i) $X_1, \dots, X_n$ are independent.

(ii) The cumulative distribution function $F_{\mathbf{X}}$ of the random vector $\mathbf{X} = (X_1, \dots, X_n)$ can be factorized as follows

$$F_{\mathbf{X}}(t_1, \dots, t_n) = F_{X_1}(t_1)\dots F_{X_n}(t_n), \quad \text{for all } (t_1, \dots, t_n)^T \in \mathbb{R}^n,$$

where F_{X_k} denotes the cumulative distribution function of the component X_k.

(iii) The characteristic function $\Phi_{\mathbf{X}}$ of the random vector $\mathbf{X} = (X_1, \dots, X_n)$ can be factorized as follows

$$\Phi_{\mathbf{X}}(t_1, \dots, t_n) = \Phi_{X_1}(t_1)\dots\Phi_{X_n}(t_n), \quad \text{for all } (t_1, \dots, t_n)^T \in \mathbb{R}^n.$$

Independence and densities

Theorem 3.3.4. *Let X and Y be two random variables.*

- *If X and Y are independent and continuous with densities f_X and f_Y, respectively, then the random vector (X, Y) is continuous with density*

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y), \quad \text{for all } x, y \in \mathbb{R}.$$

- *Conversely, if the random vector (X, Y) has a density $f_{(X,Y)}$ and if there are functions g_1 and g_2 such that the joint density can be written as a product*

$$f_{(X,Y)}(x, y) = g_1(x)g_2(y), \quad \textit{for (almost) all } x, y \in \mathbb{R},$$

then X and Y are independent.

Example. Assume that the random vector (X, Y) has density $f(x, y) = 6x^2y\mathbb{1}_{(x,y)\in[0,1]^2}$. (Exercise: check that this is a density). Then, according to Theorem 3.3.4, X and Y are independent, since one can write

$$f(x, y) = 6x^2y\mathbb{1}_{[0,1]\times[0,1]}(x, y) = 6x^2\mathbb{1}_{x\in[0,1]} \times y\mathbb{1}_{y\in[0,1]}.$$

Though, to find the marginal densities of X and Y we have to care about constants. The marginal density f_X of X is given by

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\mathrm{d}y \int_{y=0}^{1} 6x^2y\mathrm{d}y = 6x^2 \int_{y=0}^{1} y\mathrm{d}y = 3x^2 \quad \text{for } x \in [0, 1].$$

Finally,

$$f_X(x) = 6x^2y\mathbb{1}_{x,y\in[0,1]} = f_X(x)2y\mathbb{1}_{y\in[0,1]}.$$

So, the marginal density f_Y of Y is necessarily given by the remaining term, that is $f_Y(y) = 2y\mathbb{1}_{y\in[0,1]}$.

We conclude the section by the result that independent continuous random variables never take the same value.

Proposition 3.3.5 (Continuous random variables are distinct)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with densities f_k for $k = 1, 2 \ldots$. Then the random variables $X_1, X_2, \ldots$ are all pairwise distinct with probability one, that is,*

$$\mathbb{P}\left(X_i \neq X_j, \textit{ for all } i \neq j\right) = 1.$$

Proof. We prove that the complement event, that is {there exist $i \neq j$ such that $X_i = X_j$}, has probability zero. First notice that

$$\begin{aligned}\mathbb{P}(\text{there exist } i \neq j \text{ such that } X_i = X_j) &= \mathbb{P}\left(\cup_{i\neq j}\left\{X_i = X_j\right\}\right) \\ &\leq \sum_{i\neq j} \mathbb{P}\left(X_i = X_j\right),\end{aligned}$$

by the union bound (Proposition 1.2.3). Now,

$$\begin{aligned}\mathbb{P}\left(X_i = X_j\right) = \mathbb{E}\left[\mathbb{1}_{X_i=X_j}\right] &= \int_{\mathbb{R}^2} \mathbb{1}_{x=y} f_i(x)f_j(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} \mathbb{1}_{x=y} f_i(x)\mathrm{d}x\right) f_j(y)\mathrm{d}y \\ &= \int_{y\in\mathbb{R}} \left(\int_{x=y}^{y} f_i(x)\mathrm{d}x\right) f_j(y)\mathrm{d}y \\ &= \int_{y\in\mathbb{R}} 0 \times f_j(y)\mathrm{d}y = 0.\end{aligned}$$

This completes the proof. □

3.4 Sums of independent random variables

Let X and Y be random variables. What can we say about $X+Y$? First, by linearity of expectation

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

Now, assume that X and Y are independent. Then

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

More generally, for *independent* random variables $X_1, \ldots, X_n$, expectations and variances sum up, that is,

$$\begin{aligned}\mathbb{E}[X_1 + \cdots + X_n] &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n], \\ \mathrm{Var}(X_1 + \cdots + X_n) &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).\end{aligned}$$

Now concerning the distribution of the sum of independent random variables, can we say something about it in general? Let us consider the case where X and Y have densities f and g, respectively. What is the density (if any) of $X+Y$? For any bounded and continuous function ϕ we obtain

$$\begin{aligned}\mathbb{E}[\phi(X+Y)] &= \int\int \phi(x+y) f(x) g(y) \mathrm{d}x \mathrm{d}y \\ &= \int_x f(x) \left(\int_y \phi(x+y) g(y) \mathrm{d}y \right) \mathrm{d}x \quad \text{(by the second Fubini theorem)} \\ &= \int_x f(x) \left(\int_u \phi(u) g(x-u) du \right) \mathrm{d}x \quad \text{(by changing variables } u = x+y,\ \tfrac{du}{\mathrm{d}y} = 1) \\ &= \int_u \phi(u) \underbrace{\left(\int_x f(x) g(u-x) \mathrm{d}x \right)}_{\text{density of } X+Y} \mathrm{d}u \quad \text{(again by Fubini).}\end{aligned}$$

Hence, by Theorem 2.4.1, we have shown that there is a general formula for the density of $X+Y$, which is the so-called convolution of f and g. This result is summarized in the following theorem.

Theorem 3.4.1 (Convolution of densities)**.** *Let X and Y be independent random variables with densities f and g, respectively. Then the sum $X+Y$ has a density given by*

$$f_{X+Y}(t) = f * g\ (t),$$

where $u \mapsto f * g\ (u)$ is the **convolution function** *of f and g defined as*

$$f * g\ (u) = \int_{x \in \mathbb{R}} f(x) g(u-x) \mathrm{d}x.$$

Obviously, $X+Y$ has the same density as $Y+X$. So we have $f*g=g*f$, i.e.

$$\int_{x\in\mathbb{R}} f(x)g(u-x)\mathrm{d}x = \int_{x\in\mathbb{R}} g(x)f(u-x)\mathrm{d}x,$$

for every u. This can also be checked by the change of variables $v=u-x$.

Example. Let X and Y be i.i.d. with exponential distribution $\mathcal{E}(1)$, i.e. with density $f(x)=\mathrm{e}^{-x}\mathbb{1}_{x\geq 0}$. Clearly, the sum $U=X+Y$ takes its values in $[0,+\infty)$. For all $u>0$, its density f_U is given by

$$\begin{aligned} f_U(u) = f*f(u) &= \int_{x=0}^{\infty} f(x)f(u-x)\mathrm{d}x = \int_{x=0}^{\infty} \mathrm{e}^{-x}\mathrm{e}^{-(u-x)}\mathbb{1}_{u-x\geq 0}\,\mathrm{d}x \\ &= \mathrm{e}^{-u}\int_{x=0}^{\infty} \mathbb{1}_{u-x\geq 0}\,\mathrm{d}x = \mathrm{e}^{-u}\int_{x=0}^{\infty} \mathbb{1}_{x\leq u}\,\mathrm{d}x = \mathrm{e}^{-u}\int_{x=0}^{u}\mathrm{d}x = u\mathrm{e}^{-u}. \end{aligned}$$

Hence $U=X+Y$ has density $u\mathrm{e}^{-u}\mathbb{1}_{u\geq 0}$ (you can check that this is a density). By the way, we see that the sum of two independent exponentially distributed random variables is no longer exponentially distributed.

Sums of random variables and characteristic functions

Another very efficient tool to handle sums of random variables is the characteristic function. Indeed, for any independent random variables X and Y we have

$$\begin{aligned} \Phi_{X+Y}(t) &= \mathbb{E}[\mathrm{e}^{it(X+Y)}] = \mathbb{E}[\mathrm{e}^{itX}\mathrm{e}^{itY}] \\ \text{(by independence)} \qquad &= \mathbb{E}[\mathrm{e}^{itX}]\mathbb{E}[\mathrm{e}^{itY}] \\ &= \Phi_X(t)\Phi_Y(t). \end{aligned} \tag{3.2}$$

Hence, the characteristic function of the sum of independent random variables is the product of the characteristic functions of the individual random variables.

An important application is the following result.

Proposition 3.4.2 (Sum of independent gaussian random variables)**.** *If X and Y are independent and normally distributed, i.e. $X\sim\mathcal{N}(\mu_1,\sigma_1^2)$ and $Y\sim\mathcal{N}(\mu_2,\sigma_2^2)$, then $X+Y$ is also a gaussian random variable. More precisely,*

$$X+Y\sim\mathcal{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2).$$

Proof. We will later see in Proposition A.1.1 that the characteristic function of the gaussian distribution $\mathcal{N}(\mu,\sigma^2)$ is given by

$$\Phi_X(t)=\exp\left(it\mu-\frac{t^2\sigma^2}{2}\right).$$

So according to (3.2), the characteristic function of X_1+X_2 is given by

$$\begin{aligned} \Phi_{X_1+X_2}(t) &= \Phi_{X_1}(t)\Phi_{X_2}(t) \\ &= \exp\left(it\mu_1-\frac{t^2\sigma_1^2}{2}\right)\exp\left(it\mu_2-\frac{t^2\sigma_2^2}{2}\right) \\ &= \exp\left(it(\mu_1+\mu_2)-\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right). \end{aligned}$$

Obviously, the last term is the characteristic function of the gaussian distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. □

Alternatively, Proposition 3.4.2 can be shown by computing the convolution of two normal densities. However, this is much more computational than using the characteristic function.

3.5 Covariance matrix

When random variables are not independent, it may be interesting to consider the covariance.

Definition 3.5.1 (Covariance of two random variables)**.** *Let X and Y be two random variables in L^2. The* **covariance** *of X and Y is defined by*

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Proposition 3.5.2 (Properties of the covariance)**.** *Let $X, Y, X_i, Y_j \in L^2$. Then,*

(i) **(Symmetry)** $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

(ii) **(Bilinearity)** *For any constants $a_i, b_i \in \mathbb{R}$,*

$$\mathrm{Cov}\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_{i,j} a_i b_j \mathrm{Cov}\left(X_i, Y_j\right).$$

(iii) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(iv) If X and Y are independent, $\mathrm{Cov}(X,Y) = 0$.

