Probability and Statistics

V. Tassion

D-INFK Spring 2022 (Updated: May 2, 2022)

Introduction to probability

Some questions you may ask

What is probability?

➜ A mathematical language describing systems involving randomness.

Where are probabilities used?

➜ **Describe random experiments** in the real world (coin flip, dice rolling, arrival times of customers in a shop, weather in 1 year,...).

➜ **Express uncertainty**. For example, when a machine performs a measurement, the value is rarely exact. One may use probability theory in this context by saying that the value obtained is equal to the real value plus a small random error.

➜ **Decision making**. Probability theory can be used to describe a system when only part of the information is known. In such context, it may help to make a decision.

➜ **Randomized algorithms** in computer science. Sometimes, it is more efficient to add some randomness to perform an algorithm. Examples: Google web search, ants searching food.

➜ **Simplify complex systems**. Examples: water molecules in water, cars on the highway, percolation processes.

The notion of probabilistic model

If one wants a precise physical description of a coin flip one would need a lot (really!) of information: the exact position of the fingers and the coins, the initial velocity, the initial angular velocity, imperfections of the coin, the surface characteristics of the table, air currents, the brain activity of the gambler... These parameters are almost impossible to measure precisely, and a tiny change in one of them may affect completely the result. In practice, we rather use a probabilistic description, which here consists in a drastic simplification of the system: we completely forget the physical description of the throw of the coin and we only focus on the possible **outcomes** of the experiment: head or tail.

Namely, the probabilistic model for the coin flip is given by 2 possible outcomes (head and tail) and each outcome has probability $p_{\text{head}} = p_{tail} = 1/2$ to be realized. In other words,

$$\textbf{Coin flip} = \{\{\text{head}, \text{tail}\}, p_{\text{head}} = 1/2, p_{\text{tail}} = 1/2\}$$

A surprising analysis: coin flips are not fair! If one tosses a coin it has more chance to fall on the same face as its initial face! See the youtube video of Persi Diaconis: How random is a coin toss? - Numberphile

Probability laws: randomness vs ordering

If one performs a single random experiment (for example a coin flip), the result is unpredictable. In contrast, when one performs many random experiments, then some general laws can be observed. For example if one tosses 10000 independent coins, one should generally observe 5000 heads and 5000 tails approximately: This is an instance of a fundamental probability law, called the law of large numbers. One goal of probability theory is to describe how ordering can emerge out of many random experiments, and establish some general probability laws.

Chapter 1

Mathematical framework

Goals

- Basic understanding of the notion of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$

  ➜ Generalization of discrete probability spaces (introduced in[LSW21])

  ➜ Notion of sigma-algebra.

- Concept of independence, conditional probability.

1 Probability space

Sample space

We want to model a random experiment. The first mathematical object needed is the set of all possible outcomes of the experiment, denoted by Ω.

Terminology: The set Ω is called the **sample space**. An element $\omega \in \Omega$ is called an **outcome** (or **elementary experiment**).

Example : *Throw of a die*

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

Events

Reminder: The set $\mathcal{P}(\Omega)$ denotes the collection of all subsets $A \subset \Omega$.

In the previous class [LSW21], the set of events was always $\mathcal{P}(\Omega)$. In this class we will work with more general sets of events $\mathcal{F} \subset \mathcal{P}(\Omega)$, called sigma-algebras.

Definition 1.1. *A **sigma-algebra** is a subset $\mathcal{F} \subset \mathcal{P}(\Omega)$ satisfying the following properties.*

P1. $\Omega \in \mathcal{F}$

P2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ *if A is an event, "non A" is also an event.*

P3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$ *if $A_1, A_2, \ldots$ are events, then "A_1 or A_2 or ..." is an event*

Examples of sigma-algebras for $\Omega = \{1, 2, 3, 4, 5, 6\}$**:**

- $\mathcal{F} = \{\varnothing, \{1, 2, 3, 4, 5, 6\}\}.$

- $\mathcal{F} = \mathcal{P}(\Omega).$ (In this case $|\mathcal{F}| = 64$).

- $\mathcal{F} = \{\varnothing, \{1, 2\}, \{3, 4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}\}.$

Non-examples of sigma-algebras for $\Omega = \{1, 2, 3, 4, 5, 6\}$**:**

- $\mathcal{F} = \{\{1, 2, 3, 4, 5, 6\}\}$ is not a sigma-algebra because **P2** is not satisfied.

- $\mathcal{F} = \{\varnothing, \{1, 2, 3\}, \{4, 5, 6\}, \{1\}, \{2, 3, 4, 5, 6\}, \Omega\}$ is not a sigma-algebra because **P3** is not satisfied.

Probability measure

> **Definition 1.2.** *Let Ω be a sample space, let $\mathcal{F}$ be a sigma-algebra. A* **probability measure** *on $(\Omega, \mathcal{F})$ is a map*
> $$\mathbb{P} : \begin{aligned} \mathcal{F} &\to [0,1] \\ A &\mapsto \mathbb{P}[A] \end{aligned}$$
> *that satisfies the following two properties*
>
> **P1.** $\mathbb{P}[\Omega] = 1$.
>
> **P2.** *(**countable additivity**)* $\mathbb{P}[A] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$ *if $A = \bigcup_{i=1}^{\infty} A_i$ (disjoint union).*

"A probability measure is a map that associates to each event a number in $[0,1]$."

Examples for $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{F} = \mathcal{P}(\{1, 2, 3, 4, 5, 6\})$:

- The mapping $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ defined by

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \frac{|A|}{6}$$

  is a probability measure on $(\Omega, \mathcal{F})$.

- Given some numbers $p_1, \ldots, p_6$ satisfying $p_1 + \cdots + p_6 = 1$, the mapping $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ defined by
  $$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \sum_{i \in A} p_i$$
  is a probability measure on $(\Omega, \mathcal{F})$. The case $p_i = \frac{1}{6}$ (for all i) corresponds to the first example, modeling a fair die. The case $p_1 = \cdots = p_5 = \frac{1}{7}$, and $p_6 = \frac{2}{7}$ would correspond to a biased die, with twice more chance fall on 6 than on the other values.
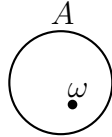
Notion of probability space

> **Definition 1.3.** *Let Ω be a sample space, $\mathcal{F}$ a sigma-algebra, and $\mathbb{P}$ a probability measure. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a* **probability space**.

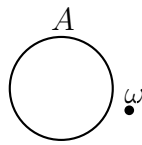To summarize, if one want to construct a probabilistic model, we give

- a sample space Ω,   "all the possible outcomes of the experiment"

- a sigma-algebra $\mathcal{F} \subset \mathcal{P}(\Omega)$,   "the set of events"

- a probability measure $\mathbb{P}$.   "gives a number in $[0,1]$ to every event"

Terminology

Let $\omega \in \Omega$ (a possible outcome). Let A be an event.
We say the event A **occurs** (for ω) if $\omega \in A$.

We say that it **does not occur** if $\omega \notin A$.

Remark 1.4. *The event $A = \varnothing$ never occurs.*   *"we never have $\omega \in \varnothing$"*
    The event $A = \Omega$ always occurs.   *"we always have $\omega \in \Omega$"*

2   Examples of probability spaces

Example with Ω finite

We now discuss a particular type of probability spaces that appear in many concrete examples. The sample space Ω is an arbitrary **finite** set, and all the outcomes have the **same** probability $p_\omega = \frac{1}{|\Omega|}$.

Definition 1.5. *Let Ω be a finite sample space. The **Laplace model** on Ω is the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where*

- *$\mathcal{F} = \mathcal{P}(\Omega)$,*

- *$\mathbb{P} : \mathcal{F} \to [0,1]$ is defined by*

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \frac{|A|}{|\Omega|}.$$

One can easily check that the mapping $\mathbb{P}$ above defines a probability measure in the sense of the definition 1.2. In this context, estimating the probability $\mathbb{P}[A]$ boils down to counting the number of elements in A and in Ω.

Example:

We consider $n \geq 3$ points on a circle, from which we select 2 at random. What is the probability that these two points selected are neighbors?
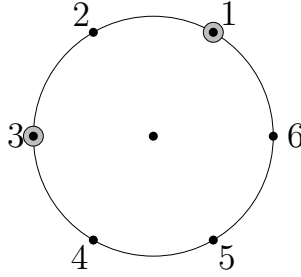
Figure 1.1: A circle with $n = 6$ points, and the subset $\{1, 3\}$ is selected.

We consider the Laplace model on

$$\Omega = \{E \subset \{1, 2, \cdots, n\} : |E| = 2\}.$$

The event "the two points of E are neighbors" is given by

$$A = \{\{1, 2\}, \{2, 3\}, \cdots, \{n - 1, n\}, \{n, 1\}\},$$

and we have

$$\mathbb{P}[A] = \frac{|A|}{|\Omega|} = \frac{n}{\binom{n}{2}} = \frac{2}{n - 1}.$$

Example with Ω infinite countable

We throw a biased coin multiple times, at each throw, the coin falls on head with probability p, and it falls on tail with probability $1 - p$ (p is a fixed parameter in $[0, 1]$). We stop at the first time we see a tail. The probability that we stop exactly at time k is given by

$$p_k = p^{k-1}(1 - p).$$

(Indeed, we stop at time k, if we have seen exactly $k - 1$ heads and 1 tail.)

For this experiment, one possible probability space is given by

- $\Omega = \mathbb{N}\backslash\{0\} = \{1, 2, 3, \ldots\}$,

- $\mathcal{F} = \mathcal{P}(\Omega)$,

- for $A \in \mathcal{F}$, $\mathbb{P}[A] = \sum_{k \in A} p_k$.

Example with Ω uncountable (for the culture)

A tempting approach to define a probability measure is to first associate to every ω the probability p_ω that the output of the experiment is ω. Then, for an event $A \subset \Omega$ define the probability of A by the formula

$$\mathbb{P}[A] = \sum_{\omega \in A} p_\omega. \tag{1.1}$$

This approach works perfectly well, when the sample space Ω is finite or countable (this is the case of the two examples above). But this approach does not work well if Ω is uncountable. For example, in the case of the droplet of water in a segment $\Omega = [0, 1]$. In this case, the probability of landing a fixed point is always 0 and the equation (1.1) does not make sense. This is for this reason that we use an axiomatic definition (in Definition 1.2) of probability measure. In this particular case, a natural choice of probability space is

- $\Omega = [0, 1]$,

- $\mathcal{F} = $ Borel σ-algebra[1]

- for $A \in \mathcal{F}$, $\mathbb{P}(A) = $ Lebesgue measure of (A).

3 Properties of Events

Operations on events and interpretation

Since events are defined as subsets of Ω, one can use operations from set theory (union, intersection, complement, symmetric difference,...). From the definition, we know that we can take the complement of an event (by H2), or a countable union of events (by H3). The following proposition asserts that the other standard set operations are allowed.

Proposition 1.6 (Consequences of the definition). *Let $\mathcal{F}$ be a sigma-algebra on Ω. We have*

P4. $\varnothing \in \mathcal{F}$,

P5. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$,

P6. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$,

P7. $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$.

Proof. We prove the items one after the other.

i. By **P1** in Definition 1.1, we have $\Omega \in \mathcal{F}$. Hence, by **P3** in Definition 1.1, we have

$$\varnothing = \Omega^c \in \mathcal{F}.$$

ii. Let $A_1, A_2, \ldots \in \mathcal{F}$. By **P2**, we also have $A_1^c, A_2^c, \ldots \in \mathcal{F}$. Then, by **P3**, we have $\bigcup_{i=1}^{\infty}(A_i)^c \in \mathcal{F}$. Finally, using **P2** again, we conclude that

$$\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty}(A_i)^c \right)^c \in \mathcal{F}.$$

---
[1]the Borel σ-algebra $\mathcal{F}$ is defined as follows: it contains all $A = [x_1, x_2] \times [y_1, y_2]$, with $0 \le x_1 \le x_2 \le 1$, $0 \le y_1 \le y_2 \le 1$, and it is the smallest collection of subsets of Ω which satisfies **P1**, **P2** and **P3** in Definition 1.1.

iii. Let $A, B \in \mathcal{F}$. Define $A_1 = A$, $A_2 = B$, and for every $i \geq 3$ $A_i = \varnothing$. By **P3**, we have
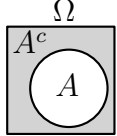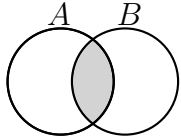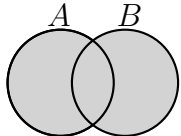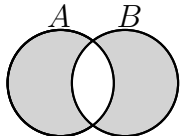
$$A \cup B = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

iv. Let $A, B \in \mathcal{F}$. By **P2**, $A^c, B^c \in \mathcal{F}$. Then by *iii.* above, we have $A^c \cup B^c \in \mathcal{F}$. Finally, by **P2**, we deduce

$$A \cap B = (A^c \cup B^c)^c \in \mathcal{F}.$$

$\square$
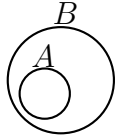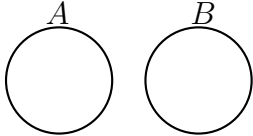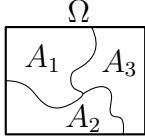
On the following tabular we consider two events A and B, and we summarize the probabilistic interpretation of the most important set operation.

Event	Graphical representation	Probab. interpretation
A^c		A does **not** occur
$A \cap B$		A **and** B occur
$A \cup B$		A **or** B occurs
$A \Delta B$		one and only one of A or B occurs

Relations between events and interpretations

Set relations (inclusion, distinctness, partition) also have probabilistic interpretations, summarized below.

Relation	Graphical representation	Probab. interpretation
$A \subset B$		If A occurs, then B occurs
$A \cap B = \varnothing$		A and B cannot occur at the same time
$\Omega = A_1 \cup A_2 \cup A_3$ with A_1, A_2, A_3 pairwise disjoint		for each outcome ω, one and only one of the events A_1, A_2, A_3 is satisfied.

Why not always working with $\mathcal{F} = \mathcal{P}(\Omega)$?

In the previous class [LSW21], the set of events was always taken to be $\mathcal{P}(\Omega)$. It may seem useless to take more general sets of events, and consider the "complicated" notion of sigma-algebra. We give here two main motivations for that:

- **First motivation: partially observed experiment.** Working with general set of event allows for natural decomposition of the probability spaces. This is particularly useful when we reveal the outcome of a random experiment algorithmically, as in the following simple example. We consider the throw of two independent dice. A possible outcome is a pair $\omega = (\omega_1, \omega_2)$ where ω_1 and ω_2 are the respective values of the first and second dies. We choose the sample space

$$\Omega = \{1, 2, 3, 4, 5, 6\}^2.$$

We can consider the following two sigma algebras

$$\mathcal{F}_1 = \{A \times \{1, 2, 3, 4, 5, 6\}, A \subset \{1, 2, 3, 4, 5, 6\}\}$$

and

$$\mathcal{F}_2 = \mathcal{P}(\Omega).$$

The first sigma-algebra $\mathcal{F}_1$ corresponds to all the events defined in terms of the first die, while the second sigma-algebra contains all the possible events in terms of the two dice. For example the event $A = \{2, 4, 6\} \times \{1, 2, 3, 4, 5, 6\}$ ("the first die is even") is in both $\mathcal{F}_1$ and $\mathcal{F}_2$, while the event $B = \{2, 4, 6\}^2$ ("both dies are even") belongs to $\mathcal{F}_2$ but not $\mathcal{F}_1$ because it requires the information of the second die.

If one reveals the outcome of the experiment algorithmically, by first revealing the first die, and then the second die. After the first step, we can say which of the events of $\mathcal{F}_1$ occur, and after the second step, we can say which of the events of $\mathcal{F}_2$ occur.