(v) For any constant c, $\mathrm{Cov}(X,c) = 0$.

(vi) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$.

The converse of property (iv) is not true. That is, a covariance equal to zero does generally not imply independence.

Definition 3.5.3 (Covariance matrix of a random vector)**.** *Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector such that $X_k \in L^2$ for $k = 1, \ldots, d$. The* **covariance matrix** $\mathrm{Cov}(\mathbf{X})$ *of $\mathbf{X}$ is the $d \times d$ matrix with entries $\mathrm{Cov}(X_i, X_j)$. More precisely,*

$$\mathrm{Cov}(\mathbf{X}) = \left(\mathrm{Cov}(X_i, X_j)\right)_{i,j} = \begin{array}{c} \\ i \end{array} \begin{array}{c} \begin{array}{c} j \end{array} \\ \begin{pmatrix} \vdots & \\ \ldots & \mathrm{Cov}(X_i, X_j) \quad \ldots \\ \vdots & \end{pmatrix} \end{array}.$$

Example. Let $X_1, \dots, X_d$ be i.i.d. with $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = \sigma^2$. Then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ by independence. Thus, the covariance matrix of $\mathbf{X} = (X_1, \dots, X_d)$ is given by

$$\mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{pmatrix}.$$

By using vector and matrix notation, the covariance matrix is easier to handle. Note that

$$(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T = \begin{pmatrix} X_1 - \mathbb{E}[X_1] \\ X_2 - \mathbb{E}[X_2] \\ \vdots \\ X_d - \mathbb{E}[X_d] \end{pmatrix} \begin{pmatrix} X_1 - \mathbb{E}[X_1] & X_2 - \mathbb{E}[X_2] & \dots & X_d - \mathbb{E}[X_d] \end{pmatrix}$$

$$= \begin{array}{c} \\ \\ i \end{array} \begin{array}{c} \begin{array}{ccc} & j & \end{array} \\ \begin{pmatrix} & \vdots & \\ \dots & (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j]) & \dots \\ & \vdots & \end{pmatrix} \end{array}.$$

By taking the expectation we see that the covariance matrix of $\mathbf{X}$ can be written as

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]. \tag{3.3}$$

Proposition 3.5.4 (Properties of the covariance matrix)**.** *Let $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector such that $X_k \in L^2$ for $k = 1, \dots, d$.*

(i) The covariance matrix $\mathrm{Cov}(\mathbf{X})$ is symmetric.

(ii) The covariance matrix $\mathrm{Cov}(\mathbf{X})$ is positive semi-definite, i.e. for all $\mathbf{t} = (t_1, \dots, t_d) \in \mathbb{R}^d$ we have $\mathbf{t}^T \mathrm{Cov}(\mathbf{X}) \mathbf{t} \geq 0$.

Proof. Property (i) is clear, since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$. To show (ii), we obtain for any $\mathbf{t} = (t_1, \dots, t_d)^T \in \mathbb{R}^d$

$$\begin{aligned} \mathbf{t}^T \mathrm{Cov}(\mathbf{X}) \mathbf{t} &= \mathbf{t}^T \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right] \mathbf{t} \\ &= \mathbb{E}\left[\mathbf{t}^T (\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{t}\right] \\ &= \mathbb{E}\left[((\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{t})^T (\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{t}\right] \qquad (\text{since } u^T v = (v^T u)^T) \\ &= \mathbb{E}\left[\|(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{t}\|^2\right] \\ &\geq 0, \end{aligned}$$

since $\|(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{t}\|^2 \geq 0$ almost surely. □

Theorem 3.5.5 (Linear transformation of a random vector)**.** *Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with finite covariance matrix. Let M be a $p \times d$ matrix and $a \in \mathbb{R}^p$ a vector. Then the mean and the covariance vector of the random vector $M\mathbf{X} + a \in \mathbb{R}^p$ are given by*

$$\mathbb{E}[M\mathbf{X} + a] = M\mathbb{E}[\mathbf{X}] + a \quad \text{and} \quad \mathrm{Cov}(M\mathbf{X} + a) = M\mathrm{Cov}(\mathbf{X})M^T.$$

Proof. As we can write

$$M\mathbf{X} = \begin{pmatrix} \sum_j M_{1,j} X_j \\ \sum_j M_{2,j} X_j \\ \vdots \\ \sum_j M_{d,j} X_j \end{pmatrix},$$

the first assertion is just a consequence of the linearity of the expectation. By applying formula (3.3), we obtain for the covariance matrix

$$\begin{aligned} \mathrm{Cov}(M\mathbf{X} + a) &= \mathbb{E}\left[(M\mathbf{X} + a - \mathbb{E}[M\mathbf{X} + a])(M\mathbf{X} + a - \mathbb{E}[M\mathbf{X} + a])^T\right] \\ &= \mathbb{E}\left[M(\mathbf{X} - \mathbb{E}[\mathbf{X}])\left(M(\mathbf{X} - \mathbb{E}[\mathbf{X}])\right)^T\right] \\ &= \mathbb{E}\left[M(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T M^T\right] \qquad (\text{since } (Au)^T = u^T A^T) \\ &= M\mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}]))^T\right] M^T \\ &= M\mathrm{Cov}(\mathbf{X})M^T. \end{aligned}$$

This completes the proof. □

3.6 Correlation

Definition 3.6.1 (Correlation)**.** *Let $X, Y \in L^2$ and $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$. The* **correlation** *or* **Pearson's correlation coefficient** *between X and Y is defined by*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

We say that X and Y have

- **positive correlation** if $\rho_{X,Y} > 0$ and
- **negative correlation** if $\rho_{X,Y} < 0$.

The correlation has the following basic properties.

Proposition 3.6.2. *Let $X, Y \in L^2$ and $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.*

- *If X and Y are independent, then $\rho_{X,Y} = 0$.*
- *The correlation is invariant with respect to affine transformations: for any $a \neq 0$, $b \neq 0$ $c, d \in \mathbb{R}$,*

  $$\rho_{aX+c,bY+d} = \mathrm{sign}(ab)\rho_{X,Y},$$

  where $\mathrm{sign}(u) = \mathbb{1}\{u > 0\} + \mathbb{1}\{u < 0\}$.

The power of the correlation lies in the following result.

Theorem 3.6.3. *Let $X, Y \in L^2$ and $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.*

- $-1 \leq \rho_{X,Y} \leq 1$.

- $\rho_{X,Y} = 1 \iff \exists a > 0, b \in \mathbb{R}$ *such that* $Y = aX + b$ *a.s.*

- $\rho_{X,Y} = -1 \iff \exists a < 0, b \in \mathbb{R}$ *such that* $Y = aX + b$ *a.s.*

We see that the correlation is a bounded indicator with values in the interval $[-1, 1]$ as opposed to the covariance taking its values in whole $\mathbb{R}$. For a covariance, it is impossible to say if its value is large or not, but for the correlation coefficient the possible extreme values are finite. Moreover, these extreme values have a clear and unique interpretation: they occur when there is an affine relationship between the random variables X and Y. Hence, the correlation $\rho_{X,Y}$ measures the degree of a **linear** relation between X and Y. A zero correlation does not necessarily mean that X and Y are independent, on the contrary, they may even be related by a deterministic relation, but then this one must be nonlinear.

Chapter 4

Convergence of random variables

In probability theory there are different ways to define what it means that a sequence of random variables $(X_n)_{n\geq 1}$ converges to some random variable X.

4.1 Different types of convergence

Definition 4.1.1. *Let X and $X_n, n \geq 1$ be random variables defined on the same probability space $(\Omega, \mathbb{P})$.*

- *The sequence $(X_n)_{n\geq 1}$ converges to X* **in probability** *if for any real number $\varepsilon > 0$*

$$\lim_{n\to+\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

  We write $X_n \xrightarrow{P} X$ or $X_n \xrightarrow{\mathbb{P}} X$.

- *The sequence $(X_n)_{n\geq 1}$ converges to X* **in L^p** *if*

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

  We write $X_n \xrightarrow{L^p} X$, and also say that (X_n) converges to X in p-th mean.

- *The sequence $(X_n)_{n\geq 1}$ converges to X* **almost surely** *if*

$$\mathbb{P}\left(\left\{\omega \text{ such that } \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

  We write $X_n \xrightarrow{a.s.} X$ or $X_n \longrightarrow X$ a.s..

These types of convergence do all concern sequences of random variables. A different type of convergence is convergence in distribution, where only distributions are considered and no sequence $(X_n)_{n\geq 1}$ is required. For example, one can state that the binomial distribution $B(n, \lambda/n)$ converges to the Poisson distribution $P(\lambda)$ when n tends to infinity, without giving a precise sequence $(X_n)_{n\geq 1}$. Convergence in distribution is studied in Chapter 4.2.

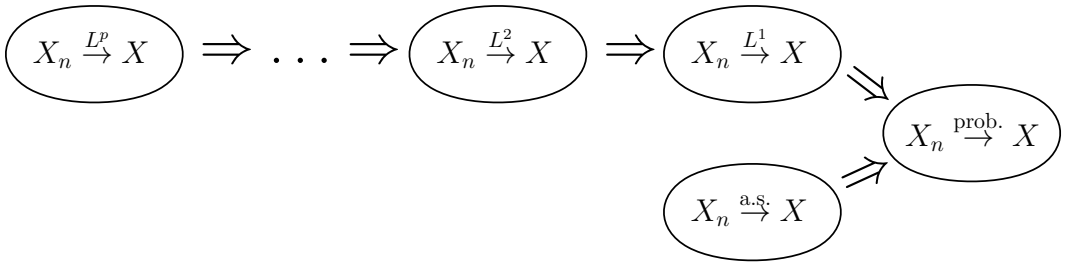


Figure 4.1: Relations between the different types of convergence.

For statistical applications, convergence in probability is the most useful among the types of convergence defined above.
Here is an example that shows that these types of convergence are indeed not equivalent.

Example. $\left(\xrightarrow{P} \neq \xrightarrow{L^p}\right)$ Take a sequence of independent random variables $X_1, X_2, \ldots$ such that

$$X_n = \begin{cases} \sqrt{n} & \text{with probability } \frac{1}{n}, \\ 0 & \text{with probability } 1 - \frac{1}{n}. \end{cases}$$

When n goes large, X_n is more and more likely to be zero, so we expect $(X_n)_{n\geq 1}$ to converge (at least in some sense) to zero.
Let us first check that $X_n \xrightarrow{P} 0$: fix a small $\varepsilon > 0$, we have

$$\mathbb{P}\left(|X_n - 0| > \varepsilon\right) = \mathbb{P}\left(X_n = \sqrt{n}\right) = 1/n \longrightarrow 0,$$

when $n \to \infty$. This proves that $X_n \xrightarrow{P} 0$.
Let us now consider the convergence to 0 in L^p.

$$\mathbb{E}\left[\ |X_n - 0|^p\ \right] = \mathbb{E}[|X_n|^p] = 0 \times (1 - \tfrac{1}{n}) + (\sqrt{n})^p \times \tfrac{1}{n} = n^{p/2-1}.$$

This goes to zero for $p < 2$. Hence, $X_n \xrightarrow{L^p} X$ for all $1 \leq p < 2$.
The different convergence types are related in the following way.