- **Second motivation: theoretical.** When the sample space is not countable (e.g. $\Omega = [0,1]$ or $\Omega = \{0,1\}^{\mathbb{N}}$), one often needs to impose some conditions on the events. This is due to the fact that we want to be able to define a probability measure on the set of events. This is not always possible on $\mathcal{F} = \mathcal{P}(\Omega)$. For example, when defining the uniform probability measure on Ω, one can construct set $A \subset [0,1]$ that are "strange enough" so that $\mathbb{P}[A]$ is not defined. Therefore we have to restrict ourselves to $\mathcal{F} \subsetneq \mathcal{P}(\Omega)$, which excludes such strange sets (see [Wil01, 2.3, p. 43] for a short discussion on this issue). For this course, this theoretical obstacle is not crucial to understand in detail. Nevertheless it is of fundamental importance in measure theory, which is the theoretical support for probability theory.

4 Properties of probability measures

Direct consequences of the definition

Proposition 1.7. *Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$.*

P3. *We have*
$$\mathbb{P}[\varnothing] = 0.$$

P4. *(**additivity**) Let $k \geq 1$, let $A_1, \ldots, A_k$ be k pairwise disjoint events, then*
$$\mathbb{P}[A_1 \cup \cdots \cup A_k] = \mathbb{P}[A_1] + \cdots + \mathbb{P}[A_k].$$

P5. *Let A be an event, then*
$$\mathbb{P}[A^c] = 1 - \mathbb{P}[A].$$

P6. *If A and B are two events (not necessarily disjoint), then*
$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

Proof. We prove the items one after the other

P3. Define $x = \mathbb{P}[\varnothing]$. We already know that $x \in [0,1]$ because x is the probability of some event. Defining $A_1 = A_2 = \cdots = \varnothing$, we have
$$\varnothing = \bigcup_{i=1}^{\infty} A_i.$$

The events A_i are disjoint and countable additivity implies
$$\sum_{i=1}^{\infty} P[A_i] = \mathbb{P}[\varnothing].$$

Since $\mathbb{P}[A_i] = x$ for every i and $P[\varnothing] \leq 1$, we have
$$\sum_{i=1}^{\infty} x \leq 1,$$

and therefore $x = 0$.

P4. Define $A_{k+1} = A_{k+2} = \cdots = \varnothing$. In this way we have

$$A_1 \cup \cdots \cup A_k = A_1 \cup \cdots \cup A_k \cup \varnothing \cup \varnothing \cup \cdots = \bigcup_{i=1}^{\infty} A_i.$$

Since the events A_i are pairwise disjoint, one can apply countable additivity as follows:

$$
\begin{aligned}
\mathbb{P}[A_1 \cup \cdots \cup A_k] \quad &= \quad \mathbb{P}[\bigcup_{i=1}^{\infty} A_i] \\
&\overset{\substack{\text{countable} \\ \text{additivity}}}{=} \quad \sum_{i=1}^{\infty} \mathbb{P}[A_i] \\
&= \quad \mathbb{P}[A_1] + \cdots + \mathbb{P}[A_k] + \underbrace{\sum_{i>k} \mathbb{P}[A_i]}_{=0}.
\end{aligned}
$$

P5. By definition of the complement, we have $\Omega = A \cup A^c$, and therefore

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[A \cup A^c].$$

Since the two events A, A^c are disjoint, additivity finally gives

$$1 = \mathbb{P}[A] + \mathbb{P}[A^c].$$

P6. $A \cup B$ is the disjoint union of A with $B\backslash A$. Hence, by additivity, we have

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B\backslash A]. \tag{1.2}$$

Also $B = (B \cap A) \cup (B \cap A^c) = (B \cap A) \cup (B\backslash A)$. Hence, by additivity,

$$\mathbb{P}[B] = \mathbb{P}[B \cap A] + \mathbb{P}[B\backslash A],$$

which give $\mathbb{P}[B\backslash A] = \mathbb{P}[B] - \mathbb{P}[A \cap B]$. Plugging this estimate in Eq. (1.2) we obtain the result.

$\square$

Useful Inequalities

In applications, it happens often that the probability $\mathbb{P}[A]$ is difficult to compute exactly: in such cases it is often useful to relate the event A to other events, and then use monotonicity (Proposition 1.8) and/or the union bound (Proposition 1.9) to obtain some bounds on $\mathbb{P}[A]$ in terms of probabilities of events that are easier to compute.

Proposition 1.8 (Monotonicity)**.** *Let $A, B \in \mathcal{F}$, then*

$$A \subset B \Rightarrow \mathbb{P}[A] \leq \mathbb{P}[B].$$

Proof. If $A \subset B$, then we have $B = A \cup (B \backslash A)$ (disjoint union). Hence, by additivity, we have

$$\mathbb{P}[B] = \mathbb{P}[A] + \mathbb{P}[B \backslash A] \geq \mathbb{P}[A].$$

$\square$

Proposition 1.9 (Union bound)**.** *Let $A_1, A_2, \ldots$ be a sequence of events (not necessarily disjoint), then we have*

$$\mathbb{P}[\bigcup_{i=1}^{\infty} A_i] \leq \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

Remark 1.10. *The union bound also applies to a finite collection of events.*

Proof. For $i \geq 1$, define

$$\widetilde{A}_i = A_i \cap A_{i-1}^c \cap \cdots \cap A_1^c.$$

One can check that

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} \widetilde{A}_i.$$

(To prove the direct inclusion, consider ω in the left hand side. Then define the smallest i such that $\omega \in A_i$. For this i, we have $\omega \in \widetilde{A}_i$, which implies that ω belongs to the right hand side. The other inclusion is clear because $\widetilde{A}_i \subset A_i$ for every i.) Now, one can apply the countable additivity to the $\widetilde{A}_i$, because they are disjoint. We get

$$\mathbb{P}[\bigcup_{i=1}^{\infty} A_i] = \mathbb{P}[\bigcup_{i=1}^{\infty} \widetilde{A}_i]$$
$$= \sum_{i=1}^{\infty} \mathbb{P}[\widetilde{A}_i]$$
$$\leq \sum_{i=1}^{\infty} \mathbb{P}[A_i].$$

$\square$

Application. We throw a die $n \geq 2$ times. We want to prove that the probability to see more than $\ell := \lceil 7 \log n \rceil$ successive 1's is small if n is large.

We consider the probability space given by

- $\Omega = \{1, 2, 3, 4, 5, 6\}^n$, an outcome is $\omega = (\underbrace{\omega_1}_{\text{die 1}}, \cdots, \underbrace{\omega_n}_{\text{die n}})$,

- $\mathcal{F} = \mathcal{P}(\Omega)$,

- for $A \in \mathcal{F}$, $\mathbb{P}[A] = \frac{|A|}{|\Omega|}$.

The event A that there exist ℓ successive 1's can be defined as follows. First, for a fixed index $k \in \{1, \ldots, n - \ell\}$, we define the event that there ℓ successive 1's between $k + 1$ and $k + \ell$ by

$$A_k = \{\omega : \omega_{k+1} = \omega_{k+2} = \cdots = \omega_{k+\ell} = 1\},$$

This way, the event A is exactly the event that there exists k such that A_k occurs. Namely,

$$A = \bigcup_{k=0}^{n-\ell} A_k.$$

Our goal is to prove that the probability of A is small. Notice that $A_k \cap A_{k'} \neq \varnothing$ for $k \neq k'$, since the element $\omega = (1, 1, \ldots, 1)$ always belongs to A_k for every index k. Hence, the event A is expressed as an non-disjoint union of events and we cannot directly Property **P2** of the probability measure to estimate its probability. Nevertheless, one can use the **union bound** to show that

$$\mathbb{P}[A] \leq \sum_{k=0}^{n-\ell} \mathbb{P}[A_k] \leq n \cdot \left(\frac{1}{6}\right)^{\ell} \leq n \cdot n^{-\log(7)/\log(6)},$$

and therefore we see that the probability of seeing more than $7 \log n$ consecutive 1's converge to 0 as n tends to infinity.

Continuity properties of probability measures

Proposition 1.11. *Let (A_n) be an increasing sequence of events (i.e. $A_n \subset A_{n+1}$ for every n). Then*

$$\lim_{n \to \infty} P[A_n] = \mathbb{P}[\bigcup_{n=1}^{\infty} A_n]. \qquad \textcolor{blue}{\textit{increasing limit}}$$

Let (B_n) be a decreasing sequence of events (i.e. $B_n \supset B_{n+1}$ for every n). Then

$$\lim_{n \to \infty} P[B_n] = \mathbb{P}[\bigcap_{n=1}^{\infty} B_n]. \qquad \textcolor{blue}{\textit{decreasing limit}}$$
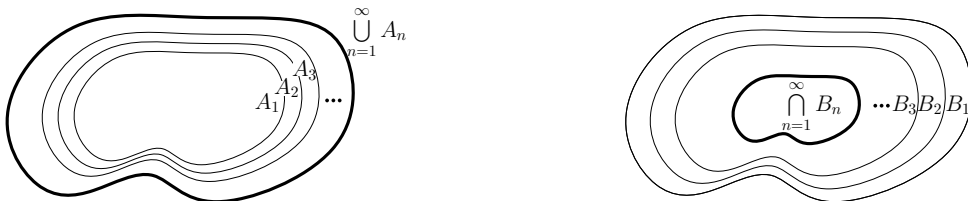
Remark 1.12. *By monotonicity, we have $\mathbb{P}[A_n] \leq \mathbb{P}[A_{n+1}]$ and $\mathbb{P}[B_n] \geq \mathbb{P}[B_{n+1}]$ for every n. Hence the limits in the proposition are well defined as monotone limits.*

Proof. Let $(A_n)_{n \geq 1}$ be an increasing sequence of events. Define $\widetilde{A}_1 = A_1$ and for every $n \geq 2$

$$\widetilde{A}_n = A_n \backslash A_{n-1}.$$

The events $\widetilde{A}_n$ are disjoint and satisfy

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \widetilde{A}_n \quad \text{and} \quad A_N = \bigcup_{n=1}^{N} \widetilde{A}_n.$$

Using first countable additivity and then additivity, we have

$$\mathbb{P}[\bigcup_{n=1}^{\infty} A_n] = \mathbb{P}[\bigcup_{n=1}^{\infty} \widetilde{A}_n]$$
$$= \sum_{n=1}^{\infty} \mathbb{P}[\widetilde{A}_n]$$
$$= \lim_{N \to \infty} \sum_{n=1}^{N} \mathbb{P}[\widetilde{A}_n]$$
$$= \lim_{N \to \infty} \mathbb{P}[A_N].$$

Now, let (B_n) be a decreasing sequence of events. Then (B_n^c) is increasing, and we can apply the previous result in the following way:

$$\mathbb{P}[\bigcap_{n=1}^{\infty} B_n] = 1 - \mathbb{P}[\bigcup_{n=1}^{\infty} B_n^c]$$
$$= 1 - \lim_{n \to \infty} \mathbb{P}[B_n^c]$$
$$= \lim_{n \to \infty} \mathbb{P}[B_n].$$

$\square$

5 Conditional probabilities

Consider a random experiment represented by some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We may sometimes possess incomplete information about the actual outcome of the experiment without knowing this outcome exactly. For example if we throw a die and a friend tells us that an even number is showing, then this information affects all our calculation of probabilities. In general, if A and B are two events and we are given that B occurs, the new probability may no longer be $\mathbb{P}[A]$. In this new circumstance, we know that A occurs if and only if $A \cap B$ occurs, suggesting that the new probability of A is proportional to $\mathbb{P}[A \cap B]$.

Definition 1.13 (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let A, B be two events with $\mathbb{P}[B] > 0$. The **conditional probability of A given B***

is defined by

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Remark 1.14. $\mathbb{P}[B \mid B] = 1.$ *Condition on B, the event B always occurs.*

Example: We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ corresponding to the throw of one die. Let $A = \{1, 2, 3\}$ be the event that the die is smaller than or equal to 3, and let $B = \{2, 4, 6\}$ be the event that the die is even. Then

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{1/6}{1/2} = 1/3.$$

Proposition 1.15. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space. Let B be an event with positive probability. Then $\mathbb{P}[\,.\,|\,B]$ is a probability measure on Ω.*

Proposition 1.16 (Formula of total probability)**.** *Let $B_1, \cdots, B_n$ be a partition[a] of the sample space Ω with $\mathbb{P}[B_i] > 0$ for every $1 \le i \le n$. Then, one has*

$$\forall A \in \mathcal{F} \quad \mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\, \mathbb{P}[B_i].$$

---
[a]i.e. $\Omega = B_1 \cup \cdots \cup B_n$ and the events are pairwise disjoint.

Proof. Using the distributivity of the intersection, we have

$$A = A \cap \Omega = A \cap (B_1 \cup \cdots \cup B_n) = (A \cap B_1) \cup \cdots \cup (A \cap B_n).$$

Since the events $A \cap B_i$ are pairwise disjoint, we have

$$\mathbb{P}[A] = \mathbb{P}[A \cap B_1] + \cdots + \mathbb{P}[A \cap B_n].$$

By definition, we have $\mathbb{P}[A \cap B_i] = \mathbb{P}[A \mid B_i]\, \mathbb{P}[B_i]$ for every i and using this expression in the equation above, we finally get

$$\mathbb{P}[A] = \mathbb{P}[A \mid B_1]\, \mathbb{P}[B_1] + \cdots + \mathbb{P}[A \mid B_n]\, \mathbb{P}[B_n].$$

$\square$

Proposition 1.17 (Bayes formula)**.** *Let $B_1, \ldots, B_n \in \mathcal{F}$ be a partition of Ω with $\mathbb{P}[B_i] > 0$ for every i. For every event A with $\mathbb{P}[A] > 0$, we have*

$$\forall i = 1, \ldots, n \quad \mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\, \mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}[A \mid B_j]\, \mathbb{P}[B_j]}.$$

Typical application: A test is performed in order to diagnose a certain rare disease, which concerns 1/10000 of a population. This test is quite reliable and gives the right answer 99 percent of the times. If a patient has a "positive test" (i.e. the test indicates that he is sick), what is the probability that he is actually sick?

The situation is modeled by setting

$$\Omega = \{0, 1\} \times \{0, 1\}.$$

and $\mathcal{F} = \mathcal{P}(\Omega)$. An outcome is a pair $\omega = (\omega_1, \omega_2)$ representing a patient, where

$$\omega_1 = \begin{cases} 0 & \text{if the patient is healthy,} \\ 1 & \text{if the patient is sick,} \end{cases}$$

$$\omega_2 = \begin{cases} 0 & \text{if the test is negative,} \\ 1 & \text{if the test is positive.} \end{cases}$$

We consider the event S that the patient is sick, and the event T that the test is positive. The elements of S are all the outcomes $\omega = (\omega_1, \omega_2)$ such that $\omega_1 = 1$, ie

$$S = \{(1, 0), (1, 1)\}.$$

Equivalently, we have

$$T = \{(0, 1), (1, 1)\}.$$

From the hypotheses, the information that we have on the probability measure is

$$\mathbb{P}[S] = \frac{1}{10000}, \quad \mathbb{P}[T \mid S] = \frac{99}{100}, \quad \mathbb{P}[T \mid S^c] = \frac{1}{100}.$$

We are looking for the a posteriori probability $\mathbb{P}[S \mid T]$ of being sick, given that the test is positive. By applying the Bayes formula to the partition $\Omega = S \cup S^c$, we obtain

$$\begin{aligned} \mathbb{P}[S \mid T] &= \frac{\mathbb{P}[T \mid S] \, \mathbb{P}[S]}{\mathbb{P}[T \mid S] \, \mathbb{P}[S] + \mathbb{P}[T \mid S^c] \, \mathbb{P}[S^c]} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \simeq 0.0098. \end{aligned}$$

This result is quite surprising: the probability to be actually sick when the test is positive is very small!

What is happening? If one looks at the whole population, there are two types of persons, who will have a positive test:

- the healthy individuals with a (wrongly) positive test, which represent roughly one percent of the population.

- the sick individuals with a (correctly) positive test, which represent roughly 1/10000 of the population.

Given that the test is positive, a person has much more chances to be in the first group of individuals.

6 Independence

Independence of events

> **Definition 1.18** (Independence of two events)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. Two events A and B are said to be* **independent** *if*
>
> $$\mathbb{P}[A \cap B] = \mathbb{P}[A] \ \mathbb{P}[B].$$

Remark 1.19. *If* $\mathbb{P}[A] \in \{0, 1\}$*, then A is independent of every event, i.e.*

$$\forall B \in \mathcal{F} \quad \mathbb{P}[A \cap B] = \mathbb{P}[A] \ \mathbb{P}[B].$$

If an event A is independent with itself (i.e. $\mathbb{P}[A \cap A] = \mathbb{P}[A]^2$), then $\mathbb{P}[A] \in \{0, 1\}$ A is independent of B if and only if A is independent of B^c.

The concept of independence is fundamental in probability: it corresponds to the intuitive idea that two events do not influence each other, as illustrated in the following proposition.

Proposition 1.20. *Let $A, B \in \mathcal{F}$ be two events with $\mathbb{P}[A], \mathbb{P}[B] > 0$. Then the following are equivalent:*

(i) $\mathbb{P}[A \cap B] = \mathbb{P}[A] \ \mathbb{P}[B],$ *A and B are independent*

(ii) $\mathbb{P}[A \,|\, B] = \mathbb{P}[A],$ *the occurrence of B has no influence on A*

(iii) $\mathbb{P}[B \,|\, A] = \mathbb{P}[B].$ *the occurrence of A has no influence on B*

Proof. Since $\mathbb{P}[B] > 0$ we have

$$(i) \Leftrightarrow \left(\frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \mathbb{P}[A] \right) \Leftrightarrow (\mathbb{P}[A \,|\, B] = \mathbb{P}[A]) \Leftrightarrow \ (ii).$$

Since (iii) is just the same as (ii) with the role of A and B reversed, the equivalence (i)$\Leftrightarrow$ (iii) can proved the same way. $\qquad\square$

Typical examples of independent events occur when one performs successively a random experiment, as illustrated below with the throw of two dice.

Example: Throw of two independent dice.
We throw two dice independently. This is modeled by the Laplace model on the sample space

$$\Omega = \{1, 2, 3, 4, 5, 6\}^2.$$

Consider the events

$$A = \{\omega : \omega_1 \in 2\mathbb{Z}\}, \qquad \text{``The first die is even''}$$
$$B = \{\omega : 1 + \omega_2 \in 2\mathbb{Z}\}, \qquad \text{``The second die is odd''}$$
$$C = \{\omega : \omega_1 + \omega_2 \leq 3\}, \qquad \text{``The sum of the two dice is at most 3''}$$
$$D = \{\omega : \omega_1 \leq 2, \omega_2 \leq 2\}. \qquad \text{``Both dice are smaller than or equal to 2''}$$

Check that:

- A and B are independent,

- A and C are not independent,

- A and D are independent.

Definition 1.21. *Let I be an arbitrary set of indices. A collection of events $(A_i)_{i \in I}$ is said to be* **independent** *if*

$$\forall J \subset I \text{ finite} \qquad \mathbb{P}[\bigcap_{j \in J} A_j] = \prod_{j \in J} \mathbb{P}[A_j].$$

Remark: Three events A, B and C are independent if the following 4 equations are satisfied (and not only the last one!):

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \, \mathbb{P}[B],$$
$$\mathbb{P}[A \cap C] = \mathbb{P}[A] \, \mathbb{P}[C],$$
$$\mathbb{P}[B \cap C] = \mathbb{P}[B] \, \mathbb{P}[C],$$
$$\mathbb{P}[A \cap B \cap C] = \mathbb{P}[A] \, \mathbb{P}[B] \, \mathbb{P}[C].$$

Example: We consider the same notation as in the example above Definition 1.21. The events A, B, and D are independent (Check that!).

Chapter 2

Random variables and distribution functions

Goals

- Understand the definition of a random variable and its distribution function.

- Use of abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

- Learn the notation allowing to define events in term of random variables.

- Explicit construction of random variables from infinite sequence of i.i.d. Bernoulli random variables.

1 Abstract definition

Most often, the probabilistic model under consideration is rather complicated, and one is only interested in certain quantities in the model. For this reason, one introduces the notion of random variables

---

Definition 2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable** (r.v.) is a map $X : \Omega \to \mathbb{R}$ such that for all $a \in \mathbb{R}$,*

$$\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}.$$

➜ The condition $\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$ is needed for $\mathbb{P}[\{\omega \in \Omega : X(\omega) \leq a\}]$ to be well-defined.

Example 1: *Gambling with one die*

We throw a fair die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ and we consider the Laplace model $(\Omega, \mathcal{F}, \mathbb{P})$ as in Definition 1.5. Suppose that we gamble on the outcome in such a way that our profit is

$$-1 \text{ if the outcome is 1, 2 or 3,}$$
$$0 \text{ if the outcome is 4,}$$
$$2 \text{ if the outcome is 5 or 6,}$$

where a negative profit correspond to a loss. Our profit can be represented by the mapping X defined by

$$\forall \omega \in \Omega \quad X(\omega) = \begin{cases} -1 & \text{if } \omega = 1, 2, 3, \\ 0 & \text{if } \omega = 4, \\ 2 & \text{if } \omega = 5, 6. \end{cases} \tag{2.1}$$

Since $\mathcal{F} = \mathcal{P}(\Omega)$, we have $\{\omega : X(\omega) \leq a\} \in \mathcal{F}$ for every a. Therefore, X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.

Example 2: *Indicator function of an event*

Let $A \in \mathcal{F}$. Consider the **indicator function** $\mathbb{1}_A$ of A, defined by

$$\forall \omega \in \Omega \quad \mathbb{1}_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A, \\ 1 & \text{if } \omega \in A. \end{cases}$$

Then $\mathbb{1}_A$ is a random variable. Indeed, we have

$$\{\omega : \mathbb{1}_A(\omega) \leq a\} = \begin{cases} \varnothing & \text{if } a < 0, \\ A^c & \text{if } 0 \leq a < 1, \\ \Omega & \text{if } a \geq 1, \end{cases}$$

and $\varnothing$, A^c and Ω are three elements of $\mathcal{F}$.

Remark: *Role of the sigma-algebra*

Consider the same notation as in Example 1. Additionally, we consider the following two sigma-algebras:

$$\mathcal{F}_1 = \{\varnothing, \{1,2,3\}, \{4,5,6\}, \{1,2,3,4,5,6\}\},$$
$$\mathcal{F}_2 = \{\varnothing, \{1,2,3\}, \{1,2,3,4\}, \{4,5,6\}, \{5,6\}, \{1,2,3,5,6\}, \{4\}, \{1,2,3,4,5,6\}\}.$$

Is X is random variable on $(\Omega, \mathcal{F}_i, \mathcal{P}\}$? To answer this question, one needs to examine the set $\{\omega : X(\omega) \le a\}$ in more details. Here we see that

$$\{\omega : X(\omega) \le a\} = \begin{cases} \varnothing & \text{if } a < -1, \\ \{1,2,3\} & \text{if } -1 \le a < 0, \\ \{1,2,3,4\} & \text{if } 0 \le a < 2, \\ \{1,2,3,4,5,6\} & \text{if } a \ge 2. \end{cases}$$

In particular, we see that X is a random variable on $(\Omega, \mathcal{F}_2, \mathcal{P})$, but not on $(\Omega, \mathcal{F}_1, \mathcal{P})$.

Notation: When events are defined in terms of random variable, we will **omit the dependence in** ω. For example, for $a \le b$ we write

$$\{X \le a\} = \{\omega \in \Omega : X(\omega) \le a\},$$
$$\{a < X \le b\} = \{\omega \in \Omega : a < X(\omega) < b\},$$
$$\{X \in \mathbb{Z}\} = \{\omega \in \Omega : X(\omega) \in \mathbb{Z}\}.$$

When consider the probability of events as above, we omit the brackets and for example simply write

$$\mathbb{P}[X \le a] = \mathbb{P}[\{X \le a\}] = \mathbb{P}[\{\omega \in \Omega : X(\omega) \le a\}].$$

---

2 Distribution function

Definition 2.2. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The **distribution function of** X is the function $F_X : \mathbb{R} \to [0,1]$ defined by*

$$\forall a \in \mathbb{R} \quad F_X(a) = \mathbb{P}[X \le a].$$

Idea: The distribution function F_X encodes the probabilistic properties of the random variable X.

Example 1: *Gambling on a die*

Let X be the random variable defined by Eq. (2.1). For $a \in \mathbb{R}$, we have

$$F_X(a) = \begin{cases} 0 & \text{if } a < -1, \\ 1/2 & \text{if } -1 \le a < 0, \\ 2/3 & \text{if } 0 \le a < 2, \\ 1 & \text{if } a \ge 2. \end{cases}$$
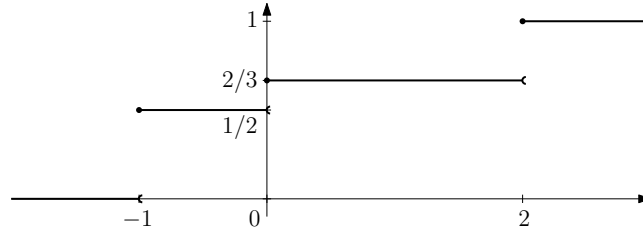

Figure 2.1: Graph of the distribution function F_X.

Example 2: *Indicator function of an event*

Let A be an event. Let $X = \mathbb{1}_A$ be the indicator function of the event A. Then

$$F_X(a) = \begin{cases} 0 & \text{if } a < 0, \\ 1 - \mathbb{P}[A] & \text{if } 0 \le a < 1, \\ 1 & \text{if } a \ge 1. \end{cases}$$

Proposition 2.3 (Basic identity). *Let $a < b$ be two real numbers. Then*

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

Proof. We have $\{X \le b\} = \{X \le a\} \cup \{a < X \le b\}$ (disjoint union). Hence

$$\mathbb{P}[X \le b] = \mathbb{P}[X \le a] + \mathbb{P}[a < X \le b],$$

which directly implies the result. □

Theorem 2.4 (Properties of distribution functions). *Let X be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution function $F = F_X : \mathbb{R} \to [0,1]$ of X satisfies the following properties.*

  (*i*) *F is nondecreasing.*

  (*ii*) *F is right continuous[a].*

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

---

^{a}i.e. $F(a) = \lim\limits_{h\downarrow 0} F(a + h)$ for every $a \in \mathbb{R}$.

Proof. We first prove (i), then (iii) and finally (ii).

(i) For $a \le b$, we have $\{X \le a\} \subset \{X \le b\}$. Hence, by monotonicity, we have $\mathbb{P}[X \le a] \le \mathbb{P}[X \le b]$, i.e.
$$F(a) \le F(b).$$

(iii) Let $a_n \uparrow \infty$. For every $\omega \in \Omega$, there exists n large enough such that $X(\omega) \le a_n$. Hence,
$$\Omega = \bigcup_{n \ge 1} \{X \le a_n\}.$$

Furthermore, we have $\{X \le a_n\} \subset \{X \le a_{n+1}\}$ and the continuity properties of probability measures imply
$$1 = \mathbb{P}[\Omega] = \mathbb{P}\Big[\bigcup_{n \ge 1} \{X \le a_n\}\Big]$$
$$= \lim_{n \to \infty} \mathbb{P}[X \le a_n]$$
$$= \lim_{n \to \infty} F(a_n).$$

In the same way, one has $\lim\limits_{a \to -\infty} F(a) = 0$. Indeed, using that for every $a_n \downarrow -\infty$,
$$\varnothing = \bigcap_{n \ge 1} \{X \le a_n\}$$

and $\{X \le a_n\} \supset \{X \le a_{n+1}\}$, it follows from the continuity properties of probability measures that
$$0 = \mathbb{P}[\varnothing] = \lim_{n \to \infty} \mathbb{P}[X \le a_n] = \lim_{n \to \infty} F(a_n).$$

(ii) Let $a \in \mathbb{R}$, let $h_n \downarrow 0$. We have
$$\{X \le a\} = \bigcap_{n \ge 1} \{X \le a + h_n\},$$

where $\{X \le a + h_n\} \supset \{X \le a + h_{n+1}\}$. Hence by the continuity properties of probability measures, we have
$$F(a) = \mathbb{P}[X \le a] = \mathbb{P}\Big[\bigcap_{n \ge 1} \{X \le a + h_n\}\Big] = \lim_{n \to \infty} \mathbb{P}[X \le a + h_n] = \lim_{n \to \infty} F[a + h_n].$$

$\square$

3   Independence

Independence of random variables

> **Definition 2.5.** *Let $X_1, \ldots, X_n$ be n random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $X_1, \ldots, X_n$ are **independent** if*
>
> $$\forall x_1, \ldots, x_n \in \mathbb{R} \qquad \mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n] = \mathbb{P}[X_1 \leq x_1] \ldots \mathbb{P}[X_n \leq x_n]. \qquad (2.2)$$

Remark: One can show that $X_1, \ldots, X_n$ are independent if and only if

$$\forall I_1 \subset \mathbb{R}, \ldots, I_n \subset \mathbb{R} \text{ intervals} \qquad \{X_1 \in I_1\}, \ldots, \{X_n \in I_n\} \text{ are independent.}$$

Example 1: *Throw of two independent dices*

We consider the Laplace model $(\Omega, \mathcal{F}, \mathbb{P})$ on $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. An element $\omega \in \Omega$ is a pair $\omega = (\omega, \omega_2)$, where the first coordinate represents the value of the first die and the second coordinate represents the value of the second die. We define the random variables $X, Y, Z : \Omega \to \mathbb{R}$ by

$$X(\omega) = \omega_1, \quad Y(\omega) = \omega_2, \quad Z(\omega) = \omega_1 + \omega_2.$$

X and Y represent the values of the first and second die, respectively. Z corresponds to the sum of the two dices. In this case, we have that X and Y are independent. To see this, observe that for every $I, J \subset \{1, \ldots, 6\}$, we have

$$\mathbb{P}[X \in I, Y \in J] = \mathbb{P}[I \times J] = \frac{|I \times J|}{|\Omega|} = \frac{|I|}{6} \cdot \frac{|J|}{6}$$
$$= \frac{|I \times \{1, 2, 3, 4, 5, 6\}|}{36} \cdot \frac{|I \times \{1, 2, 3, 4, 5, 6\}|}{36} = \mathbb{P}[X \in I]\mathbb{P}[Y \in J].$$

For every $x, y \in \mathbb{R}$, there exist $I, J \subset \{1, 2, 3, 4, 5, 6\}$ such that $\{X \leq x\} = \{X \in I\}$ and $\{X \leq y\} = \{X \in J\}$. Therefore,

$$\mathbb{P}[X \leq x, Y \leq y] = \mathbb{P}[X \leq x]\mathbb{P}[Y \leq y].$$

On the other hand, X and Z are not independent since

$$\frac{1}{6^2} = \mathbb{P}[X \leq 1, Z \leq 2] \neq \mathbb{P}[X \leq 1]\mathbb{P}[Z \leq 2] = \frac{1}{6^3}.$$

Grouping

If we have a set of independent random variables, and we make disjoint groups of such random variables, then these groups are also independent form each other. This idea is formalized by the following proposition.