Proposition 4.1.2. *(i) If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{P} X$.*

(ii) If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.

(iii) if $X_n \xrightarrow{L^q} X$ for some q, then $X_n \xrightarrow{L^p} X$ for every $p < q$.

We see that convergence in probability is in fact the weakest type of convergence, see Figure 4.1.

Proof. Assume that $X_n \xrightarrow{L^p} X$ and fix $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(|X_n - X| > \varepsilon\right) &= \mathbb{P}\left(|X_n - X|^p > \varepsilon^p\right), && \text{(this is the same event)} \\ &\leq \frac{\mathbb{E}[\ |X_n - X|^p\ ]}{\varepsilon^p}, && \text{(by Markov's inequality)} \end{aligned}$$

which goes to zero by assumption (ε is fixed and $n \to +\infty$). This implies $X_n \xrightarrow{P} X$.
Assertion (ii) is admitted.
To prove (iii), according to Theorem 2.5.2, for $p < q$,

$$\|X_n - X\|_p \leq \|X_n - X\|_q.$$

Thus, when the right-hand side goes to zero, so does the left-hand side. □

The following proposition is sometimes useful.

Proposition 4.1.3 (Preservation of convergence)**.** *Let (X_n) and (Y_n) be sequences of random variables defined on the same probability space $(\Omega, \mathbb{P})$.*

- *Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.*
- *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$ and $X_n Y_n \xrightarrow{P} XY$.*

The proposition also holds for almost sure convergence.

4.2 Convergence of distributions

We now discuss a different type of convergence: convergence of distributions of random variables instead of convergence of random variables themselves.

Definition

To introduce the concept, we may still consider a sequence $X_1, X_2, \ldots$ of random variables, but unlike to the previous part of the chapter, they may be defined on different probability spaces.

Definition 4.2.1. *One says that $(X_n)_{n\geq 1}$ converges* **in distribution** *(or* **in law***) to X when n tends to infinity, if for every bounded and continuous function ϕ we have*

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

We write $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$.

This kind of convergence is different as it concerns distributions of random variables rather than the random variables themselves. To see why, observe that if $X_1, X_2, \ldots$ are identically distributed, then for all n, X_n has the same distribution as X_1 and as X_2, so that

$$X_n \xrightarrow{d} X_1 \quad \text{but also} \quad X_n \xrightarrow{d} X_2.$$

Thus, the limit is not unique. In fact, it would be more appropriate to write that *the distribution* of X_n converges to *the distribution* of X: you will sometimes find the notation

$$\mathbb{P}_{X_n} \xrightarrow{d} \mathbb{P}_X.$$

With this notation the limit is unique.

One also says that $\mathbb{P}_{X_n}$ **converges weakly** to $\mathbb{P}_X$.

Properties

Convergence in distribution is in fact the weakest of all kinds of convergence.

Proposition 4.2.2. *Let X and $(X_n)_{n\geq 1}$ be random variables defined on the same probability space. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.*

Proof. Let ϕ be a continuous function bounded by some constant $A > 0$, i.e. $|\phi(x)| < A$ for all $x \in \mathbb{R}$. For the sake of simplicity, we furthermore assume that ϕ is also Lipschitz, that is, there exists a constant $c > 0$ such that for all $x, y \in \mathbb{R}$

$$|\phi(x) - \phi(y)| \leq c|x - y|.$$

Fix $\varepsilon > 0$ and write

$$\begin{array}{rcccc} |\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X)]| & \leq & \mathbb{E}\left[|\phi(X_n) - \phi(X)|\right] & & \\ & = & \mathbb{E}\Big[\underbrace{|\phi(X_n) - \phi(X)|}_{\leq c|X_n - X| \text{ (Lipschitz)}} \mathbb{1}_{|X_n - X| \leq \varepsilon}\Big] & + & \mathbb{E}\Big[\underbrace{|\phi(X_n) - \phi(X)|}_{\leq |\phi(X_n)| + |\phi(X)| \leq 2A} \mathbb{1}_{|X_n - X| > \varepsilon}\Big] \\ & \leq & \mathbb{E}\left[c|X_n - X|\mathbb{1}_{|X_n - X| \leq \varepsilon}\right] & + & \mathbb{E}\left[2A\mathbb{1}_{|X_n - X| > \varepsilon}\right] \\ & \leq & \mathbb{E}[c\varepsilon\mathbb{1}_{|X_n - X| \leq \varepsilon}] & + & 2A\mathbb{E}\left[\mathbb{1}_{|X_n - X| > \varepsilon}\right] \\ & \leq & c\varepsilon & + & 2A\mathbb{P}(|X_n - X| > \varepsilon). \end{array}$$

The last probability tends to zero by assumption. This proves that

$$\lim_{n \to +\infty} |\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X)]| \leq c\varepsilon$$

for any $\varepsilon > 0$. Hence, $\lim_{n \to +\infty} |\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X)]| = 0$. □

In the particular case, where the limit is a constant, convergence in distribution and convergence in probability are equivalent.

Proposition 4.2.3. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables. If there is a constant c such that $X_n \xrightarrow{d} c$, then $X_n \xrightarrow{P} c$ as $n \to \infty$.*

To show convergence in distribution in practice, we rarely use the definition, but rather one of the following two criteria.

Theorem 4.2.4. *The following three assertions are equivalent:*

1. *$X_n \xrightarrow{d} X$ as $n \to \infty$.*
2. *$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ for every $t \in \mathbb{R}$ such that F_X is continuous at t.*
3. *$\lim_{n\to\infty} \Phi_{X_n}(t) = \Phi_X(t)$ for every $t \in \mathbb{R}$.*

As an application, let us prove the **law of rare events**: the sum of independent Bernoulli random variables with a small parameter is approximately distributed as a Poisson random variable.

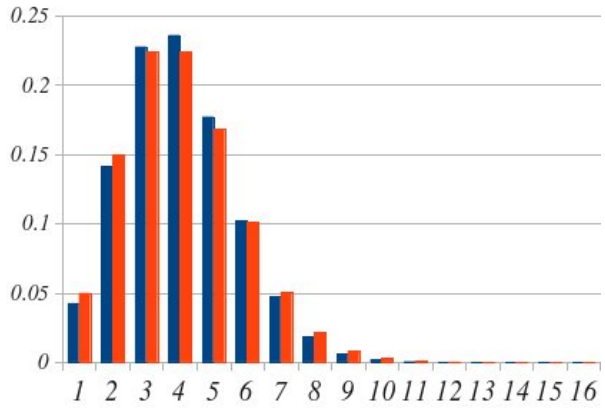


Figure 4.2: Probability mass function of the binomial distribution Binom$(30, 3/30)$ (black) and the Poisson distribution Poisson(3) (red).

Proposition 4.2.5 (Law of rare events)**.** *Let $\lambda > 0$. Then*

$$\text{Binom}(n, \lambda/n) \xrightarrow{d} \text{Poisson}(\lambda), \qquad \text{as } n \to \infty.$$

Proof. Let $B_n \sim \text{Binom}(n, \lambda/n)$. For the characteristic function of B_n we obtain

$$\begin{aligned}
\mathbb{E}[\exp(itB_n)] = \sum_{k=0}^{n} \mathrm{e}^{itk} \mathbb{P}(B_n = k) &= \sum_{k=0}^{n} \mathrm{e}^{itk} \binom{n}{k} (\tfrac{\lambda}{n})^k (1 - \tfrac{\lambda}{n})^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} (\mathrm{e}^{it} \tfrac{\lambda}{n})^k (1 - \tfrac{\lambda}{n})^{n-k} \\
&= \left(1 - \tfrac{\lambda}{n} + \mathrm{e}^{it} \tfrac{\lambda}{n}\right)^n \text{ (by the binomial identity).}
\end{aligned}$$

Now, for each t,

$$\mathbb{E}[\exp(itB_n)] = \left(1 + \frac{\lambda \mathrm{e}^{it} - \lambda}{n}\right)^n \longrightarrow \exp\left(\lambda \mathrm{e}^{it} - \lambda\right) \quad (n \to \infty),$$

where we used that $(1 + u/n)^n \to \exp(u)$, which is true for any real number u and also (but this is not so easy to prove) for complex numbers. Now it remains to show that $\exp(\lambda \mathrm{e}^{it} - \lambda)$ is the characteristic function of the Poisson distribution with parameter λ. Let $X \sim$Poi(λ). Then we obtain

$$\mathbb{E}[\exp(itX)] = \sum_{k \geq 0} \mathrm{e}^{itk} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda} \sum_{k \geq 0} \frac{(\mathrm{e}^{it}\lambda)^k}{k!} = \exp\left(\lambda \mathrm{e}^{it} - \lambda\right).$$

This completes the proof. □

The law of rare events partly explains the relevance of the Poisson distribution in modeling data arising in various fields of application.

The proposition is illustrated in Figure 4.2, where we observe that the probabilities of the number of successes in 30 Bernoulli trials with 10% success are well approximated by the Poisson distribution Poisson(3).

Convergence preservation

Here are two useful properties regarding convergence in distribution.

Proposition 4.2.6 (Preservation of convergence in law)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables, and let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$ as $n \to \infty$.*

In general, it is false to say that the sum of two sequences that converge in distribution also converges in distribution. Only the following weaker result holds in general.

Proposition 4.2.7 (Slutsky's lemma)**.** *Let $(X_n)_{n\geq 1}, (Y_n)_{n\geq 1}$ be sequences of random variables, and let $c \in \mathbb{R}$ be a constant. Assume that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, then the sequence of random vectors $((X_n, Y_n))_{n\geq 1}$ converges in distribution, that is,*

$$(X_n, Y_n) \xrightarrow{d} (X, c) \quad \text{as } n \to \infty.$$

In particular, this implies that

$$X_n Y_n \xrightarrow{d} Xc, \qquad X_n + Y_n \xrightarrow{d} X + c, \qquad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}.$$

4.3 Law of large numbers

In statistics, an important quantity is the sample mean $\bar{X}_n$ of i.i.d. random variables $X_1, X_2, \ldots, X_n$. It is defined as

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The **law of large numbers** (LLN) tells us that the sample mean is an approximation of the expectation $\mathbb{E}[X_1]$. In other words, the sample mean gets closer and closer to $\mathbb{E}[X]$ when the sample size n increases. The precise statement is the following.