Proposition 2.6 (grouping). *Let $X_1, \ldots, X_n$ be n independent random variables. Let $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ be some indices and $\phi_1, \ldots, \phi_k$ some functions. Then*

$$Y_1 = \phi_1(X_1, \ldots, X_{i_1}), Y_2 = \phi_2(X_{i_1+1}, \ldots, X_{i_2}) \ldots, Y_k = \phi_k(X_{i_{k-1}+1}, \ldots, X_{i_k})$$

are independent.

Proof. Admitted $\qquad \square$

Sequences of i.i.d. random variables

> **Definition 2.7.** *An infinite sequence $X_1, X_2, \ldots$ of random variables is said to be*
>
> - ***independent*** *if $X_1, \ldots, X_n$ are independent, for every n.*
>
> - ***independent and identically distributed (iid)*** *if they are independent and they have the same distribution function, i.e.*
>
> $$\forall i, j \quad F_{X_i} = F_{X_j}.$$

4   Transformation of random variables

Once we have some random variables $X_1, X_2, \ldots$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we can create and consider many new random variables on the same probability space by using operations. For example, one can consider $Z_1 = \exp(X_1)$, $Z_2 = X_1 + X_2, \ldots$ One should not completely forget that random variables are maps $\Omega \to \mathbb{R}$. For example, the random variables Z_1 and Z_2 correspond to the maps defined by for every $\omega \in \Omega$

$$Z_1(\omega) = \exp(X_1(\omega)), \qquad Z_2(\omega) = X_1(\omega) + X_2(\omega).$$

Formally, we introduce the following notation, which allows us to work with random variables as if the were just real numbers. If X is a random variable, and $\phi : \mathbb{R} \to \mathbb{R}$, the we write

$$\phi(X) := \phi \circ X.$$

This way, $\phi(X)$ is a new mapping $\Omega \to \mathbb{R}$ as shown on the diagram.

$$
\begin{array}{ccccc}
\Omega & \xrightarrow{X} & \mathbb{R} & \xrightarrow{\phi} & \mathbb{R} \\
\omega & \longmapsto & X(\omega) & \longmapsto & \phi(X(\omega)).
\end{array}
$$

More generally, we can also consider functions of several variables. If $X_1, \ldots, X_n$ are n random variables and $\phi : \mathbb{R}^n \to \mathbb{R}$, then we write

$$\phi(X_1, \ldots, X_n) := \phi \circ (X_1, \ldots, X_n).$$

5   Construction of random variables

In Section 1, we defined random variables. In Section 2, we saw that we can associate to any random variable X a distribution function $F = F_X : \mathbb{R} \to [0, 1]$, which encodes it probabilistic properties, and satisfies

(i) F is nondecreasing,

(ii) F is right continuous,

(iii) $\lim_{a \to -\infty} F(a) = 0$ and $\lim_{a \to \infty} F(a) = 1$.

Conversely, given a function $F : \mathbb{R} \to [0,1]$ satisfying Items (i)–(iii), does there exist a random variable X such that $F_X = F$?

The goal of this section is construct general random variables, and answer to the question above positively. A complete construction would require some tools from measure theory which are beyond the scope of this class: Our approach will rely on an abstract theorem of Kolmogorov, that guarantees existences of iid sequences. This Theorem will be admitted, but the rest of the construction will be rigorously detailed. Our motivation is twofold

- On a theoretical level, the existence of random variables is fundamental: "it is more satisfying the objects we are talking about exist!"

- On a practical level, the explicit construction provided here gives a general recipe to construct random variables. This can be used to simulate an arbitrary random variable, provided its distribution function.

The construction proceeds in 4 steps.

Step 1: Kolmogorov theorem and iid sequence of Bernoulli random variables

Our construction start with Bernoulli random variables, that we now define.

Definition 2.8. *Let $p \in [0,1]$. A random variable X is said to be a **Bernoulli random variable with parameter** p if*

$$\mathbb{P}[X = 0] = 1 - p \quad and \quad \mathbb{P}[X = 1] = p.$$

In this case, we write $X \sim \mathrm{Ber}(p)$.

Example: *Flipping n coins*

We wish to define a model for n successive independent coin flips. Consider the sample space $\Omega = \{0,1\}^n$ equipped with the Laplace model $(\mathcal{F}, \mathbb{P})$. Define the random variables

$$X_i : \begin{array}{ccc} \Omega & \to & \{0,1\} \\ (\omega_1, \ldots, \omega_n) & \mapsto & \omega_i \end{array} .$$

(X_i represents the result of the i-th coin flip, $X_i = 1$ if the i-th coin flip is a head, $X_i = 0$ if it is a tail.) Then the random variables $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter $1/2$.

To prove that $X_1 \sim \mathrm{Ber}(1/2)$, we compute the probability of the events $\{X_1 = 0\} = \{0\} \times \{0,1\}^{n-1}$ and $\{X_1 = 1\} = \{1\} \times \{0,1\}^{n-1}$ using the definition of $\mathbb{P}$ for the Laplace model:

$$\mathbb{P}[X_1 = 0] = \frac{|\{0\} \times \{0,1\}^{n-1}|}{|\Omega|} = \frac{1}{2} \quad \text{and} \quad \mathbb{P}[X_1 = 1] = \frac{|\{1\} \times \{0,1\}^{n-1}|}{|\Omega|} = \frac{1}{2}.$$

Equivalently, one can prove that each X_i, $1 \le i \le n$ is a Bernoulli random variable with parameter $1/2$.

To prove independence, it suffices to prove Equation (2.2) for $x_1, \ldots, x_n \in \{0, 1\}$. For such numbers, using $|\{0, x_i\}| = 1 + x_i$, we have

$$
\begin{aligned}
\mathbb{P}[X_1 \le x_1, \ldots, X_n \le x_n] &= \mathbb{P}[\{0, x_1\} \times \cdots \times \{0, x_n\}] \\
&= \frac{|\{0, x_1\} \times \cdots \times \{0, x_n\}|}{|\Omega|} \\
&= \frac{1 + x_1}{2} \cdots \frac{1 + x_n}{2} = \mathbb{P}[X_1 \le x_1] \cdots \mathbb{P}[X_n \le x_n].
\end{aligned}
$$

For every $n \ge 1$ the example above constructs a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameter $1/2$. Similarly, it is natural to consider an infinite sequence of independent Bernoulli random variables $X_1, X_2, \ldots$. The construction of a suitable probability space is much more delicate and it is the content of the following theorem.

> **Theorem 2.9** (Existence theorem of Kolmogorov)**.** *There exists a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an infinite sequence of random variables* $X_1, X_2, \ldots$ *(on this probability space) that is an iid sequence of Bernoulli random variables with parameter* $1/2$*.*

Proof. Admitted. $\qquad\square$

Step 2: Construction of a uniform random variable in $[0, 1]$

Here we use Bernoulli random variables to construct a uniform random variable in $[0, 1]$. Intuitively, one can imagine a droplet of water falling in the interval $[0, 1]$. We assume that the droplet falls on the interval homogeneously. For example, the probability to fall in $[0, 1, 0.2]$ is the same as to fall in $[0.8, 0.9]$. A uniform random variable in $[0, 1]$ represents the position at which such a droplet falls.

Definition 2.10. *A random variable U is said to be a* **uniform random variable in** $[0, 1]$ *if its distribution function is equal to*

$$
F_U(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1. \end{cases}
$$

In this case, we write $U \sim \mathcal{U}([0, 1])$.

Let $X_1, X_2, \ldots$ be a sequence of independent Bernoulli random variables with parameter $1/2$. For every fixed ω, we have $X_1(\omega), X_2(\omega) \cdots \in \{0, 1\}$. Hence the infinite series

$$
Y(\omega) = \sum_{n=1}^{\infty} 2^{-n} X_n(\omega) \tag{2.3}
$$

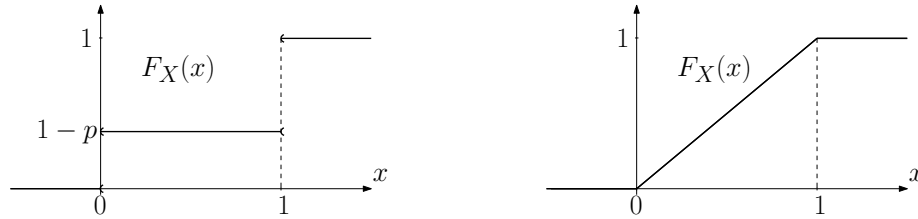is absolutely convergent, and we have $Y(\omega) \in [0, 1]$.

Figure 2.2: Left: distribution function of a Bernoulli r.v. with parameter p . Right: distribution function of a uniform random variable in $[0, 1]$.

Proposition 2.11. *The mapping $Y : \Omega \to [0, 1]$ defined by Equation (2.3) is a uniform random variable in $[0, 1]$.*

Step 3: Construction of a random variable with an arbitrary distribution F

Let $F : \mathbb{R} \to [0, 1]$ satisfying Items (i)–(iii) at the beginning of the section.

If F is strictly increasing and continuous then F is one to one and one can define its inverse F^{-1}. For every $\alpha \in [0, 1]$, $F^{-1}(\alpha)$ is the unique real number x such that $F(x) = \alpha$. In such a case, this defines the inverse distribution function. More generally, we can define a generalized inverse for F.

Definition 2.12 (Generalized inverse). *The generalized inverse of F is the mapping $F^{-1} : (0, 1) \to \mathbb{R}$ defined by*

$$\forall \alpha \in (0, 1) \quad \boxed{F^{-1}(\alpha) = \inf\{x \in \mathbb{R} \,:\, F(x) \geq \alpha\}.}$$

By definition of the infimum and using right continuity of F, we have for every $x \in \mathbb{R}$ and $\alpha \in (0, 1)$

$$(F^{-1}(\alpha) \leq x) \iff (\alpha \leq F(x)).$$

Relying on this general inverse function, the following theorem provides a way to a construct random variable with arbitrary distribution functions.

Theorem 2.13 (inverse transform sampling). *Let $F : \mathbb{R} \to [0, 1]$ satisfying Items (i)–(iii) at the beginning of the section. Let U be a uniform random variable in $[0, 1]$. Then the random variable*

$$X = F^{-1}(U) \tag{2.4}$$

has distribution $F_X = F$.

Remark 2.14. *Formally, there is an issue in the definition of X in Eq. (2.4). Indeed, we have $U : \Omega \to [0, 1]$ and $F^{-1} : (0, 1) \to \mathbb{R}$. Nevertheless, we have $\mathbb{P}[U \in (0, 1)] = 1$, and therefore X is well defined on a set of probability 1, and we can easily fix the issue by defining*

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

(the value 0 in the second case plays no role and could be replaced by any real number).

Proof. For every $x \in \mathbb{R}$, we have

$$\mathbb{P}[X \leq x] = \mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[U \leq F(x)] = F(x).$$

$\square$

Step 4: General sequence of independent random variables

Theorem 2.15. *Let $F_1, F_2 \dots$ be a sequence of functions $\mathbb{R} \to [0, 1]$ satisfying Items (i)– (iii) at the beginning of the section. Then there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of independent random variables $X_1, X_2, \dots$ on this probability space such that*

- *for every i X_i has distribution function F_i (i.e. $\forall x \ \mathbb{P}[X_i \leq x] = F_i(x)$), and*

- *$X_1, X_2, \dots$ are independent.*

Proof. See exercise. $\square$

The theorem above is important in the theory because it allows us to work with random variables directly without defining precisely the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For example, if F and G are two given distribution functions, it allows us for example to write:

"Let X, Y be two independent random variables with distribution function F and G resp.".

Chapter 3

Discrete and continuous random variables

Goals

- Definition of discrete and continuous random variables.

- Classical examples of discrete and continuous random variables: motivation, relation between them.

- Probabilistic interpretation of the analytic properties of F_X.

- Density f_X of a random variable: interpretation, relation with the distribution function F_X.

Framework We fix some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All the random variables considered in this chapter will be defined on this reference probability space.

1 Discontinuity/continuity points of F

We have seen that the distribution function $F = F_X$ of a random variable X is always right continuous. What about the left continuity?

For a Bernoulli random variable $X \sim \text{Ber}(p)$ with $p < 1$, we have $F_X(-h) = 0$ for every $h > 0$, but $F_X(0) = 1 - p \neq 0$. Therefore, F_X is not left continuous at 0, i.e.

$$\lim_{h \downarrow 0} F_X(-h) = 0 \neq F_X(0).$$

One can see this on Fig. 5, which shows a jump of $F_X(x)$ at $x = 0$.

In contrast, the distribution function of the Uniform random variable, represented on Fig. 5 is continuous on $\mathbb{R}$, in particular, it is left continuous at every point: for a uniform random variable U, we have

$$\forall a \in \mathbb{R} \quad \lim_{h \downarrow 0} F_U(a - h) = F_U(a).$$

The following proposition gives an interpretation of the left limit

$$F(a-) \coloneqq \lim_{h \downarrow 0} F(a - h)$$

at a given point a for a general distribution function.

Proposition 3.1 (probability of a given value)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution function F. Then for every a in $\mathbb{R}$ we have*

$$\mathbb{P}[X = a] = F(a) - F(a-)$$

We omit the proof, which can easily be obtained using the basic identity of Proposition 2.3 together with the continuity properties of probability measures (Prop. 1.11). We rather insist on the interpretation of this proposition.

Fix $a \in \mathbb{R}$.

➜ If F is not continuous at a point $a \in \mathbb{R}$, then the "jump size" $F(a) - F(a-)$ is equal to the probability that $X = a$.

➜ If F is continuous at a point $a \in \mathbb{R}$, then $\mathbb{P}[X = a] = 0$.

---

2 Almost sure events

An important notion when working with random variables is the notion of almost sure occurrence for an event.

Definition 3.2. *Let $A \in \mathcal{F}$ be an event. We say that A occurs **almost surely (a.s.)** if*

$$\mathbb{P}[A] = 1.$$

Remark 3.3. *This notion can be extended to any set $A \subset \Omega$ (not necessarily an event): We say that A occurs **almost surely** if there exists an event $A' \in \mathcal{F}$ such that $A' \subset A$ and $\mathbb{P}[A'] = 1$.*

In other words, something occurs a.s. if it occurs with probability 1. For example, if X, Y are two random variables, we write

$$X \leq Y \quad \text{a.s.}$$

if $\mathbb{P}[X \leq Y] = 1$, and

$$X \leq a \quad \text{a.s.}$$

if $\mathbb{P}[X \leq a] = 1$.

3 Discrete random variables

Definition 3.4 (Discrete random variables)**.** *A random variable $X : \Omega \to \mathbb{R}$ is said to be **discrete** if there exists some set $W \subset \mathbb{R}$ finite or countable such that*

$$X \in W \quad \text{a.s.}$$

Remark 3.5. *If the sample space Ω is finite or countable, then every random variable $X : \Omega \to \mathbb{R}$ is discrete. Indeed, the image $X(\Omega) = \{x \in \mathbb{R} : \exists \omega \in \Omega \ X(\omega) = x\}$ is finite or countable and we have $\mathbb{P}[X \in W] = 1$, with $W = X(\Omega)$.*

Definition 3.6. *Let X be a discrete random variable taking some values in some finite or countable set $W \subset \mathbb{R}$. The **distribution of** X is the sequence of numbers $(p(x))_{x \in W}$ defined by*

$$\forall x \in W \quad p(x) := \mathbb{P}[X = x].$$

Proposition 3.7. *The distribution $(p(x))_{x \in W}$ of a discrete random variable satisfies*

$$\sum_{x \in W} p(x) = 1.$$

Proof. We have

$$\{X \in W\} = \bigcup_{x \in W} \{X = x\}.$$

Since the union is disjoint and the set W is at most countable, we have

$$1 = \mathbb{P}[X \in W] = \mathbb{P}\big[\bigcup_{x \in W} \{X = x\}\big] = \sum_{x \in W} \mathbb{P}[X = x] = \sum_{x \in W} p(x).$$

$\square$

Let us give 3 examples of discrete random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, the Laplace model on $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 1: *Value of the die*

Consider the random variable $X : \Omega \to \mathbb{R}$ defined by

$$\forall \omega \in \Omega \quad X(\omega) := \omega$$

(X represents the value of the die). Then X takes values in

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

almost surely. Hence it is discrete and its distribution is given by

$$\forall x \in W \quad p(x) = \mathbb{P}[X = x] = \frac{1}{6}.$$

Example 2: *Gambling with one die*

Consider the random variable defined by

$$\forall \omega \in \Omega \quad X(\omega) := \begin{cases} -1 & \text{if } \omega = 1, 2, 3, \\ 0 & \text{if } \omega = 4, \\ 2 & \text{if } \omega = 5, 6. \end{cases}$$

as in Example 1 Page 21. Then X takes values in

$$W = \{-1, 0, 2\}$$

almost surely and its distribution is given by

$$p(-1) = \frac{1}{2}, \quad p(0) = \frac{1}{6}, \quad p(2) = \frac{1}{3}.$$

Example 3: *Multiple of 3*

Consider the random variable defined by

$$\forall \omega \in \Omega \quad X(\omega) := \begin{cases} 1 & \text{if } \omega \in \{3, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

(X is the indicator function that the die is a multiple of 3). Then X takes values in

$$W = \{0, 1\}$$

almost surely and its distribution is given by

$$p(0) = \frac{2}{3}, \quad p(1) = \frac{1}{3}.$$

Following Definition 3.4, X is a Bernoulli random variable with parameter 1/3.

Remark 3.8. *Conversely, if we are given a sequence of numbers $(p(x))_{x \in W}$ with values in $[0, 1]$ and such that*

$$\sum_{x \in W} p(x) = 1, \tag{3.1}$$

then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X with associated distribution $(p(x))$. This is a consequence of the existence theorem 2.15 in Chapter 2. This observation is important in practice, it allows us to write:

"Let X be a discrete random variable with distribution $(p(x))_{x \in W}$."

Distribution p vs distribution function F_X

From p to F_X

Proposition 3.9. *Let X be a discrete random variable with values in a finite or countable set W almost surely, and distribution p. Then the distribution function of X is given by*

$$\boxed{\forall x \in \mathbb{R} \quad F_X(x) = \sum_{\substack{y \leq x \\ y \in W}} p(y)} \tag{3.2}$$

Proof. For every $x \in \mathbb{R}$ we have

$$\mathbb{P}[X \leq x] = \mathbb{P}[X \in (-\infty, x] \cap W]] + \underbrace{\mathbb{P}[X \in (-\infty, x] \cap W^c]}_{\leq \mathbb{P}[X \in W^c] = 0} = \mathbb{P}[\bigcup_{\substack{y \leq x \\ y \in W}} \{X = y\}] = \sum_{\substack{y \leq x \\ y \in W}} \mathbb{P}[X = y].$$

$\square$

From F_X to p   Given a discrete random variable X, Equation (3.2) expresses the distribution function F_X in terms of p as a piecewise constant function. Conversely, a random variable with a piecewise constant distribution function F is discrete and W and p are given by

$$W = \{\text{positions of the jumps of } F_X\},$$
$$p(x) = \text{"height of the jump" at } x \in W.$$

4 Examples of discrete random variables

Bernoulli distribution

The simplest (non constant) random variable is the Bernoulli random variable. It was defined already in the previous chapter. We recall its definition here.

Definition 3.10 (Bernoulli)**.** *Let $0 \le p \le 1$. A random variable X is said to be a* **Bernoulli random variable with parameter** *p if it takes values in $W = \{0, 1\}$ and*

$$\mathbb{P}[X = 0] = 1 - p \quad and \quad \mathbb{P}[X = 1] = p.$$

In this case, we write $X \sim \mathrm{Ber}(p)$.

Binomial distribution

Another fundamental example is the binomial distribution, which appears in applications when we consider the number of successes in a repetition of Bernoulli experiments.

Definition 3.11 (Binomial)**.** *Let $0 \le p \le 1$, let $n \in \mathbb{N}$. A random variable X is said to be a* **binomial random variable with parameters** *n* **and** *p if it takes values in $W = \{0, \ldots, n\}$ and*

$$\forall k \in \{0, \ldots, n\} \quad \mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

In this case, we write $X \sim \mathrm{Bin}(n, p)$.

Remark 3.12. *If we define $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$, we have*

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

hence the equation (3.1) is satisfied. This guarantees the existence of binomial random variables.

Proposition 3.13 (Sum of independent Bernoulli and binomial)**.** *Let $0 \le p \le 1$, Let $n \in \mathbb{N}$. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Then*

$$S_n := X_1 + \cdots + X_n$$

is a binomial random variable with parameter n and p.

Proof. One can easily check that S_n is a random variable which takes values in $\{0, \ldots, n\}$. Furthermore, for every $k \in \{0, \ldots, n\}$ we have

$$\{S_n = k\} = \bigcup_{\substack{x_1, \ldots, x_n \in \{0,1\} \\ x_1 + \cdots + x_n = k}} \{X_1 = x_1, \cdots, X_n = x_n\}.$$

Since the union is disjoint, we get

$$
\begin{aligned}
\mathbb{P}[S_n = k] &= \sum_{\substack{x_1,\ldots,x_n \in \{0,1\} \\ x_1 + \cdots + x_n = k}} \mathbb{P}[X_1 = x_1, \cdots, X_n = x_n] \\
&= \sum_{\substack{x_1,\ldots,x_n \in \{0,1\} \\ x_1 + \cdots + x_n = k}} \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n] \\
&= \sum_{\substack{x_1,\ldots,x_n \in \{0,1\} \\ x_1 + \cdots + x_n = k}} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}
$$

$\square$

Remark 3.14. *In particular, the distribution* $\mathrm{Bin}(1,p)$ *is the same as the distribution* $\mathrm{Ber}(p)$. *One can also check that if* $X \sim \mathrm{Bin}(m,p)$ *and* $Y \sim \mathrm{Bin}(n,p)$ *and* X,Y *are independent, then* $X + Y \sim \mathrm{Bin}(m+n,p)$.

Geometric distribution

Definition 3.15 (Geometric). *Let* $0 < p \le 1$. *A random variable* X *is said to be a* ***geometric random variable with parameter*** p *if it takes values in* $W = \mathbb{N}\backslash\{0\}$ *and*

$$
\boxed{\forall k \in \mathbb{N}\backslash\{0\} \quad \mathbb{P}[X = k] = (1-p)^{k-1} \cdot p}.
$$

In this case, we write $X \sim \mathrm{Geom}(p)$.

Remark 3.16. *For* $p = 1$, *and* $k = 1$, *a term* 0^0 *appears in the equation above, we use the convention* $0^0 = 1$ *and therefore* $\mathbb{P}[X = 1] = 1$ *in this case.*

Remark 3.17. *If we define* $p(k) = (1-p)^{k-1} p$, *we have*

$$
\sum_{k=1}^{\infty} p(k) = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p \cdot \frac{1}{p} = 1,
$$

hence the equation (3.1) *is satisfied. This guarantees the existence of geometric random variables.*

The geometric random variable appears naturally as the first success in an infinite sequence of Bernoulli experiments with parameter p. This is formalized by the following proposition.

Proposition 3.18. *Let* $X_1, X_2, \ldots$ *be a sequence of infinitely many independent Bernoulli r.v.'s with parameter* p. *Then*

$$
T := \min\{n \ge 1 : X_n = 1\}
$$

is a geometric random variable with parameter p.

Remark 3.19. *When saying that T is a geometric random variable, we make a slight abuse: indeed, the random variable T may take the value $+\infty$ if all the random variables X_i's are equal to 0. Nevertheless, this is not a problem for the calculations because one can check that $\mathbb{P}[T = \infty] = 0$.*

Proof. We have $T = k$ if the first $k - 1$ trials fail, and the k's one is a success. Formally, we have

$$\{T = k\} = \{X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1\}.$$

Hence, by independence,

$$\begin{aligned}
\mathbb{P}[T = k] &= \mathbb{P}[X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1] \\
&= \mathbb{P}[X_1 = 0]\cdots\mathbb{P}[X_{k-1} = 0]\,\mathbb{P}[X_k = 1] \\
&= (1 - p)^{k-1}p.
\end{aligned}$$

$\square$

The previous proposition gives us an easy way to remember the definition of the geometric r.v., and also some simple formulas related to the geometric distribution. For example, if T is a geometric distribution with parameter p, we have $T > n$ if the n first Bernoulli experiments fail, and therefore

$$\mathbb{P}[T > n] = (1 - p)^n. \tag{3.3}$$

Also, it gives an important interpretation to the equation (3.4) in the proposition below: if we are waiting for a first success in a sequence of experiments, and if we know that the first n steps were a failure, then the remaining time to wait is again a geometric random variable with parameter p.

Proposition 3.20 (Absence of memory of the geometric distribution)**.** *Let $T \sim Geom(p)$ for some $0 < p < 1$. Then*

$$\forall n \geq 0 \quad \forall k \geq 1 \qquad \mathbb{P}[T \geq n + k \,|\, T > n] = \mathbb{P}[T \geq k]. \tag{3.4}$$

Proof. It follows directly from the formula (3.3). $\square$

Poisson distribution

Definition 3.21. *Let $\lambda > 0$ be a positive real number. A random variable X is said to be a **Poisson random variable with parameter** λ if it takes values in $W = \mathbb{N}$ and*

$$\boxed{\forall k \in \mathbb{N} \quad \mathbb{P}[X = k] = \frac{\lambda^k}{k!}e^{-\lambda}.}$$

In this case, we write $X \sim \text{Poisson}(\lambda)$.

Remark 3.22. *If we define* $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, *we have*

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

hence the equation (3.1) *is satisfied. This guarantees the existence of Poisson random variables.*

The Poisson distribution appears naturally as an approximation of a binomial distribution when the parameter n is large and the parameter p is small, as stated formally in the following proposition.

Proposition 3.23 (Poisson approximation of the binomial)**.** *Let* $\lambda > 0$. *For every* $n \geq 1$, *consider a random variable* $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$. *Then*

$$\forall k \in \mathbb{N} \quad \lim_{n \to \infty} \mathbb{P}[X_n = k] = \mathbb{P}[N = k], \tag{3.5}$$

where N *is a Poisson random variable with parameter* λ.

Remark 3.24. *The convergence* (3.5) *is called a convergence in distribution. Intuitively, it says that* X_n *and* N *have very similar probabilistic properties for* n *large.*

Proof. Fix $k \in \mathbb{N}$. For every $n \geq 1$, we have

$$\mathbb{P}[X_n = k] = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}}_{\xrightarrow[n \to \infty]{} 1} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\xrightarrow[n \to \infty]{} e^{-\lambda}},$$

which concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

This approximation may be useful in practice. For example, consider a single page of the "Neue Zürcher Zeitung" containing, say, $n = 10^4$ characters, and suppose that the typesetter mis-sets approximately $1/1000$ of the characters. In other words, each character has a probability $p = 10/n$ to be mis-set. The number M of mistakes in the page corresponds to a binomial random variable with parameters n and $p = 10/n$. Hence by the Poisson approximation, for example we have

$$\mathbb{P}[M = 5] \simeq \frac{10^5}{5!} e^{-10} \simeq 0,0378.$$

5   Continuous random variables

Definition 3.25 (Continuous random variables)**.** *A random variable* $X : \Omega \to \mathbb{R}$ *is said to be* **continuous** *if its distribution function* F_X *can be written as*

$$F_X(a) = \int_{-\infty}^{a} f(x)dx \quad \text{for all } a \text{ in } \mathbb{R} \tag{3.6}$$

for some nonnegative function $f : \mathbb{R} \to \mathbb{R}_+$*, called the* **density** *of* X*.*

Intuition: $f(x)dx$ represent the probability that X takes a value in the infinitesimal interval $[x, x + dx]$.

To understand the terminology "continuous", observe that the formula (3.6) implies that F_X is a continuous function. In particular, by Proposition 3.1, the r.v. X satisfies

$$\forall x \in \mathbb{R} \quad \mathbb{P}[X = x] = 0.$$

Proposition 3.26. *The density* f *of a random variable satisfies*

$$\boxed{\int_{-\infty}^{+\infty} f(x)dx = 1}.$$

Proof. We have

$$\int_{-\infty}^{+\infty} f(x)dx = \lim_{y \to \infty} \int_{-\infty}^{y} f(x)dx = \lim_{y \to \infty} F_X(y) = 1.$$

$\square$

Conversely, if we are given a nonnegative function $f : \mathbb{R} \to \mathbb{R}^+$ such that

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X with associated density f. This is a consequence of the existence theorem 2.15 in Chapter 2.

Density f vs distribution function F_X

From f to F_X   Let X be a continuous random variable X with density f. By definition, the distribution function F_X can be calculated as the integral

$$F_X(x) = \int_{-\infty}^{x} f(y)dy.$$

From F_X to f Since one goes from f to F_X by integrating, it is natural to expect that the reverse operation is to take a derivative. This is in general the case, provided F_X is regular enough.

The following theorem will be useful in applications to calculate densities.

Theorem 3.27. *Let X be a random variable. Assume the distribution function F_X is continuous and piecewise $\mathcal{C}^1$, i.e. that there exist $x_0 = -\infty < x_1 < \cdots < x_{n-1} < x_n = +\infty$ such that F_X is $\mathcal{C}^1$ on every interval (x_i, x_{i+1}). Then X is a continuous random variable and a density f can be constructed by defining*

$$\forall x \in (x_i, x_{i+1}) \quad f(x) = F_X'(x)$$

and setting arbitrary values at $x_1, \ldots, x_{n-1}$.

Proof. We simply write $F = F_X$. Let $0 \le i < n$. If $x_i < a < b < x_{i+1}$, the fundamental theorem of calculus implies that

$$F(b) - F(a) = \int_a^b F'(y)dy = \int_a^b f(y)dy$$

Now, let $x \in \mathbb{R}$ and let i be such that $x \in [x_i, x_{i+1})$. Using the convention $F(x_0) = 0$, we can write $F(x)$ as a telescopic sum

$$F(x) = F(x) - F(x_0) = (F(x) - F(x_i)) + \cdots + (F(x_1) - F(x_0)). \tag{3.7}$$

By continuity of F, we have

$$(F(x) - F(x_i)) = \lim_{a\downarrow x_i}(F(x) - F(a)) = \lim_{a\downarrow x_i}\int_a^x f(y)dy = \int_{x_i}^x f(y)dy,$$

and equivalently

$$F(x_i) - F(x_{i-1}) = \int_{x_{i-1}}^{x_i} f(y)dy$$

Plugging these identities in (3.7), we get

$$F(x) = \int_{x_i}^x f(y)dy + \int_{x_{i-1}}^{x_i} f(y)dy + \cdots + \int_{x_0}^{x_1} f(y)dy$$
$$= \int_{-\infty}^x f(y)dy.$$

$\square$

6 Examples of continuous random variables

Uniform distributions

Definition 3.28 (Uniform distribution in $[a, b]$, $a < b$.)**.** *A continuous random variable X is said be* **uniform in** $[a, b]$ *if its density is equal to*

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b], \\ 0 & x \notin [a, b]. \end{cases}$$
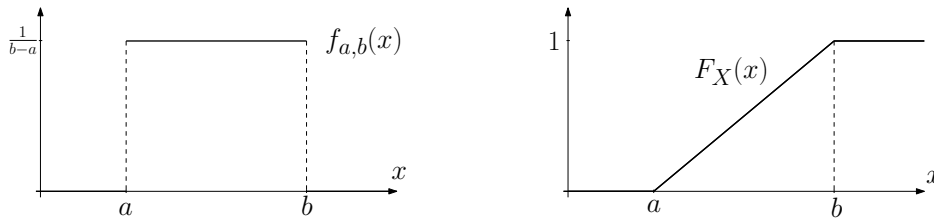
In this case, we write $X \sim \mathcal{U}([a, b])$.