Theorem 4.3.1 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. integrable random variables with expectation $\mu = \mathbb{E}[X_1]$ and such that $\mathrm{Var}(X_1) < +\infty$. Then*

$$\bar{X}_n \xrightarrow{L^2} \mu, \quad \text{as } n \to +\infty.$$

Obviously, under the assumptions of the theorem, convergence holds also in probability, that is,

$$\bar{X}_n \xrightarrow{P} \mu, \quad \text{as } n \to +\infty.$$

The term"weak" refers to the fact that the convergence only holds in L^2 and in probability. Just below we will see the "strong" LLN.

Proof. Note that by linearity of the expectation

$$\mathbb{E}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n} \times n\mu = \mu.$$

To prove the theorem, one has to show that the following term tends to zero when n goes to infinity. Indeed, by the properties of the variance stated in Proposition 2.2.6, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(\bar{X}_n - \mu\right)^2\right] &= \mathbb{E}\left[\left(\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right]\right)^2\right] \\ &= \mathrm{Var}\left(\bar{X}_n\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= \frac{1}{n^2} n\mathrm{Var}(X_1) = \frac{1}{n}\mathrm{Var}(X_1) \longrightarrow 0.\end{aligned}$$

This proves the L^2-convergence of $\bar{X}_n$ to μ. □

Example. A fair coin is tossed infinitely many times. Denote by H_n the number of *heads* seen in the first n tosses. Then

$$\mathbb{P}\left(0.49n \leq H_n \leq 0.51n\right) = \mathbb{P}\left(\left|\tfrac{H_n}{n} - 0.5\right| \leq 0.01\right) \longrightarrow 1,$$

applying convergence in probability derived from the weak LLN.

Theorem 4.3.2 (Strong law of large numbers)**.** *Let $X_1, X_2, \dots$ be a sequence of i.i.d. integrable random variables with expectation μ. Then*

$$\bar{X}_n \longrightarrow \mu \ a.s. \qquad as\ n \to \infty.$$

The strong LLN is a much deeper result than the weak one and comes with less assumptions, since the random variable X_i may not have finite variance. We omit the proof.

Example. Let us turn back to the example of tossing a coin. The strong LLN shows that the frequency $\frac{H_n}{n}$ of heads converges to $1/2$, almost surely. Here you may see the meaning of "almost surely": $\Omega = \{H, T\}^{\mathbb{N}}$, and there are $\omega \in \Omega$ such that $\frac{H_n(\omega)}{n}$ does not go to $1/2$, for instance

$$\omega = (H, H, H, H, H, \dots),$$

for which $\dfrac{H_n(\omega)}{n} = 1$. The strong LLN shows that such ω's form a set of measure zero.

4.4 Central limit theorem

Let $X_1, X_2, \dots$ be i.i.d. random variables with mean μ and finite variance σ^2. According to the law of large numbers, the sample mean $\bar{X}_n$ approximates the mean μ, so that $\bar{X}_n - \mu$

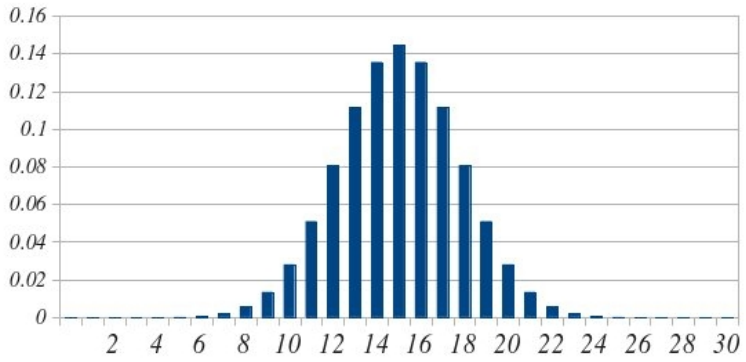


Figure 4.3: Probability mass function of the binomial distribution Bin$(30, 1/2)$.

tends to 0. The question here is the rate of convergence. Is the convergence of $(\bar{X}_n - \mu)_{n\geq 1}$ of order $\sqrt{n}, \sqrt[3]{n}, \log(n), \ldots$?
Let's try to find some intuitive answer by analysing the sequence $(n^\alpha(\bar{X}_n - \mu))_{n\geq 1}$. We see that

$$\begin{aligned}\mathbb{E}\left[\left(n^\alpha(\bar{X}_n - \mu)\right)^2\right] &= n^{2\alpha}\mathbb{E}\left[\left(\bar{X}_n - \mathbb{E}[\bar{X}_n]\right)^2\right] = n^{2\alpha}\mathrm{Var}(\bar{X}_n) \\ &= n^{2\alpha}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = n^{2\alpha-1}\sigma^2 \\ &\overset{n\to+\infty}{\longrightarrow} \begin{cases} 0 & \text{if } 2\alpha - 1 < 0 \\ \sigma^2 & \text{if } 2\alpha - 1 = 0 \\ +\infty & \text{if } 2\alpha - 1 > 0 \end{cases}\end{aligned}$$

Obviously, something interesting happens when $2\alpha - 1 = 0$, i.e. for $\alpha = 1/2$. This is described in more detail by the central limit theorem.

Theorem 4.4.1 (Central limit theorem)**.** *Let $(X_n)_{n\geq 1}$ be i.i.d. random variables with finite variance. Denote $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$. Then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } n \to \infty.$$

When the variance σ^2 is non zero, the CLT says that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \to \infty.$$

This means, that for large sample size n, the sample mean behaves approximately like a normal distribution with mean μ and variance σ^2/n. That is,

$$\bar{X}_n \approx \mu + \frac{\sigma}{\sqrt{n}}Z \quad \text{for large } n,$$

where $Z \sim \mathcal{N}(0, 1)$. According to the CLT, $\bar{X}_n$ has **gaussian fluctuations** around its mean μ. To put it differently, one can also use the CLT to describe the behavior of the sum of i.i.d. random variables. Denote $S_n = \sum_{i=1}^n X_i = n\bar{X}_n$. Then we have

$$S_n \approx n\mu + \sigma\sqrt{n}Z,$$

that is, S_n has gaussian fluctuations around its mean $n\mu$. This is also illustrated in Figure 4.3 that shows the binomial distribution of S_n in the case of i.i.d. random variables X_i with Bernoulli distribution. The probability mass function of S_n is clearly symmetric around the mean $\mu = 15$ and may be well approached by a normal distribution as suggested by the CLT.

Proof. For simplicity, we prove the CLT only in a very particular case. The proof in the general case is similar. We consider discrete i.i.d. random variables X_n with

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

Thus, here we have $\mu = 0$ and $\sigma^2 = 1$. We will prove that the characteristic function of $\sqrt{n}\bar{X}_n$ tends to the characteristic function of the standard normal distribution $\mathcal{N}(0,1)$. More precisely, for any real number t,

$$\lim_{n\to\infty} \Phi_{\sqrt{n}\bar{X}_n}(t) = \exp(-t^2/2).$$

We obtain

$$\begin{aligned}
\Phi_{\sqrt{n}\bar{X}_n}(t) &= \mathbb{E}\left[\mathrm{e}^{it\sqrt{n}\bar{X}_n}\right] = \mathbb{E}\left[\exp\left\{\frac{it}{\sqrt{n}}\sum_{k=1}^{n} X_k\right\}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^{n}\exp\left\{\frac{it}{\sqrt{n}}X_k\right\}\right] \\
&= \left(\mathbb{E}\left[\exp\left\{\frac{it}{\sqrt{n}}X_1\right\}\right]\right)^n \qquad \text{(since } X_i \text{ i.i.d.).}
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\frac{it}{\sqrt{n}}X_1\right\}\right] &= \mathrm{e}^{it/\sqrt{n}}\mathbb{P}(X_1 = +1) + \mathrm{e}^{-it/\sqrt{n}}\mathbb{P}(X_1 = -1) \\
&= \frac{1}{2}(\mathrm{e}^{it/\sqrt{n}} + \mathrm{e}^{-it/\sqrt{n}}).
\end{aligned}$$

Recall that $\mathrm{e}^{it} = \cos(t) + i\sin(t)$, so that $\frac{1}{2}(\mathrm{e}^{it} + \mathrm{e}^{-it}) = \cos(t)$. Thus,

$$\mathbb{E}\left[\mathrm{e}^{it\frac{X_1}{\sqrt{n}}}\right] = \cos\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \mathrm{o}(1/n),$$

since $\cos(u) = 1 - u^2/2 + \mathrm{o}(u^2)$. Finally,

$$\begin{aligned}
\Phi_{\sqrt{n}\bar{X}_n}(t) &= \left(1 - \frac{t^2}{2n} + \mathrm{o}(1/n)\right)^n \\
&= \exp\left(n\log\left(1 - \frac{t^2}{2n} + \mathrm{o}(1/n)\right)\right) \\
&= \exp\left(n\left(-\frac{t^2}{2n} + \mathrm{o}(1/n)\right)\right) \qquad \text{(since } \log(1+u) = u + \mathrm{o}(u)\text{)} \\
&\longrightarrow \exp(-t^2/2),
\end{aligned}$$

when $n \to +\infty$. This completes the proof. □

Many interesting statistics are functions of the sample mean of the form $g(\bar{X}_n)$. The following result is a kind of extension of the CLT to such statistics.

Theorem 4.4.2 (Delta method)**.** *Let $(X_n)_{n\geq 1}$ be i.i.d. random variables with finite variance. Denote $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$. Moreover, let $g(\cdot)$ be a continuously differentiable function, that is, g is differentiable and its derivative is continuous. Then*

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) \xrightarrow{d} \mathcal{N}(0, (g'(\mu))^2\sigma^2), \qquad \text{when } n \to \infty.$$

Applications of the CLT

Asymptotic confidence interval

An unfair coin that turns head with probability p is flipped n times. Let H_n be the number of heads in the first n tosses. Clearly, $H_n \sim \mathrm{Binom}(n, p)$ as we can write

$$H_n = \sum_{i=1}^{n} X_i,$$

where the random variables X_i are i.i.d. with Bernoulli distribution $B(p)$. Recall that $\mathbb{E}[X_k] = p$ and $\mathrm{Var}(X_k) = p(1-p)$. Then, according to the CLT, we have

$$\sqrt{n}\frac{H_n/n - p}{\sqrt{p(1-p)}} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$. Theorem 4.2.4 (ii) yields that, for any reals a, b,

$$\begin{aligned}
&\lim_{n\to+\infty} \mathbb{P}\left(a \leq \sqrt{n}\frac{H_n/n - p}{\sqrt{p(1-p)}} \leq b\right) \\
&\quad = \lim_{n\to+\infty} \mathbb{P}\left(\sqrt{n}\frac{H_n/n - p}{\sqrt{p(1-p)}} \leq b\right) - \lim_{n\to+\infty} \mathbb{P}\left(\sqrt{n}\frac{H_n/n - p}{\sqrt{p(1-p)}} \leq a\right) \\
&\quad = F_Z(b) - F_Z(a) \\
&\quad = \mathbb{P}(a \leq Z \leq b),
\end{aligned}$$

where F_Z is the cumulative distribution function of the standard normal distribution, which is continuous on $\mathbb{R}$. Typically we choose $b = -a$ such that $\mathbb{P}(-a \leq Z \leq a) = 0.95$. This is the case for $a \approx 1.96$. Then we obtain

$$\begin{aligned}
0.95 &= \mathbb{P}(-1.96 \leq Z \leq 1.96) \\
&= \lim_{n\to+\infty} \mathbb{P}\left(-1.96 \leq \sqrt{n}\frac{H_n/n - p}{\sqrt{p(1-p)}} \leq 1.96\right) \\
&= \lim_{n\to+\infty} \mathbb{P}\left(\frac{H_n}{n} \in \left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right).
\end{aligned}$$

That is, when the sample size n is large, the mean number of heads H_n/n lies in a small interval around p. Note that for $0 < p < 1$, one has $p(1-p) \leq 1/4$, so that $1.96\sqrt{p(1-p)} \leq 1.96/2 < 1$. This implies that

$$\left[p - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right],$$

and so

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{H_n}{n} \in \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]\right) \geq 0.95.$$

In other words, with more than 95% chance, the frequency of heads after n flips is at most at a distance $\frac{1}{\sqrt{n}}$ to p.