Figure 3.1: Density and distribution function of a uniform random variable in $[a, b]$.

Intuition: X represents a uniformly chosen point in $[a, b]$.

Properties of a uniform random variable X in $[a, b]$:

- The probability to fall in a an interval $[c, c + \ell] \subset [a, b]$ depends only on its length ℓ:

$$\mathbb{P}[X \in [c, c + \ell]] = \frac{\ell}{b - a}.$$

- The distribution function of X is equal to

$$F_X(x) = \begin{cases} 0 & x < a, \\ \frac{x-a}{b-a} & a \le x \le b, \\ 1 & x > b. \end{cases}$$

Proof.   □

Exponential distribution

The exponential distribution is the continuous analogue of the geometric distribution.

Definition 3.29 (Exponential distribution with $\lambda > 0$). *A continuous random variable T is said be* **exponential with parameter** $\lambda > 0$ *if its density is equal to*

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$
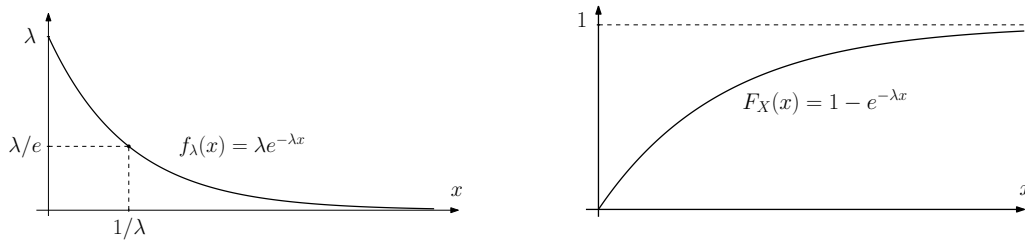
In this case, we write $T \sim Exp(\lambda)$.

Figure 3.2: Density and distribution function of an exponential random variable with parameter λ.

Intuition/application: T represents the time of a "clock ring". For example, the time at which a first customer arrives in a shop is well modeled by an exponential random variable.

Properties of an exponential random variable T with parameter λ.

- The waiting probability is exponentially small:

$$\forall t \geq 0 \quad \boxed{\mathbb{P}[T > t] = e^{-\lambda t}}.$$

- It has the absence of memory property:

$$\forall t, s \geq 0 \quad \mathbb{P}[T > t + s | T > t] = [T > s].$$

The first item follows from the definition:

$$\mathbb{P}[T \geq t] = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}.$$

The second item is a direct computation of the conditional probability:

$$\mathbb{P}[T > t + s | T > t] = \frac{\mathbb{P}[T > t + s]}{\mathbb{P}[T > t]} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s}.$$

Normal distribution

Definition 3.30. *A continuous random variable X is said be **normal with parameters** m **and** $\sigma^2 > 0$ if its density is equal to*

$$f_{m,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$
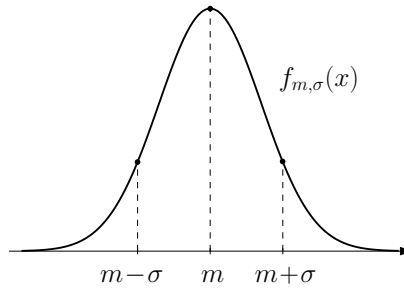
In this case, we write $X \sim \mathcal{N}(m, \sigma^2)$.

Figure 3.3: Density of a normal random variable with parameters m and σ^2.

Intuition/application: The normal distribution arises in many applications. For example, imagine that we measure a physical quantity: the real value is m and the measured value is in general very well modeled by a normal random variable X with parameters m and σ. The quantity σ, which represents the fluctuations for X can also be interpreted as the quality of the measurement in this context. A small σ corresponds to small fluctuations of X, which means that X is typically closed to m. In contrary, a large σ corresponds to large fluctuations and can be interpreted as inaccurate measurement. We will see later a mathematical justification explaining why the normal random variable appears in many places.

Properties of normal random variables

- If $X_1, \ldots, X_n$ are independent random variables with parameters $(m_1, \sigma_1^2), \ldots, (m_n, \sigma_n^2)$ respectively, then
  $$Z = m_0 + \lambda_1 X_1 + \ldots + \lambda_n X_n$$
  is a normal random variable with parameters $m = m_0 + \lambda_1 m_1 + \cdots + \lambda_n m_n$ and $\sigma^2 = \lambda_1^2 \sigma_1^2 + \cdots + \lambda_n^2 \sigma_n^2$.

- In particular, if $X \sim \mathcal{N}(0, 1)$ (in this case we say that X is a **standard normal random variable**), then
  $$Z = m + \sigma \cdot X$$
  is a normal random variable with parameters m and σ^2.

- If X is a normal random variable with parameters m and σ^2, then all the "probability mass" is mainly in the interval $[m - 3\sigma, m + 3\sigma]$. Namely, we have

$$\mathbb{P}[|X - m| \geq 3\sigma] \leq 0.0027.$$

At a first look, it may be surprising that the right hand side (0.0027) does not depend on the parameters σ and m... This is explained as follows: consider the random variable $Z = \frac{X-m}{\sigma}$. The first property above implies that $Z = \frac{1}{\sigma}X - \frac{m}{\sigma}$ is a standard normal random variable. The left hand side can be rewritten as

$$\mathbb{P}[|X - m| \geq 3\sigma] = \mathbb{P}[|\frac{X - m}{\sigma}| \geq 3] = \mathbb{P}[|Z| \geq 3]$$

Then the inequality $\mathbb{P}[|Z| \geq 3] \leq 0.0027$ can be directly checked from a table.

Chapter 4

Expectation

Goals

- Definition of the expectation and intuition.

- Rules of calculus (sum/product of random variables).

- Inequalities, relations between expectation and probability of events

- Definition of the variance and intuition.

Framework   We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All the random variables considered in this chapter will be defined on this reference probability space.

In this class, we focus on discrete and continuous random variables, and we will give the expectation for these two different types of random variables by two different formulas. There is a unified theory (based on measure theory) of expectation, that defines the expectation for general random variables. In this class, we will keep the focus on the important results from this theory and their applications, without giving the proofs.

1 Expectation for general random variables

Definition 4.1. *Let $X : \Omega \to \mathbb{R}_+$ be a random variable with nonnegative values. The expectation of X is defined as*

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))dx. \tag{4.1}$$

Remark 4.2. *The expectation may be finite or infinite.*

Proposition 4.3. *Let X be a nonnegative random variable. Then we have*

$$\mathbb{E}[X] \geq 0,$$

with equality if and only if $X = 0$ almost surely.

Proof. The expectation $\mathbb{E}[X]$ is defined as the integral of the nonnegative function $G(x) = 1 - F_X(x) \geq 0$. Hence $\mathbb{E}[X] \geq 0$. Now, assume that $\mathbb{E}[X] = 0$. This implies that $G(x) = 0$ for every $x > 0$ (by contradiction, if $G(x) = \alpha > 0$ for some $x > 0$, then $G(y) \geq G(x) = \alpha$ for all $y \in [0, x]$ by monotonicity, which implies that $\int_0^x G(y)dy \geq x\alpha > 0$). By continuity of probability measures, we have

$$\mathbb{P}[X > 0] = \lim_{x \downarrow 0} \mathbb{P}[X > x] = \lim_{x \downarrow 0} G(x) = 0.$$

Therefore,

$$\mathbb{P}[X \leq 0] = 1 - \mathbb{P}[X > 0] = 1.$$

Hence $X \geq 0$ and $X \leq 0$ almost surely, which implies that $X = 0$ almost surely. $\qquad \square$

For general random variables (not necessarily with a constant sign), we define the expectation by decomposing into positive and negative parts. The positive and negative parts of X are the random variables X_-, X_+ defined by

$$X_+(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0, \\ 0 & \text{if } X(\omega) < 0, \end{cases} \quad \text{and} \quad X_-(\omega) = \begin{cases} -X(\omega) & \text{if } X(\omega) \leq 0, \\ 0 & \text{if } X(\omega) > 0. \end{cases}$$

Notice that both X_+ and X_- take nonnegative values. Furthermore, we have $X = X_+ - X_-$, and $|X| = X_+ + X_-$.

Definition 4.4. *Let X be a random variable. If $\mathbb{E}[|X|] < \infty$, then the expectation of X is defined by*

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]. \tag{4.2}$$

Remark 4.5. *The condition* $\mathbb{E}[|X|] < \infty$ *ensures that* $\mathbb{E}[X_-], \mathbb{E}[X_+] < \infty$ *(because* $|X| = X_+ + X_-$*), and therefore the difference in Eq. (4.2) makes sense.*

If $X \geq 0$, the expectation of X is always defined. It may be finite or infinite.

If X does not have a constant sign, the expectation of X is well defined if $\mathbb{E}[|X|] < \infty$. When this condition is not satisfied, we say that the expectation of X is **undefined**.

2   Expectation of a discrete random variable

> **Proposition 4.6.** *Let* $X : \Omega \to \mathbb{R}$ *be a discrete random variable with values in* W *(finite or countable) almost surely. We have*
>
> $$\boxed{\mathbb{E}[X] = \sum_{x \in W} x \cdot \mathbb{P}[X = x]},$$
>
> *provided the sum is well defined.*

Proof. We first assume that $X \geq 0$ almost surely. By Propositions 3.7 and 3.9 we have for every $x \in \mathbb{R}$

$$1 - F_X(x) = \sum_{\substack{y > x \\ y \in W}} p(y) = \sum_{y \in W} \mathbf{1}_{y > x} \cdot p(y).$$

By using this identity in the definition of the expectation, we get

$$\mathbb{E}[X] = \int_0^\infty \Big(\sum_{y \in W} \mathbf{1}_{y > x} \cdot p(y) \Big) dx = \sum_{y \in W} \Big(\int_0^\infty \mathbf{1}_{y > x} dx \Big) \cdot p(y) = \sum_{y \in W} y \cdot p(y).$$

Now if X is not of constant sign, we use the decomposition $X = X_+ - X_-$ and we apply the formula above to X_+ and X_-. We obtain

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-] = \sum_{y \in W} y \mathbb{P}[X_+ = y] - \sum_{y \in W} y \mathbb{P}[X_- = y].$$

The definitions of X_+ and X_- imply that $\{X = y\}$ is equal to the disjoint union $\{X_+ = y\} \cup \{X_- = -y\}$ $\square$

Example 1: *Bernoulli r.v.*

Let X be a Bernoulli random variable with parameter p. We have

$$\boxed{\mathbb{E}[X] = p}.$$

Indeed,

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

Example 2: *Bet on a die*

Consider the random variable $X : \Omega \to \{-1, 0, +2\}$ defined by Eq. (2.1) page 21. Then

$$\mathbb{E}[X] = -1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{6} + 2 \cdot \frac{1}{3} = \frac{1}{6}.$$

Example 3: *Poisson r.v.*

Let X be a Poisson random variable with parameter $\lambda > 0$, then

$$\boxed{\mathbb{E}[X] = \lambda}.$$

Indeed,

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = \lambda \cdot \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \cdot e^{-\lambda} = \lambda.$$

Example 4: *Indicator of an event*

Let $A \in \mathcal{F}$ be an event. Consider the **indicator function** $\mathbb{1}_A$ of A, defined by

$$\forall \omega \in \Omega \quad \mathbb{1}_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A, \\ 1 & \text{if } \omega \in A. \end{cases}$$

Then $\mathbb{1}_A$ is a random variable. Indeed, we have

$$\{\mathbb{1}_A \leq a\} = \begin{cases} \varnothing & \text{if } a < 0, \\ A^c & \text{if } 0 \leq a < 1, \\ \Omega & \text{if } a \geq 1, \end{cases}$$

and $\varnothing$, A^c and Ω are three elements of $\mathcal{F}$. Furthermore, writing $X = 1_A$, we have

$$\mathbb{P}[X = 0] = 1 - \mathbb{P}[A] \quad \text{and} \quad \mathbb{P}[X = 1] = \mathbb{P}[A].$$

Therefore $\mathbb{1}_A$ is a Bernoulli r.v. with parameter $\mathbb{P}[A]$. Hence,

$$\boxed{\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A]}.$$

Proposition 4.7. *Let $X : \Omega \to \mathbb{R}$ be a discrete random variable with values in W (finite or countable) almost surely. For every $\phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\boxed{\mathbb{E}[\phi(X)] = \sum_{x \in W} \phi(x) \cdot \mathbb{P}[X = x]},$$

provided the sum is well defined.

Proof. Admitted. $\square$

3 Expectation of a continuous random variable

Proposition 4.8. *Let X be a continuous random variable with density f. Then, we have*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx, \tag{4.3}$$

provided the integral is well defined.

Proof. We assume that $X \geq 0$ almost surely. The general case (without sign constraint) can be deduced from the positive case by using the decomposition $X = X_+ - X_-$ (similarly as in the proof of Proposition 4.6). By definition of the density and Proposition 3.26, for every $x \in \mathbb{R}$ we have

$$1 - F_X(x) = \int_x^{+\infty} f(y)dy = \int_{-\infty}^{+\infty} \mathbf{1}_{x<y} \cdot f(y)dy.$$

By using this identity in the definition of the expectation, we get

$$\mathbb{E}[X] = \int_0^{\infty} \left(\int_{-\infty}^{+\infty} \mathbf{1}_{x<y} \cdot f(y)dy \right)dx = \int_{-\infty}^{+\infty} \left(\int_0^{\infty} \mathbf{1}_{y>x} \cdot dx \right) \cdot f(y)dy = \int_{-\infty}^{\infty} y \cdot f(y)dy.$$

$\square$

Example 1: *Uniform random variable in $[a,b]$, $a < b$*

We have

$$\mathbb{E}[X] = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \cdot \left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right).$$

Therefore,

$$\boxed{\mathbb{E}[X] = \frac{a+b}{2}.}$$

Example 2: *Exponential random variable with parameter $\lambda > 0$*

By integration by parts, we have

$$\mathbb{E}[X] = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \left[-xe^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx.$$

Therefore,

$$\boxed{\mathbb{E}[X] = \frac{1}{\lambda}.}$$

Proposition 4.9. *Let X be a continuous random variable with density f. Let $\phi : \mathbb{R} \to \mathbb{R}$ be such that $\phi(X)$ is a random variable. Then, we have*

$$\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)f(x)dx, \tag{4.4}$$

provided the integral is well defined.

4 Calculus

One of the reasons why the expectation is a such a powerful tool in probability theory is that we can do calculations: for example, one can calculate the expectation of $X + Y$ if one knows the expectations of X and Y. In this section we give the rules of calculus for the basic operations on random variables.

Linearity

Theorem 4.10 (Linearity of the expectation)**.** *Let $X, Y : \Omega \to \mathbb{R}$ be random variables, let $\lambda \in \mathbb{R}$. Provided the expectations are well defined, we have*

1. $\boxed{\mathbb{E}[\lambda \cdot X] = \lambda \cdot \mathbb{E}[X]}$.

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

Remark 4.11. *The random variables X and Y do not need to be independent.*

Remark 4.12. *More generally, by induction we have: for every integer $n \geq 1$*

$$\mathbb{E}[\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n] = \lambda_1 \mathbb{E}[X_1] + \lambda_2 \mathbb{E}[X_2] + \cdots + \lambda_n \mathbb{E}[X_n],$$

for any n random variables $X_1, X_2, \ldots, X_n : \Omega \to E$, and any $\lambda_1, \lambda_2, \cdots, \lambda_n \in \mathbb{R}$, provided the expectations are well defined.