In statistics, the question is reversed. Here, the parameter p is unknown and the frequency of heads H_n/n is used as an estimator of p.
Obviously, we have

$$\frac{H_n}{n} \in \left[p \pm \frac{1}{\sqrt{n}}\right] \iff \left|\frac{H_n}{n} - p\right| \leq \frac{1}{\sqrt{n}} \iff p \in \left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right].$$

Hence,

$$\lim_{n\to+\infty} \mathbb{P}\left(p \in \left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right]\right) \geq 0.95,$$

and we say that $\mathcal{I} := \left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right]$ is an **asymptotic confidence interval** for p with confidence level 95%. This means that the interval $\mathcal{I}$ contains the unknown parameter p with a chance of 95%, that is, $\mathcal{I}$ gives an approximation of p and comes with an explicit confidence level. That is, we can quantify the uncertainty of our estimator, provided that our assumption on the distribution of the observations are exact and the sample size is large enough.

Limit distribution of a variance estimate of the Bernoulli distribution

In the example of tossing a coin, we consider i.i.d. random variables $X_1, X_2, \ldots$ from the Bernoulli distribution with parameter p. Denote by

$$v = \mathrm{Var}(X_1) = p(1-p)$$

the variance of the Bernoulli distribution $B(p)$. As the sample mean $\bar{X}_n$ is a good estimator of parameter p, it seems to be reasonable to consider

$$\hat{v}_n = \bar{X}_n(1 - \bar{X}_n)$$

as an estimator of v. Indeed, as $\bar{X}_n$ converges to p in probability when n tends to infinity, by the continuous mapping theorem $\hat{v}_n = g(\bar{X}_n)$ with continuous function $g(y) = y(1-y)$ converges to v in probability, that is

$$\hat{v}_n = g(\bar{X}_n) \xrightarrow{P} g(p) = p(1-p) = v, \quad \text{as } n \to \infty.$$

Now, g is continuous differentiable with derivative $g'(y) = 1 - 2y$. According to the delta method, we obtain that

$$\sqrt{n}(\hat{v}_n - v) = \sqrt{n}(g(\bar{X}_n) - g(p)) \xrightarrow{d} \mathcal{N}(0, (g'(p))^2 \mathrm{Var}(X_1)) = \mathcal{N}(0, (1-2p)^2 v).$$

Chapter 5

Conditioning

Conditional probabilities

Recall that for events A, B of $(\Omega, \mathbb{P})$ such that $\mathbb{P}(B) > 0$, the conditional probability of A given B is defined by

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

5.1 Conditional distributions

Definition 5.1.1. *Let (X, Y) be a random vector with joint density $f_{(X,Y)}(x, y)$. Denote by f_Y the marginal density of Y. For $y \in \mathbb{R}$, the* **conditional density** *of X given $Y = y$ is defined as the function*

$$f_{X|Y=y}(x) = \begin{cases} \frac{f_{(X,Y)}(x,y)}{f_Y(y)}, & \text{if } f_Y(y) \neq 0 \\ 0 & \text{if } f_Y(y) = 0 \end{cases}$$

Proposition 5.1.2. *For any fixed y, the conditional density $x \mapsto f_{X|Y=y}(x)$ of X given that $Y = y$ is a probability density function.*

Proof. Clearly, $f_{X|Y=y}(x) \geq 0$ for all x. Moreover,

$$\int_x f_{X|Y=y}(x)\mathrm{d}x = \int_x \frac{f_{(X,Y)}(x,y)}{f_Y(y)}\mathrm{d}x = \frac{1}{f_Y(y)} \int_x f_{(X,Y)}(x,y)\mathrm{d}x = \frac{1}{f_Y(y)} f_Y(y) = 1.$$

□

Remark. An important case is when X and Y are independent. Then

$$f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} \overset{\text{(by indep.)}}{=} \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x),$$

as expected.

Example. We continue the example of page 58, where X is picked uniformly in $[0,1]$, and then, given X, Y is chosen uniformly in $[0,X]$. This means that

$$f_X(x) = \mathbb{1}_{0\leq x\leq 1}$$
$$f_{Y|X=x}(y) = \frac{1}{x}\mathbb{1}_{0\leq y\leq x}.$$

According to Definition 5.1.1 the joint density of (X,Y) is given by

$$f_{(X,Y)}(x,y) = f_X(x)f_{Y|X=x}(y) = \mathbb{1}_{0\leq x\leq 1}\times\frac{1}{x}\mathbb{1}_{0\leq y\leq x} = \frac{1}{x}\mathbb{1}_{0\leq y\leq x\leq 1}.$$

Finally, for the density of Y we find for $y\in[0,1]$,

$$f_Y(y) = \int_{x=0}^{1}\frac{1}{x}\mathbb{1}_{0\leq y\leq x}\mathrm{d}x = \int_{x=y}^{1}\frac{1}{x}\mathrm{d}x = [\log(x)]_{x=y}^{x=1} = -\log(y).$$

Bayes' formula

The relation between the conditional probabilities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$, is given by the Bayes formula for events, that is,

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

There a similar formula for densities, that describes the relation between the conditional densities $f_{X|Y}$ and $f_{Y|X}$.

Proposition 5.1.3 (Bayes' formula for conditional densities)**.** *We have*

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}.$$

Proof. We have

$$\begin{aligned}\frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)} &= \frac{f(x,y)f_X(x)}{f_X(x)f_Y(y)}\\ &= \frac{f(x,y)\cancel{f_X(x)}}{\cancel{f_X(x)}f_Y(y)}\\ &= f_{X|Y=y}(x).\end{aligned}$$

This completes the proof. □

5.2 Conditional expectation

We now consider the problem of finding the best prediction that can be made on X, given the value of another random variable Y. The solution should depend on Y, and therefore be a random variable.

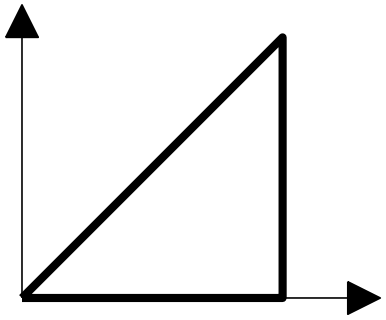


Figure 5.1: Unit triangle.

Definition 5.2.1 (Conditional expectation)**.**

- **(Discrete case)** *Let (X, Y) be a pair of discrete random variables. Denote*

$$p(x|y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}, \qquad \text{if} \quad \mathbb{P}(Y = y) > 0,$$

and $p(x|y) = 0$ if $\mathbb{P}(Y = y) = 0$. The **conditional expectation** *of X given Y is defined by*

$$\mathbb{E}[X|Y] = \sum_x xp(x|Y).$$

- **(Continuous case)** *Let (X, Y) be a random vector with* **continuous** *distribution. The* **conditional expectation** *of X given Y is defined by*

$$\mathbb{E}[X|Y] = \int_x x f_{X|Y}(x|Y)\mathrm{d}x.$$

Note that in both cases, by construction $\mathbb{E}[X|Y]$ is a function of Y, and therefore is a random variable.
More generally, for any integrable function ϕ,

$$\mathbb{E}[\phi(X, Y)|Y] = \begin{cases} \sum_x \phi(x)p(x|Y), & \text{in the discrete case} \\ \int_x \phi(x, Y) f_{X|Y}(x|Y)\mathrm{d}x, & \text{in the continuous case} \end{cases}$$

In the continuous case we also introduce the following useful notation. We denote

$$\mathbb{E}[X|Y = y] = \int x f_{X|Y}(x|y)\mathrm{d}x.$$

Here, $\mathbb{E}[X|Y = y]$ is a simple function of y, while $\mathbb{E}[X|Y]$ is a random variable.

Example. Let (X, Y) have uniform distribution on the unit triangle of area $1/2$, see Figure 5.1. More precisely, the joint density of (X, Y) is $f_{(X,Y)}(x, y) = 2\mathbb{1}_{0\le y\le x\le 1}$. Then

$$f_Y(y) = \int_x 2\mathbb{1}_{0\le y\le x\le 1}\mathrm{d}x = \int_{x=y}^1 2\mathrm{d}x = 2(1 - y).$$

Let us apply the definition to compute $\mathbb{E}[X|Y]$. First,

$$\int_x xf(x,Y)\mathrm{d}x = \int_x 2x\mathbb{1}_{0\leq Y\leq x\leq 1}\mathrm{d}x = 2\int_{x=Y}^{1} x\mathrm{d}x = 2\left[x^2/2\right]_{x=Y}^{x=1} = 1-Y^2.$$

Then the formula reads

$$\mathbb{E}[X|Y] = \frac{\int_x xf_{(X,Y)}(x,Y)\mathrm{d}x}{f_Y(Y)} = \frac{1-Y^2}{2(1-Y)} = \frac{1+Y}{2}.$$

We see that $\mathbb{E}[X|Y]$ is a function of Y. Besides, $\mathbb{E}[X|Y=y] = (1+y)/2$.

Proposition 5.2.2 (Properties of conditional expectations)**.** *Let X, X', Y be random variables. In both the discrete and the continuous case, we have*

(i) **(Linearity)** $\mathbb{E}[aX+X'|Y] = a\mathbb{E}[X|Y] + \mathbb{E}[X'|Y]$ *for any constant a.*

(ii) **(Averaging)**

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X|Y]\right] = \begin{cases} \sum_y \mathbb{E}[X|Y=y]p_Y(y), & \text{in the discrete case} \\ \int_y \mathbb{E}[X|Y=y]f_Y(y)\mathrm{d}y, & \text{in the continuous case} \end{cases}$$

(iii) **('Taking out what is known')** *For any function g, we have $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$. In particular, $\mathbb{E}[g(Y)|Y] = g(Y)$.*

(iv) **(Independence)** *If X and Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.*

Proof. We skip the proof of (i), which is intuitive but not obvious in the setting of this course. Let us briefly treat the other properties in the continuous case. Property (ii) is obtained as follows. By the definition of $\mathbb{E}[X|Y]$,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[X|Y]\right] &= \mathbb{E}\left[\frac{\int_x xf(x,Y)\mathrm{d}x}{f_Y(Y)}\right] \\
&= \int_y \left(\frac{\int_x xf(x,y)\mathrm{d}x}{f_Y(y)}\right) f_Y(y)\mathrm{d}y \qquad \text{(note that this is equal to } \int_y \mathbb{E}[X|Y=y]f_Y(y)\mathrm{d}y.\text{)} \\
&= \int_y \left(\frac{\int_x xf(x,y)\mathrm{d}x}{\cancel{f_Y(y)}}\right) \cancel{f_Y(y)}\mathrm{d}y \\
&= \int_y\int_x xf(x,y)\mathrm{d}x\mathrm{d}y = \mathbb{E}[X].
\end{aligned}$$

To show (iii), we see that

$$\begin{aligned}
\mathbb{E}[Xg(Y)|Y] &= \frac{\int_x xg(Y)f(x,Y)\mathrm{d}x}{f_Y(Y)} \\
&= g(Y)\frac{\int_x xf(x,Y)\mathrm{d}x}{f_Y(Y)} \\
&= g(Y)\mathbb{E}[X|Y].
\end{aligned}$$

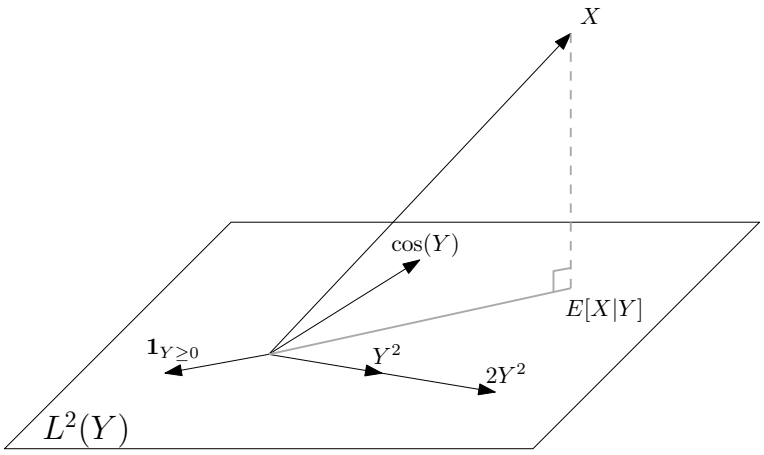


Figure 5.2: Illustration of the conditional expectation as the orthogonal projection of X onto $L^2(Y)$.

To show (iv) it is intuitive that Y doesn't bring any information on X. Indeed,

$$\begin{aligned}
\mathbb{E}[X|Y] &= \frac{\int_x x f(x,Y)\mathrm{d}x}{f_Y(Y)} \\
&= \frac{\int_x x f_X(x) f_Y(Y)\mathrm{d}x}{f_Y(Y)} \qquad \text{(by independence of } X, Y\text{)} \\
&= \frac{\int_x x f_X(x) \cancel{f_Y(Y)}\mathrm{d}x}{\cancel{f_Y(Y)}} \\
&= \mathbb{E}[X].
\end{aligned}$$

□

We give an example that shows how we can use $\mathbb{E}[Y|X]$ in order to evaluate $\mathbb{E}[Y]$.

Example. Again we consider the case, where we first sample X uniformly in $[0,1]$, and then pick Y uniformly in $[0,X]$. We want to compute $\mathbb{E}[Y]$. According to Proposition 5.2.2 (ii), we obtain by averaging that

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[\mathbb{E}[Y|X]\right] \\
&= \mathbb{E}\left[X/2\right] \qquad (Y \text{ is uniform in } [0,X]) \\
&= \int_{x=0}^{1} \tfrac{1}{2} x \mathrm{d}x = \left[x^2/2\right]_{x=0}^{x=1} = 1/4.
\end{aligned}$$

Conditional expectation as the best predictor

Let $X, Y \in L^2$. The conditional expectation $\mathbb{E}[X|Y]$ coincides with the orthogonal projection of X onto the vector space

$$L^2(Y) = \left\{\text{random variables of the form } g(Y) \text{ such that } \mathbb{E}[g(Y)^2] < +\infty\right\},$$

equipped with the inner product $\langle X_1, X_2 \rangle = \mathbb{E}[X_1 X_2 | Y]$, see Figure 5.2.
As we have seen in Section 2.5, orthogonal projections can be considered as the solution of a minimization problem. This gives the following interpretation of the conditional expectation $\mathbb{E}[X|Y]$.

Theorem 5.2.3 (Conditional expectation as the best predictor)**.** *Let $X, Y \in L^2$. The conditional expectation $\mathbb{E}[X|Y]$ is the best predictor of X among all possible predictors which are functions of Y, i.e.*

$$\mathbb{E}\Big[\big(X - \mathbb{E}[X|Y]\big)^2 \Big] \leq \mathbb{E}\Big[\big(X - g(Y)\big)^2 \Big]$$

for every function g such that $g(Y) \in L^2$.

Appendix A

Gaussian vectors

In this chapter we introduce gaussian vectors, which play a central role in statistics.

A.1 Gaussian random variables

To start with, we recall the basic properties of gaussian random variables and the normal distribution on $\mathbb{R}$.

Proposition A.1.1 (Properties of the univariate normal distribution)**.** *Let X be a random variable with gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, i.e. X has density*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \text{for } x \in \mathbb{R}.$$

1. $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.
2. *For any constants $a, b \in \mathbb{R}$, $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.*
3. *The characteristic function of X is given by*

$$\Phi_X(t) = \mathbb{E}[\mathrm{e}^{itX}] = \exp(it\mu - t^2\sigma^2/2). \tag{A.1}$$

4. *If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \ldots, X_d \sim \mathcal{N}(\mu_d, \sigma_d^2)$ are independent, then*

$$\sum_{k=1}^{d} X_k \sim \mathcal{N}\left(\sum_{k=1}^{d} \mu_k, \sum_{k=1}^{d} \sigma_k^2\right).$$

Proof. Property (i) is obtained by direct computation of the moments of the gaussian distribution. Property (ii) can be shown by computing the density of $aX + b$ using a change of variables. The last item can be proved computing the characteristic function of the sum $\sum_{k=1}^{d} X_k$.

For (iii) we first show that the statement holds for the standard normal distribution, i.e. $X \sim \mathcal{N}(0,1)$. Hence, we have to show that $\Phi_X(t) = \exp(-t^2/2)$. By differentiating the characteristic function Φ_X we obtain

$$\begin{aligned}
\Phi'_X(t) = \frac{\partial}{\partial t}\Phi_X(t) &= \mathbb{E}\left[\frac{\partial}{\partial t}\mathrm{e}^{itX}\right] \qquad \text{(using Theorem 2.6.4)} \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial t}\mathrm{e}^{itx}\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\mathrm{d}x \\
&= \int_{\mathbb{R}} ix\mathrm{e}^{itx}\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\mathrm{d}x \\
&= \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}} \underbrace{\mathrm{e}^{itx}}_{v}\underbrace{x\exp(-x^2/2)}_{u'}\mathrm{d}x = \frac{i}{\sqrt{2\pi}}\left(\cancel{[uv]_{-\infty}^{+\infty}} - \int_{-\infty}^{+\infty} uv'\right) \\
&= \frac{i}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \exp(-x^2/2)it\mathrm{e}^{itx} = -t\Phi_X(t).
\end{aligned}$$

Thus we have to solve the differential equation $\frac{\Phi'_X(t)}{\Phi_X(t)} = -t$, which is equivalent to

$$(\log(\Phi_X(t)))' = -t.$$

Hence $\log(\Phi_X(t)) = -t^2/2 + c$ for some constant c, i.e. $\Phi_X(t) = \exp(-t^2/2)\mathrm{e}^c$. Recall that $\Phi_X(0) = 1$, so $\mathrm{e}^c = 1$.
Now we turn to the general case where $X \sim \mathcal{N}(\mu, \sigma^2)$. We can write $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$. So,

$$\mathbb{E}\left[\mathrm{e}^{itX}\right] = \mathbb{E}\left[\mathrm{e}^{it\mu+it\sigma Z}\right] = \mathrm{e}^{it\mu}\mathbb{E}\left[\mathrm{e}^{it\sigma Z}\right] = \mathrm{e}^{it\mu}\mathbb{E}\left[\mathrm{e}^{i(t\sigma)Z}\right] = \mathrm{e}^{it\mu}\Phi_Z(t\sigma) = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right).$$

This completes the proof. □

A.2 Gaussian vectors

We now introduce the multivariate gaussian distribution.

Definition A.2.1 (Gaussian vectors)**.** *A random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ *is a* **gaussian vector***, and we say that* $\mathbf{X}$ *follows the* **multivariate normal distribution***, if for any* $\mathbf{t} = (t_1, t_2, \ldots, t_d)^T \in \mathbb{R}^d$ *the linear combination*

$$\mathbf{t}^T\mathbf{X} = \sum_{k=1}^{d} t_k X_k$$

is a gaussian random variable in $\mathbb{R}$*.*

By convention, we include the constant random variables of the form $X = \mu$ *a.s.* in the family of normal distributions. In this case we write $X \sim \mathcal{N}(\mu, 0)$.

Example. Gaussian vectors.

- Let $X_1, \dots, X_d$ be independent normally distributed random variables with $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ for $k = 1, \dots, d$. Then, the random variable

$$\mathbf{t}^T\mathbf{X} = \sum_{k=1}^{d} t_k X_k \sim \mathcal{N}\left(\sum_{k=1}^{d} t_k\mu_k, \sum_{k=1}^{d} t_k^2\sigma_k^2\right)$$

  has normal distribution for any $\mathbf{t} = (t_1, \dots, t_d) \in \mathbb{R}^d$. Hence, $\mathbf{X}$ is a gaussian vector. In other words, any vector made of **independent** gaussian random variables is a gaussian vector.

- Let X have standard normal distribution $\mathcal{N}(0, 1)$. Then the vector $(X, -X)$ is a gaussian vector, since

$$t_1X + t_2(-X) = (t_1 - t_2)X,$$

  is a gaussian random variable for all $t_1, t_2 \in \mathbb{R}$.