Proof. The proof of Item (i) follows from the definition. The proof of Item (ii) for general random variables belongs to the abstract framework of measure theory and we admit it. We give here the proof for a finite sample space, which illustrates well the key idea behind the linearity property. Let us assume that Ω is finite, and is equipped with the sigma-algebra $\mathcal{F} = \mathcal{P}(\Omega)$. In this case the two random variables X and Y are necessarily discrete (see Remark 3.5). For every $x \in X$,

$$\mathbb{P}[X = x] = \mathbb{P}[\bigcup_{\omega \in \Omega : X(\omega) = x} \{\omega\}] = \sum_{\omega \in \Omega} \mathbf{1}_{X(\omega) = x} \cdot \mathbb{P}[\omega],$$

where we make the abuse of notation $\mathbb{P}[\omega] = \mathbb{P}[\{\omega\}]$. Therefore,

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x \cdot \mathbb{P}[X = x] = \sum_{x \in X(\Omega)} x \cdot \left(\sum_{\omega \in \Omega} \mathbf{1}_{X(\omega) = x} \cdot \mathbb{P}[\omega] \right)$$

$$= \sum_{x \in X(\Omega)} \sum_{\omega \in \Omega} x \cdot \mathbf{1}_{X(\omega) = x} \cdot \mathbb{P}[\omega]$$

$$= \sum_{\omega \in \Omega} \mathbb{P}[\omega] \cdot \underbrace{\sum_{x \in X(\Omega)} x \cdot \mathbf{1}_{X(\omega) = x}}_{= X(\omega)}$$

$$= \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\omega].$$

Using this formula for $X + Y$ and Y, we obtain

$$\mathbb{E}[X + Y] = \sum_{\omega \in \Omega} \big(X(\omega) + Y(\omega) \big) \cdot \mathbb{P}[\omega] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}[\omega] + \sum_{\omega \in \Omega} Y(\omega) \cdot \mathbb{P}[\omega] = \mathbb{E}[X] + \mathbb{E}[Y].$$

$\square$

Application 1: *Expectation of a Binomial random variable.*

Let $n \geq 1$ and $0 \leq p \leq 1$. Let S be a binomial random variable with parameters n and p. What is the expectation of S?

By definition we have

$$\mathbb{E}[S] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1 - p)^{n-k}$$

and this sum does not look so nice... However, we can use that S has the same distribution as $S_n = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are n i.i.d. Bernoulli random variables with parameter p. By linearity we have

$$\mathbb{E}[S_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].$$

Using that $\mathbb{E}[X_i] = p$ for every i, we deduce directly

$$\boxed{\mathbb{E}[S] = \mathbb{E}[S_n] = np.}$$

Application 2: *Expectation of a normal random variable*

If X is a normal distribution with parameters m and σ^2, then it has the same distribution as $m + \sigma \cdot Y$ where Y is a standard normal random variable. By Proposition 4.9, we have

$$\mathbb{E}[X] = \mathbb{E}[m + \sigma \cdot Y] = m + \sigma \mathbb{E}[Y],$$

hence it suffices to compute the expectation of Y. Writing $f_{0,1}$ for the density of Y, we have

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} x \cdot f_{0,1}(x) dx = 0$$

because $x \cdot f_{0,1}(x)$ is an odd function. Finally, we obtain

$$\boxed{\mathbb{E}[X] = m.}$$

> **Theorem 4.13.** *Let X, Y be two random variables. If <u>X and Y are independent</u>,*
> *then*
> $$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Proof. Admitted. $\square$

5   Tailsum formulas

Proposition 4.14 (Tailsum formula for nonnegative random variables)**.** *Let X be a random variable, such that $X \geq 0$ almost surely. Then, we have*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]dx.$$

Proof. It follows form the definition of the expectation (Eq. (4.1)) and the identity

$$1 - F_X(x) = 1 - \mathbb{P}[X \leq x] = \mathbb{P}[X > x].$$

$\square$

Application: Alternative computation of the expectation of an exponential random variable.

We proved in the previous paragraph that the expectation of an exponential random variable T with parameter $\lambda \geq 0$ is equal to $1/\lambda$. We here give an alternative derivation of this result. We have $T \geq 0$ almost surely, and $\mathbb{P}[T > x] = e^{-\lambda x}$. Hence, the proposition above gives

$$\mathbb{E}[T] = \int_0^\infty e^{-\lambda x}dx = \frac{1}{\lambda}.$$

Proposition 4.15 (Tailsum formula for discrete random variables)**.** *Let X be a discrete r.v. <u>taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$</u>. Then*

$$\boxed{\mathbb{E}[X] = \sum_{n=1}^\infty \mathbb{P}[X \geq n].}$$

Proof. Since $X \geq 0$ almost surely, we can apply proposition 4.14 to write

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]dx = \sum_{n=1}^\infty \int_{n-1}^n \mathbb{P}[X > x]dx.$$

Now, for $x \in [n-1, n)$ we have $\mathbb{P}[X > x] = \mathbb{P}[X \geq n]$. Therefore,

$$\int_{n-1}^n \mathbb{P}[X > x]dx = \mathbb{P}[X \geq n] \int_{n-1}^n dx = \mathbb{P}[X \geq n].$$

$\square$

Application: Computation of the expectation of a geometric random variable.

Let T be a geometric random variable with parameter $0 < p \leq 1$. Then

$$\boxed{\mathbb{E}[T] = \frac{1}{p}.}$$

Indeed, T takes values in $\mathbb{N}$ and the proposition above gives

$$\mathbb{E}[T] = \sum_{n \geq 1} \mathbb{P}[T \geq n] = \sum_{n \geq 1} (1 - p)^{n-1} = \frac{1}{1 - (1 - p)} = \frac{1}{p}.$$

6 Characterizations via expectations

Density

If a random variable X has a density f, we can calculate the expectation of X via the formula (4.4). Notice that two random variables with different densities may have the same expectation. For example a uniform random variable in $[-1, 1]$ and a Gaussian random variable with parameters $m = 0$ and $\sigma^2 > 0$ have the same expectation, but different densities. In other words, the expectation of a random variable does not characterize the density.

Nevertheless it is possible to characterize the density of a random variable X, by considering all the expectations of images $\phi(X)$ for a sufficiently large class of functions ϕ. This is the content of the proposition below.

For this course we consider functions ϕ that are piecewise continuous, bounded functions. Recall that a function $\phi : \mathbb{R} \to \mathbb{R}$ is **piecewise continuous** if there exists $a_1 < a_2 < \cdots < a_n$ such that ϕ is continuous on each interval (a_i, a_{i+1}). It is **bounded** if there exists $C > 0$ such that

$$\forall x \in \mathbb{R} \quad |\phi(x)| \leq C.$$

Proposition 4.16. *Let X be a random variable. Let $f : \mathbb{R} \to \mathbb{R}_+$ such that $\int_{-\infty}^{+\infty} f(x)dx = 1$. Then the following are equivalent:*

(i) X is continuous with density f,

(ii) For every function $\phi : \mathbb{R} \to \mathbb{R}$ piecewise continuous, bounded,

$$\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)f(x)dx. \tag{4.5}$$

Proof.

$\boxed{(i) \Rightarrow (ii)}$ It follows from Prop. 4.9.

$\boxed{(ii) \Rightarrow (i)}$ Let $a \in \mathbb{R}$, and consider the function ϕ_a defined by $\phi_a(x) = \mathbf{1}_{x \leq a}$.

By applying Eq. (4.5), we find

$$\mathbb{E}[\mathbf{1}_{X \leq a}] = \mathbb{E}[\phi_a(X)] = \int_{-\infty}^{\infty} \phi_a(x) f(x) dx = \int_{-\infty}^{a} f(x) dx.$$

By applying the identity $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A]$ to the event $A = \{X \leq a\}$, we finally get

$$\mathbb{P}[X \leq a] = \int_{-\infty}^{a} f(x) dx,$$

which concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ □

Independence

If two random variables X and Y are independent, then we have seen in Theorem 6.1 that

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{4.6}$$

Conversely, the formula above does **not** imply that X and Y are independent. See for example Exercise 6.6. Nevertheless, if we consider a stronger form of Eq. (4.6), allowing to take arbitrary images of X and Y we obtain a characterization of independence, as stated below.

Theorem 4.17. *Let X, Y be 2 discrete random variables. Then the following are equivalent*

(i) X, Y are independent,

(ii) For every $\phi : \mathbb{R} \to \mathbb{R}, \psi : \mathbb{R} \to \mathbb{R}$ piecewise continuous, bounded,

$$\boxed{\mathbb{E}[\phi(X)\psi(Y)] = \mathbb{E}[\phi(X)]\mathbb{E}[\psi(Y)]}. \tag{4.7}$$

Proof.

(i)$\Rightarrow$ (ii) Admitted.

(ii)$\Rightarrow$ (i) Let $a, b \in \mathbb{R}$. By applying Eq. (4.7) to the two function defined by $\phi_a(x) = \mathbf{1}_{x \leq a}$ and $\psi_b(y) = \mathbf{1}_{y \leq b}$, we get

$$\mathbb{E}[\mathbf{1}_{X \leq a, Y \leq b}] = \mathbb{E}[\mathbf{1}_{X \leq a}]\mathbb{E}[\mathbf{1}_{X \leq b}].$$

Using $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A]$, this concludes that X and Y are independent.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ □

Above, we only considered a pair (X, Y) of random variables, but the same ideas apply to n random variables $X_1, \ldots, X_n$, as in the following theorem.

Theorem 4.18. *Let $X_1, \ldots, X_n$ be n random variables. Then the following are equivalent*

(i) *$X_1, \ldots, X_n$ are independent,*

(ii) *For every $\phi_1 : \mathbb{R} \to \mathbb{R}, \ldots, \phi_n : \mathbb{R} \to \mathbb{R}$ piecewise continuous, bounded*

$$\mathbb{E}[\phi_1(X_1)\cdots\phi_n(X_n)] = \mathbb{E}[\phi_1(X_1)]\cdots\mathbb{E}[\phi_n(X_n)].$$

7 Inequalities

Monotonicity

Proposition 4.19. *Let X, Y be two random variables such that*

$$X \leq Y \, a.s.$$

Then

$$\mathbb{E}[X] \leq \mathbb{E}[Y],$$

provided the two expectations are well defined.

Proof. Consider the random variable $Z = Y - X$. By hypothesis, we have $Z \geq 0$, which implies that $\mathbb{E}[Z] \geq 0$ (by Prop. 4.3). By linearity, we have

$$\mathbb{E}[Y] - \mathbb{E}[X] = \mathbb{E}[Z] \geq 0.$$

$\square$

Markov's inequality

Theorem 4.20 (Markov's inequality)**.** *Let X be a **nonnegative** random variable. Then for every $a > 0$, we have*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{4.8}$$

Proof. Let $a > 0$. By monotonicity of the expectation, we have

$$\mathbb{E}[X] \geq \mathbb{E}[X \cdot \mathbb{1}_{X \geq a}] \geq \mathbb{E}[a \cdot \mathbb{1}_{X \geq a}] = a\mathbb{P}[X \geq a].$$

$\square$

Jensen's inequality

Theorem 4.21 (Jensen's inequality)**.** *Let X be a random variable. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. If $\mathbb{E}[\phi(X)]$ and $\mathbb{E}[X]$ are well defined, then*

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

The Jensen inequality has several important consequences. First, by applying it to $\phi(x) = |x|$, we obtain the triangle inequality. For every integrable discrete random variable X, we have

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

Another important consequence relates the average of $|X|$ with the average of X^2. By applying it to the convex function $\phi(x) = x^2$, we obtain that for every discrete random variable, we have

$$\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}. \tag{4.9}$$

8   Variance

Definition 4.22. *Let X be a variable such that $\mathbb{E}[X^2] < \infty$. The **variance of** X is defined by*

$$\boxed{\sigma_X^2 = \mathbb{E}[(X - m)^2], \quad where\ m = \mathbb{E}[X].}$$

*The square root σ_X of the variance is called the **standard deviation of** X.*

Remark 4.23. *If $\mathbb{E}[X^2] < \infty$, then we also have $\mathbb{E}[|X|] < \infty$ by Eq. (4.9) and therefore the average $m = \mathbb{E}[X]$ is well-defined.*

The standard deviation is an indicator of how large the fluctuations of X around $m = \mathbb{E}[X]$ are. We illustrate this fact on two simple examples.

Example 1: *Deterministic random variable*

Let $a \in \mathbb{R}$. Consider the random variable defined by $X(\omega) = a$ for every ω. Then $m = \mathbb{E}[X] = a$ and $\sigma_X^2 = \mathbb{E}[(X - m)^2] = 0$.

Example 2: *Uniform random variable on two points*

Let $a < b$ be two real numbers. Consider a random variable X with distribution given by $\mathbb{P}[X = a] = \mathbb{P}[X = b] = 1/2$. Then $m = \mathbb{E}[X] = (a + b)/2$ and

$$\sigma_X = \sqrt{\mathbb{E}[(X - m)^2]} = \frac{a - b}{2}.$$

In general, a random variable X with a small variance is well concentrated on values close to its expectation $m = \mathbb{E}[X]$. This concentration phenomena can be quantified using the Chebyshev's inequality.

Theorem 4.24. *Let X be a random variable such that $\mathbb{E}[X^2] < \infty$. Then for every $a \geq 0$ we have*

$$\mathbb{P}[|X - m| \geq a] \leq \frac{\sigma_X^2}{a^2}, \qquad \text{where } m = \mathbb{E}[X].$$

Proof. Consider the random variable $Y = (X - m)^2$. By definition, we have $\sigma_X^2 = \mathbb{E}[Y]$. Furthermore, for every $a \geq 0$

$$\mathbb{P}[|X - m| \geq a] = \mathbb{P}[Y \geq a^2].$$

By applying Markov's inequality to Y (which is nonnegative), we obtain

$$\mathbb{P}[|X - m| \geq a] \leq \frac{\mathbb{E}[Y]}{a^2} \leq \frac{\sigma_X^2}{a^2}.$$

$\square$

Proposition 4.25 (basic properties of the variance)**.**

1. *Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then*

$$\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

2. *Let X be a random variable with $\mathbb{E}[X^2] < \infty$, let $\lambda \in \mathbb{R}$. Then*

$$\sigma_{\lambda X}^2 = \lambda^2 \cdot \sigma_X^2.$$

3. *Let $X_1, \ldots, X_n$ be n pairwise independent random variables and $S = X_1 + \cdots + X_n$. Then*

$$\sigma_S^2 = \sigma_{X_1}^2 + \cdots + \sigma_{X_n}^2.$$

Proof.

1. Let $m = \mathbb{E}[X]$. Using linearity of the expectation, we get

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

2. Using the formula of Item 1 and linearity of the expectation, we obtain

$$\sigma_{\lambda X}^2 = \mathbb{E}[(\lambda X)^2] - (\mathbb{E}[\lambda X])^2 = \lambda^2 \cdot \mathbb{E}[X^2] - \lambda^2 \cdot \mathbb{E}[X]^2 = \lambda^2 \cdot \sigma_X^2.$$

3. Writing $m_i = \mathbb{E}[X_i]$, we have

$$S - \mathbb{E}[S] = \sum_{i=1}^{n}(X_i - m_i),$$

and therefore

$$\sigma_S^2 = \sum_{1 \leq i,j \leq n} \mathbb{E}[(X_i - m_i)(X_j - m_j)].$$

However, for $i \neq j$, independence implies $\mathbb{E}[(X_i - m_i)(X_j - m_j)] = \mathbb{E}[(X_i - m_i)]\mathbb{E}[(X_j - m_j)] = 0$. Hence only the diagonal terms (for which $i = j$) survive in the sum and we obtain

$$\sigma_S^2 = \sum_{i=1}^{n} \mathbb{E}[(X_i - m_i)^2] = \sum_{i=1}^{n} \sigma_{X_i}^2.$$

$\square$

Application: Let S be a binomial random variables with parameters n and p. What is the variance of S?