- Let $a > 0$ be a constant and X have standard normal distribution $\mathcal{N}(0, 1)$. Set

$$Y = \begin{cases} X, & \text{if } |X| > a \\ -X, & \text{if } |X| \le a \end{cases}$$

  One can show that $Y \sim \mathcal{N}(0, 1)$. However, (X, Y) is **not** a gaussian vector. Indeed $X+Y$ is not normally distributed, since $\mathbb{P}(X+Y = 0) = \mathbb{P}(X \in (-a, a)) = 2F_X(-a) > 0$, while $X + Y$ is not a constant either. So this is an example of gaussian random variables X and Y, where the random vector (X, Y) is not gaussian.

If $\mathbf{X} = (X_1, \dots, X_d)$ is a gaussian vector, then every sub-vector of $\mathbf{X}$ is a gaussian vector as well: for instance, (X_2, X_5, X_{11}) is a gaussian vector. To show this using the definition, it is sufficient to take $t_2 = t_5 = t_{11} = 1$ and every other t_i equal to zero. In particular, by taking $t_i = 1$ and $t_j = 0$ for $j \neq i$, we obtain that all components X_k are normally distributed. That is, every component of a gaussian vector is gaussian.

Theorem A.2.2. *Let $\mathbf{X}$ be a gaussian vector in $\mathbb{R}^d$ with mean vector μ and covariance matrix C. Then the multivariate characteristic function of $\mathbf{X}$ is given by*

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T C\mathbf{t}\right), \qquad \textit{for } \mathbf{t} \in \mathbb{R}^d.$$

Proof. We fix a vector $\mathbf{t} = (t_1, \dots, t_d)^T \in \mathbb{R}^d$. As $\mathbf{X}$ is a gaussian vector, the random variable Y defined by the linear combination $Y = \mathbf{t}^T\mathbf{X} = t_1X_1 + \cdots + t_dX_d$ is normally distributed. Hence,

$$\begin{aligned}\Phi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}\left[\exp\left(i\mathbf{t}^T\mathbf{X}\right)\right] = \mathbb{E}\left[\exp(i \times 1 \times Y)\right] \\ &= \Phi_Y(1) = \exp\left\{i\mathbb{E}[Y] - \frac{\mathrm{Var}(Y)}{2}\right\},\end{aligned}$$

by Proposition A.1.1. We obtain, by linearity of the expectation,

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbf{t}^T\mathbf{X}\right] = \mathbf{t}^T\mathbb{E}[\mathbf{X}] = \mathbf{t}^T\mu,$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbf{t}^T\mathbf{X}) = \mathbf{t}^T\mathrm{Var}(\mathbf{X})\mathbf{t} = \mathbf{t}^T C\mathbf{t},$$

which concludes the proof. □

A consequence of Theorem A.2.2 is that if $\mathbf{X}$ and $\mathbf{Y}$ are two gaussian vectors with the same mean vector and the same covariance matrix, then $\Phi_{\mathbf{X}}(\mathbf{t}) = \Phi_{\mathbf{Y}}(\mathbf{t})$ for all $\mathbf{t}$, implying that $\mathbf{X}$ and $\mathbf{Y}$ have the same law. This gives us the following proposition.

Proposition A.2.3. *The distribution of a gaussian vector $\mathbf{X} = (X_1, \ldots, X_d)$ is fully characterized by its mean vector μ and its covariance matrix C, and we denote*

$$\mathbf{X} \sim \mathcal{N}_d(\mu, C).$$

Another consequence of the expression of the characteristic function is that for gaussian vectors independence of two components is equivalent to a covariance equal to zero.

Proposition A.2.4. *Let $\mathbf{X}$ be a gaussian vector. If for i and j, $\mathrm{Cov}(X_i, X_j) = 0$, then X_i and X_j are independent.*

Example. Let X and Y be independent with standard normal distribution $\mathcal{N}(0,1)$. We use the proposition to show that $X+Y$ and $X-Y$ are independent. First, we see that for all t_1, t_2 we have $t_1(X+Y) + t_2(X-Y) = (t_1+t_2)X + (t_1-t_2)Y$ has normal distribution. Hence, $(X+Y, X-Y)$ is a gaussian vector. Hence, to study independence it is enough to study the covariance. Using bilinearity repeatedly we get

$$\begin{aligned}
\mathrm{Cov}(X+Y, X-Y) &= \mathrm{Cov}(X, X-Y) + \mathrm{Cov}(Y, X-Y) \\
&= \mathrm{Cov}(X,X) - \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) - \mathrm{Cov}(Y,Y) \\
&= \mathrm{Cov}(X,X) - \mathrm{Cov}(Y,Y) \\
&= \mathrm{Var}(X) - \mathrm{Var}(Y) \\
&= 0.
\end{aligned}$$

Proposition A.2.4 also holds for subvectors of $\mathbf{X}$. That means that zero entries or a block structure of a covariance matrix indicates the corresponding parts of the gaussian vector are independent.

A useful property is that any linear transformation of gaussian vectors is also gaussian.

Proposition A.2.5 (Linear transformation of a gaussian vector)**.** *Let $\mathbf{X}$ be a gaussian vector with $\mathbf{X} \sim \mathcal{N}_d(\mu, C)$. Then, for any $p \times d$-matrix M and any vector $a \in \mathbb{R}^p$, the random vector $M\mathbf{X} + a \in \mathbb{R}^p$ is a gaussian vector. More precisely,*

$$M\mathbf{X} + a \sim \mathcal{N}_d(M\mu + a, MCM^T).$$

Proof. For any $\mathbf{t} \in \mathbb{R}^d$, we have $\mathbf{t}^T(M\mathbf{X}+a) = (\mathbf{t}^T M)\mathbf{X}+\mathbf{t}^T a$, that is, the linear combination of the components of $M\mathbf{X}+a$ can also be considered as a linear combination of the components of $\mathbf{X}$ (plus some constant). As $\mathbf{X}$ is a gaussian vector, $\mathbf{t}^T(M\mathbf{X}+a)$ is normally distributed. It follows that $M\mathbf{X}+a$ is a gaussian vector. Now, it is sufficient to compute the mean and the covariance matrix of $M\mathbf{X}+a$. This is done by Theorem 3.5.5. □

We conclude this section with the formula of the density of a gaussian vector (if it exists).

Proposition A.2.6 (Density of the multivariate normal distribution)**.** *Let $\mathbf{X}$ be a gaussian vector with distribution $\mathcal{N}_d(\mu, C)$. If C is invertible, i.e. there exists C^{-1} such that $C^{-1}\times C = I_d$, where I_d denotes the identity matrix of size $d \times d$, then $\mathbf{X}$ has a density $f_{\mathbf{X}}$ that is given by*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)\right), \tag{A.2}$$

for all $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$.

Proof. First show that the result holds for the standard multivariate normal distribution, i.e. where $\mu = (0, \ldots, 0)^T \in \mathbb{R}^d$ and $C = I_d$ is the identity matrix. In this case, the components X_k are i.i.d. with standard normal distribution $\mathcal{N}(0,1)$. Hence, by Theorem 3.3.4, for all $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, the joint density of $\mathbf{X}$ can be written as

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1}(x_1) \cdots f_{X_d}(x_d) \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_1^2/2} \cdots \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_d^2/2} \\ &= \frac{1}{\sqrt{(2\pi)^d}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right\} \\ &= \frac{1}{\sqrt{(2\pi)^d}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right\}. \end{aligned}$$

The last expression coincides with the density given in (A.2) using $\mu = (0, \ldots, 0)^T \in \mathbb{R}^d$ and $C = I_d$, since $\det(I_d) = 1$. Now, the general formula can be obtained by a multivariate change of variables. □

The fact that C is invertible implies that $\det(C) \neq 0$.

When $d = 1$, we recover the density of the univariate gaussian distribution, since the matrix C is just $\big(\mathrm{Var}(X)\big)$, so that $\det(C) = \mathrm{Var}(X)$.

One can show that when C is not invertible, then the gaussian vector $\mathbf{X}$ has no density. Indeed, in this case, $\mathbf{X}$ takes its values in a strict subspace of $\mathbb{R}^d$ (whose Lebesgue measure is zero) and so $\mathbf{X}$ does not admit a density.

Figure A.1 illustrates the joint density of a gaussian vector with negatively correlated components. More precisely, the figure shows the density of the normal distribution

$$\mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}4 & -3\\ -3 & 4\end{pmatrix}\right).$$

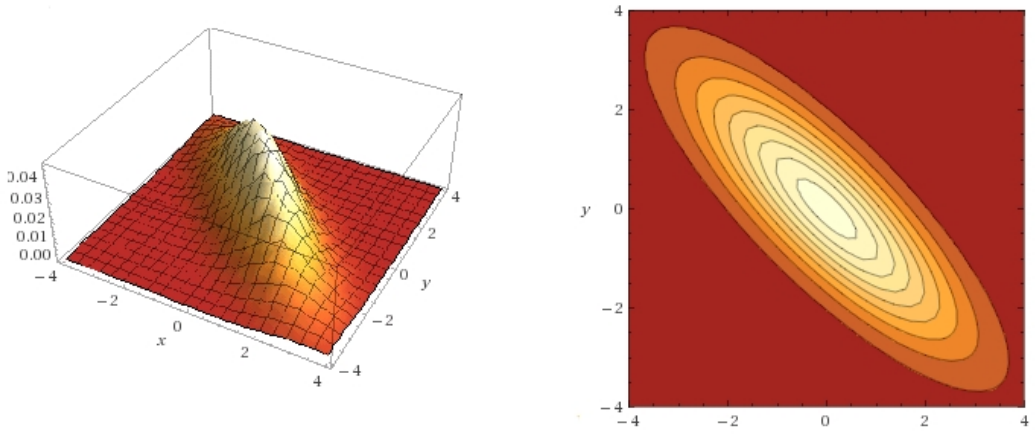


Figure A.1: Density of a gaussian vector in $\mathbb{R}^2$ with negatively correlated components.

Indeed, the negative covariance $\mathrm{Cov}(X_1, X_2) = -3$ implies that X_1 and X_2 tend to have opposite signs, which is also observed in the figure. You may compare this distribution to the one displayed in Figure 3.1, where the density of a gaussian vector with independent components is illustrated.

Conditional expectation and gaussian vector

Recall that for a gaussian vector (X, Y), X and Y are independent if and only if $\mathrm{Cov}(X, Y) = 0$. This is a useful property for the computation of conditional expectations.

Example. Let (X, Y) be a gaussian vector with normal distribution $\mathcal{N}_2\left(\begin{pmatrix}3\\2\end{pmatrix}, \begin{pmatrix}3 & 1\\1 & 2\end{pmatrix}\right)$.
We want to compute $\mathbb{E}[X|Y]$.
1st step. Let us find a constant a such that $X - aY$ and Y are independent. Since $\binom{X-aY}{Y}$ is also a gaussian vector (recall Proposition A.2.5), it is sufficient to consider the covariance given by

$$\mathrm{Cov}(Y, X - aY) = \mathrm{Cov}(Y, X) - a\mathrm{Cov}(Y, Y) = 1 - 2a.$$

Therefore, taking $a = 1/2$, we deduce that the random variable $X - Y/2$ is independent from Y.
2nd step. We write

$$\begin{array}{rcccc} \mathbb{E}[X|Y] = \mathbb{E}[X - Y/2 + Y/2|Y] & = & \mathbb{E}[Y/2|Y] & + & \mathbb{E}[X - Y/2|Y] \\ & = & Y/2 & + & \mathbb{E}[X - Y/2] \\ & & \text{('taking out what is known')} & & \text{(independence)} \end{array}$$

Finally, $\mathbb{E}[X - Y/2] = \mathbb{E}[X] - \mathbb{E}[Y]/2 = 3 - 2/2 = 2$ We get $\mathbb{E}[X|Y] = Y/2 + 2$.

A.3 How to simulate a gaussian vector?

Suppose that we know how to simulate from the standard normal distribution $\mathcal{N}(0,1)$. How can we simulate a gaussian vector (X,Y) with arbitrary mean μ and covariance matrix C? Consider the example where μ and C are given by

$$\mu = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad C = \begin{pmatrix} 5 & -1 \\ -1 & 10 \end{pmatrix}.$$

Let Z_1 and Z_2 be to independent random variables with standard normal distribution $\mathcal{N}(0,1)$ (this is what we know to simulate). Then (Z_1, Z_2) is a gaussian vector. According to Proposition A.2.5, if is sufficient to find a 2×2 matrix M such that $MM^T = C$. Then $M\binom{Z_1}{Z_2} + \mu$ has the desired normal distribution $\mathcal{N}_2(\mu, C)$. In our example we check that $M = \begin{pmatrix} -1 & 2 \\ 3 & 1 \end{pmatrix}$ is a solution of our problem. Indeed,

$$\begin{pmatrix} -1 & 2 \\ 3 & 1 \end{pmatrix} \times \begin{pmatrix} -1 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -1 \\ -1 & 10 \end{pmatrix}.$$

Appendix B

Multivariate version of change of variables formula

B.1 Bivariate change of variables

Let (X, Y) be a random vector with density $f(x, y)$. Here the aim is to determine the joint density of $(U, V) = (u(X, Y), v(X, Y))$, where u and v are known functions. This can be achieved by the following theorem, which is the analogous result of Theorem 2.4.1 for random vectors.

Theorem B.1.1 (Characterization with bounded and continuous ϕ)**.** *Let $f : \mathbb{R}^2 \to \mathbb{R}_+$ be a function such that, for every bounded and continuous function $\phi : \mathbb{R}^2 \to \mathbb{R}$,*

$$\mathbb{E}\left[\phi(X, Y)\right] = \iint \phi(x, y) f(x, y) \mathrm{d}x\mathrm{d}y.$$

then the random vector (X, Y) has joint density f.

Hence, to determine the density of $(U, V) = (u(X, Y), v(X, Y))$, we consider

$$\mathbb{E}[(\phi(U, V)] = \mathbb{E}\left[\phi\left(u(X, Y), v(X, Y)\right)\right] = \iint_{\mathbb{R}^2} \phi\left(u(x, y), v(x, y)\right) f(x, y) \mathrm{d}x\mathrm{d}y,$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$ is any bounded and continuous function. The general method consists in a bivariate change of variables. But instead of going into the details of the theory, let us rather work out two examples.

A simple example

Let (X, Y) have joint density

$$f(x, y) = \tfrac{3}{4} \exp\left(-|x + 2y| - |x - y|\right).$$

What is the joint density of $(X + 2Y, X - Y)$? We set $U = X + 2Y$ and $V = X - Y$ and consider

$$\mathbb{E}\left[\phi(U, V)\right] = \iint_{\mathbb{R}^2} \phi\left(x + 2y, x - y\right) \tfrac{3}{4} \exp\left(-|x + 2y| - |x - y|\right) \mathrm{d}x\mathrm{d}y.$$

In the latter integral we make the change of variables

$$\begin{cases} u = x + 2y \\ v = x - y \end{cases} \iff \begin{cases} x = (u + 2v)/3 \\ y = (u - v)/3 \end{cases}$$

In the integral (x, y) runs over all $\mathbb{R}^2$, so does (u, v). Hence, the integration domain is still $\mathbb{R}^2$ after the change of variable. For the change $\mathrm{d}x\mathrm{d}y \leftrightarrow \mathrm{d}u\mathrm{d}v$, we have to compute the so-called **Jacobian matrix**

$$\mathrm{Jac}(x, y) = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial u}\left(\frac{u+2v}{3}\right) & \frac{\partial}{\partial v}\left(\frac{u+2v}{3}\right) \\ \frac{\partial}{\partial u}\left(\frac{u-v}{3}\right) & \frac{\partial}{\partial v}\left(\frac{u-v}{3}\right) \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & -1/3 \end{pmatrix}.$$

Now the general formula is (don't forget absolute values)

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v} = \left|\det(\mathrm{Jac}(x, y))\right|.$$

In the example, we thus obtain

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v} = \left|\begin{pmatrix} 1/3 & 2/3 \\ 1/3 & -1/3 \end{pmatrix}\right| = \left|\frac{1}{3} \times \left(-\frac{1}{3}\right) - \frac{1}{3} \times \frac{2}{3}\right| = |-3/9| = 1/3.$$

Finally, we get

$$\mathbb{E}\left[\phi(U, V)\right] = \iint_{\mathbb{R}^2} \phi(u, v)\, \tfrac{3}{4}\mathrm{e}^{-|u|-|v|} \frac{\mathrm{d}u\mathrm{d}v}{3} = \iint_{\mathbb{R}^2} \phi(u, v)\, \tfrac{1}{4}\mathrm{e}^{-|u|-|v|} \mathrm{d}u\mathrm{d}v.$$

Thus, according to Theorem B.1.1, the random vector $(U, V) = (X + 2Y, X - Y)$ has density $(u, v) \mapsto \frac{1}{4}\mathrm{e}^{-|u|-|v|}$.

Remark.

1. It is not necessary to check explicitly that $\frac{1}{4}\mathrm{e}^{-|u|-|v|}$ is a density on $\mathbb{R}^2$. This is always the case (if the change of variables is done correctly).
2. One can also compute $\frac{\mathrm{d}u\mathrm{d}v}{\mathrm{d}x\mathrm{d}y}$ in the reverse way. Just interchange $(u, v) \leftrightarrow (x, y)$, then we obtain

   $$\frac{\mathrm{d}u\mathrm{d}v}{\mathrm{d}x\mathrm{d}y} = \left|\det(\mathrm{Jac}(u, v))\right| = \left|\det\begin{pmatrix} \frac{\partial}{\partial x}(x + 2y) & \frac{\partial}{\partial y}(x + 2y) \\ \frac{\partial}{\partial x}(x - y) & \frac{\partial}{\partial y}(x - y) \end{pmatrix}\right| = \left|\det\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}.\right| = |-3| = 3,$$

   which is consistent with the previous computation.

Another example: polar coordinates

Let the random vector (X, Y) have uniform distribution on the quarter disc

$$D = \left\{(x, y) \in \mathbb{R}_+^2, \sqrt{x^2 + y^2} \leq 1\right\}.$$

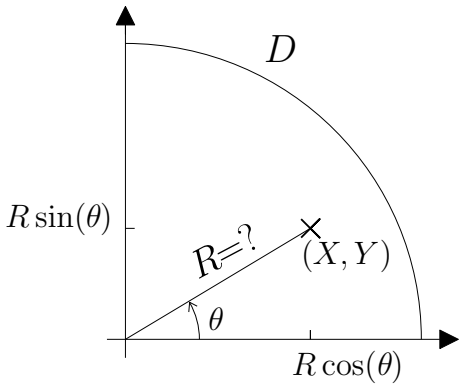


Figure B.1: Polar coordinates and the quarter disc.

In other words, (X, Y) has the density

$$f(x,y) = \frac{1}{\pi/4}\mathbb{1}_D(x,y).$$

Note that f is a density, since the area of D is $\pi/4$. Now let $R = \sqrt{X^2+Y^2}$ be the distance from (X, Y) to the origin, see Figure B.1. We are interested in the distribution of R. To this end, we compute

$$\mathbb{E}\left[\phi\left(R\right)\right] = \iint_{\mathbb{R}^2} \phi\left(\sqrt{x^2+y^2}\right)\frac{1}{\pi/4}\mathbb{1}_D(x,y)\mathrm{d}x\mathrm{d}y,$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a bounded and continuous function. To obtain $r = \sqrt{x^2+y^2}$, we make the **polar** change of variables, that is,

$$\begin{cases} r\cos(\theta) = x, \\ r\sin(\theta) = y. \end{cases}$$

In this way we get

$$\sqrt{x^2+y^2} = \sqrt{r^2\cos^2(\theta) + r^2\sin^2(\theta)} = \sqrt{r^2\times 1} = r.$$

From Figure B.1 we have

$$(x,y) \in D \Leftrightarrow \begin{cases} r \le 1, \\ 0 \le \theta \le \pi/2. \end{cases}$$

To make the change $\mathrm{d}x\mathrm{d}y \leftrightarrow \mathrm{d}r\mathrm{d}\theta$, we determine the Jacobian matrix. We find that

$$\mathrm{Jac}(x,y) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial r}r\cos(\theta) & \frac{\partial}{\partial \theta}r\cos(\theta) \\ \frac{\partial}{\partial r}r\sin(\theta) & \frac{\partial}{\partial \theta}r\sin(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}.$$

Hence, we get

$$\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}r\mathrm{d}\theta} = \left|\det(\mathrm{Jac}(x,y))\right| = \left|r\cos^2(\theta) + r\sin^2(\theta)\right| = r.$$

Finally,

$$\begin{aligned}\mathbb{E}\left[\phi(R)\right] &= \int_{r=0}^{1}\left(\int_{\theta=0}^{\pi/2}\phi(r)\frac{r}{\pi/4}\mathrm{d}\theta\right)\mathrm{d}r \\ &= \int_{r=0}^{1}\phi(r)r\left(\int_{\theta=0}^{\pi/2}\frac{1}{\pi/4}\mathrm{d}\theta\right)\mathrm{d}r \\ &= \int_{r=0}^{1}\phi(r)2r\mathrm{d}r.\end{aligned}$$

It follows by Theorem 2.4.1 that R has density $f_R(r) = 2r\mathbb{1}_{[0,1]}$.