Here, again, we can use that S has the same distribution as $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables with parameter p. Hence

$$\sigma_S^2 = \sigma_{S_n}^2 \overset{\text{independence}}{=} \sigma_{X_1}^2 + \cdots + \sigma_{X_n}^2$$
$$\overset{\text{ident. distrib.}}{=} n \cdot \sigma_{X_1}^2.$$

One has $\sigma_{X_1}^2 = \mathbb{E}[X_i^2] - p^2 = p - p^2 = p(1-p)$. Hence

$$\boxed{\sigma_S^2 = n \cdot p(1-p)}.$$

Here, we have discovered an important effect of summing i.i.d. random variables. One has

$$\mathbb{E}[S] = n \cdot p \quad \text{and} \quad \sigma_S = \sqrt{n} \cdot \sqrt{p(1-p)}$$

so the expectation of S grows like n, while the fluctuations of S_n grow like $\sqrt{n}$ thanks to cancellations in the sum $S_n - np = (X_1 - p) + \cdots + (X_n - p)$.

9 Covariance

We introduce the notion of covariance, which can be used in some cases as to quantify the dependence between two random variables.

Definition 4.26. *Let X, Y be two random variables. Assume that $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ (finite second moment). We define the **covariance between X and Y** as*

$$\boxed{\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].}$$

Remark: The condition that X and Y have finite second moment ensures that the covariance is well defined. Indeed by the elementary inequality $|XY| \leq \frac{1}{2}X^2 + \frac{1}{2}Y^2$ and monotonie and linearity of the expectation, we have

$$\mathbb{E}[|XY|] \leq \frac{1}{2}\mathbb{E}[X^2] + \frac{1}{2}\mathbb{E}[Y^2] < \infty.$$

As we have seen in Section 4, the covariance between two independent random variables vanishes:

$$X, Y \text{ independent} \implies \text{Cov}(X, Y) = 0.$$

The reciprocal implication is not true in general (see Exercise 6.6). Nevertheless, as we have seen in Section 6, we can obtain a characterization by using a stronger property involving test functions. By Theorem 4.17, we have

$$X, Y \text{ independent} \iff \forall \phi, \psi \text{ piecewise continuous, bounded} \quad \text{Cov}(\phi(X), \psi(Y)) = 0.$$

Chapter 5

Joint distribution

Goals

- Definition of the joint distribution for discrete/continuous random variables.

- Calculation of marginals

- Interpretation of dependence/independence of random variables.

Framework   We fix some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All the random variables considered in this chapter will be defined on this reference probability space.

1   Discrete joint distributions

1.1   Definition

> **Definition 5.1.** *Let $X_1, \ldots, X_n$ be n discrete random variables with $X_i \in W_i$ almost surely, for some $W_i \subset \mathbb{R}$ finite or countable. The joint distribution of $(X_1, \ldots, X_n)$ is the collection $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$ defined by*
>
> $$\boxed{p(x_1, \ldots, x_n) = \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]}$$

Example:

Let X, Y be two independent Bernoulli random variables with parameter $1/2$. The joint distribution of (X, Y) is given by

$$\forall x, y \in \{0, 1\} \qquad p(x, y) = \frac{1}{4}.$$

The joint distribution of (X, X) is equal to

$$\forall x, y \in \{0, 1\} \qquad p(x, y) = \begin{cases} \frac{1}{2} & x = y, \\ 0 & x \neq y. \end{cases}$$

Let $Z = X + Y$, then the joint distribution $p = (p(x, z))_{x \in \{0,1\}, z \in \{0,1,2\}}$ of (X, Z) is given by the following table:

x \ z	0	1	2
0	1/4	1/4	0
1	0	1/4	1/4

> **Proposition 5.2.** *The joint distribution of some random variables $X_1, \ldots, X_n$ satisfies*
>
> $$\boxed{\sum_{x_1 \in W_1, \ldots, x_n \in W_n} p(x_1, \ldots, x_n) = 1.} \tag{5.1}$$

Proof. Consider the event $A = \{X_1 \in W_1, \ldots, X_n \in W_n\}$. A is a finite intersection of almost sure events, hence $\mathbb{P}[A] = 1$ (see Exercise 2.1, in Sheet 2). Furthermore, it can be written as the disjoint union

$$A = \bigcup_{x_1 \in W_1, \ldots, x_n \in W_n} \{X_1 = x_1, \ldots, X_n = x_n\}.$$

Therefore,

$$1 = \mathbb{P}[A] = \sum_{x_1 \in W_1, \ldots, x_n \in W_n} \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n].$$

$\square$

Conversely, given some finite or countable sets $W_1, \ldots, W_n$ and a function $p : W_1 \times \cdots \times W_n \to [0, 1]$ satisfying (5.1), there exists a probability space and some discrete random variables with distribution p (see exercises).

1.2 Distribution of the image

One of the main advantages of working with random variables is that we can "manipulate" them as numbers. For instance, if we are given n random variables $X_1, X_2, \ldots, X_n$, we can think of them as n "random" numbers and we can make operations with them.

The following proposition gives the distribution of such random variables as images of discrete random variables.

Proposition 5.3. *Let $n \geq 1$ and $\phi : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary function. Let $X_1, \ldots, X_n$ be n discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with respective values in some finite or countable sets $W_1, \ldots, W_n$ almost surely. Then $Z = \phi(X_1, \ldots, X_n)$ is a discrete random with values in $W = \phi(W_1 \times \cdots \times W_n)$ almost surely and with distribution given by*

$$\forall z \in W \quad \mathbb{P}[Z = z] = \sum_{\substack{x_1 \in W_1, \ldots, x_n \in W_n \\ \phi(x_1, \ldots, x_n) = z}} \mathbb{P}[X_1 = x_1, \ldots X_n = x_n].$$

Proof. First notice that the set W is finite or countable, as an image of a finite and countable set. To prove that Z is a discrete random variable, it suffices to show that

- it takes values in W almost surely, and

- for every z in W, we have $\{Z = z\} \in \mathcal{F}$.

Indeed, the second item implies that for every $a \in \mathbb{R}$

$$\{Z \leq a\} = \bigcup_{\substack{z \in W \\ z \leq a}} \{Z = z\}$$

is also an event (as a countable union of events).

Now, the first item follows from the inclusion $\{X_1 \in W_1, \ldots, X_n \in W_n\} \subset \{Z \in W\}$. For the second item, let $z \in W$ and observe that

$$\{Z = z\} = \bigcup_{\substack{x_1 \in W_1, \ldots, x_n \in W_n \\ \phi(x_1, \ldots, x_n) = z}} \{X_1 = x_1, \ldots X_n = x_n\}. \tag{5.2}$$

Hence, $\{Z = z\} \in \mathcal{F}$ since it is a countable union of events. This implies that Z is a discrete random vartiable. To compute the distribution, observe that the union in (5.2) is disjoint and at most countable, hence,

$$\mathbb{P}[Z = z] = \sum_{\substack{x_1 \in W_1, \ldots, x_n \in W_n \\ \phi(x_1, \ldots, x_n) = z}} \mathbb{P}[X_1 = x_1, \ldots X_n = x_n].$$

$\square$

Example:

Consider the random variable $Z = X + Y$ defined (as in Section 1.1) as the sum of two independent Bernoulli random variables with parameter 1/2. By applying the proposition above to $\phi(x,y) = x + y$ we get

$$\mathbb{P}[Z = 0] = \sum_{\substack{x,y \in \{0,1\} \\ x+y=0}} \mathbb{P}[X = x, Y = y] = \mathbb{P}[X = 0, Y = 0] = 1/4$$

$$\mathbb{P}[Z = 1] = \sum_{\substack{x,y \in \{0,1\} \\ x+y=1}} \mathbb{P}[X = x, Y = y] = \mathbb{P}[X = 0, Y = 1] + \mathbb{P}[X = 1, Y = 0] = 1/2$$

$$\mathbb{P}[Z = 2] = \sum_{\substack{x,y \in \{0,1\} \\ x+y=2}} \mathbb{P}[X = x, Y = y] = \mathbb{P}[X = 1, Y = 1] = 1/4.$$

1.3 Marginal distributions

If one knows the joint distribution of $X_1, \ldots, X_n$, one can recover the distribution of each X_i separately. In this context the distribution of X_i is called the distribution of the i-th marginal.

Proposition 5.4. *Let $X_1, \ldots, X_n$ be n discrete random variables with joint distribution $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. For every i, we have*

$$\forall z \in W_i \qquad \mathbb{P}[X_i = z] = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} p(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n).$$

Proof. Apply Proposition 5.3 to $\phi(x_1, \ldots, x_n) = x_i$. $\qquad\square$

By the proposition above, if we know the joint distribution p of two random variables X, Y, then we can compute the distribution of X and the distribution of Y, but the converse is not true. Knowing the marginal distributions is not sufficient to compute the joint distribution. For example, let X, Y be two independent random variables. Then (X, Y) and (X, X) have the same marginal distributions (both Bernoulli $(1/2)$, but they have different joint distributions.

This notion of marginal distributions may help to understand joint distributions: Heuristically, the joint distribution of $X_1, \ldots, X_n$ encodes the distribution of each X_i separately, **as well as** how the random variables depend on each other.

1.4 Expectation of the image

Proposition 5.5. *Let $X_1, \ldots, X_n$ be n discrete random variables with joint distribu-*

tion $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. Let $\phi : \mathbb{R}^n \to \mathbb{R}$, then

$$\boxed{\mathbb{E}[\phi(X_1, \ldots, X_n)] = \sum_{x_1, \ldots, x_n} \phi(x_1, \ldots, x_n) p(x_1, \ldots, x_n)},$$

whenever the sum is well-defined.

Proof. Set $W = \phi(W_1, \ldots, W_n)$. Using the formula of Prop. 3.9, we have

$$\sum_{z \in F} z \cdot \mathbb{P}[Z = z] = \sum_{z \in F} \sum_{x_1, \ldots, x_n \in E} z \cdot \mathbf{1}_{\phi(x_1, \ldots, x_n) = z} \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n]$$

$$= \sum_{x_1, \ldots, x_n \in E} \underbrace{\left(\sum_{z \in F} z \cdot \mathbf{1}_{\phi(x_1, \ldots, x_n) = z} \right)}_{= \phi(x_1, \ldots, x_n)} \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n].$$

(the permutation of the two sums can be justified by the Fubini's theorem, provided the sum are well defined). $\qquad\square$

1.5 Independence

Proposition 5.6. *Let $X_1, \ldots, X_n$ be n discrete random variabless with joint distribution $p = (p(x_1, \ldots, x_n))_{x_1 \in W_1, \ldots, x_n \in W_n}$. The following are equivalent*

   (i) $X_1, \ldots, X_n$ are independent,

  (ii) $p(x_1, \ldots, x_n) = \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n]$ for every $x_1 \in W_1, \ldots, x_n \in W_n$.

Proof.

(i)$\Rightarrow$ (ii) Consider the functions $\phi_1, \ldots, \phi_n$ defined by $\phi_i(z) = \mathbf{1}_{z = x_i}$. We have

$$\begin{aligned}
\mathbb{P}[X = x_1, \ldots, X_n = x_n] &= \mathbb{E}[\mathbf{1}_{X_1 = x_1, \ldots, X_n = x_n}] \\
&= \mathbb{E}[\phi_1(X_1) \cdots \phi_n(X_n)] \\
&\overset{(i)}{=} \mathbb{E}[\phi_1(X_1)] \cdots \mathbb{E}[\phi_n(X_n)] \\
&= \mathbb{P}[X_1 = x_1] \cdots \mathbb{P}[X_n = x_n]
\end{aligned}$$

(ii)$\Rightarrow$ (i) Let $\phi_1 : W_1 \to \mathbb{R}, \ldots, \phi_n : W_n \to \mathbb{R}$. By Proposition 5.3 applied to $\phi(x_1, \ldots, x_n) =$

$\phi_1(x_1)\cdots\phi_n(x_n)$, we have

$$
\begin{aligned}
\mathbb{E}[\phi_1(X_1)\cdots\phi_n(X_n)] &= \sum_{x_1,\ldots,x_n} \phi_1(x_1)\cdots\phi_n(x_n)\cdot p(x_1,\ldots,x_n) \\
&\stackrel{(ii)}{=} \sum_{x_1,\ldots,x_n} \phi_1(x_1)\cdots\phi_n(x_n)\cdot \mathbb{P}[X_1 = x_x]\cdots\mathbb{P}[X_n = x_n] \\
&= \Big(\sum_{x_1}\phi_1(x_1)\cdot\mathbb{P}[X = x]\Big)\cdots\Big(\sum_{x_n}\phi_n(x_n)\cdot\mathbb{P}[X_n = x_n]\Big) \\
&= \mathbb{E}[\phi_1(X_1)]\cdots\mathbb{E}[\phi_n(X_n)].
\end{aligned}
$$

By Theorem 4.18, $X_1, \cdots, X_n$ are independent.

$\square$

2 Continuous joint distribution

2.1 Definition

Definition 5.7. *Let $n \geq 1$, some random variables $X_1,\ldots,X_n : \Omega \to \mathbb{R}$ have a **continuous joint distribution** if there exists a function $f : \mathbb{R}^n \to \mathbb{R}_+$ such that*

$$
\mathbb{P}[X_1 \leq a_1,\ldots,X_n \leq a_n] = \int_{-\infty}^{a_1}\cdots\int_{-\infty}^{a_n} f(x_1,\cdots,x_n)dx_n\cdots dx_1
$$

*for every $a_1,\ldots,a_n \in \mathbb{R}$. A function f as above is called a **joint density of** (X,Y).*

Proposition 5.8. *Let f be the joint density of n random variables $X_1,\ldots,X_n$. Then we have*

$$
\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)dx_n\ldots dx_1 = 1. \tag{5.3}
$$

Conversely, given a non negative function f satisfying (5.3), one can always construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and n random variables $X_1,\ldots,X_n : \Omega \to \mathbb{R}$ with joint density f (admitted).

Proof. By continuity of probability measures (Proposition 1.11) we have

$$
\begin{aligned}
1 = \lim_{a_1,\ldots,a_n\to\infty}\mathbb{P}[X_1 \leq a_1,\ldots,X_n \leq a_n] &= \lim_{a_1,\ldots,a_n\to\infty}\int_{-\infty}^{a_1}\cdots\int_{-\infty}^{a_n} f(x_1,\cdots,x_n)dx_n\cdots dx_1 \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)dx_n\ldots dx_1.
\end{aligned}
$$

$\square$

Interpretation: Informally, $f(x_1, \ldots, x_n)dx_1 \cdots dx_n$ represents the probability that the random vector $(X_1, \ldots, X_n)$ lies in the small region $[x_1, x_1 + dx_1] \times \cdots \times [x_n, x_n + dx_n]$.

Example 1: *Uniform point in the square*

Consider two random variables X and Y with joint density $f(x, y) = \mathbf{1}_{0 \leq x, y \leq 1}$, i.e.

$$f(x, y) = \begin{cases} 1 & (x, y) \in [0, 1]^2 \\ 0 & (x, y) \notin [0, 1]^2. \end{cases}$$

Example 2: *Uniform point in the disk*

Let $D = \{(x, y) : x^2 + y^2 \leq 1\}$ be the disk of radius 1 around 0. Consider two random variables X and Y with joint density $f(x, y) = \frac{1}{\pi}\mathbf{1}_{x^2+y^2 \leq 1}$, i.e.

$$f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & x^2 + y^2 > 1. \end{cases}$$

2.2 Expectation of the image

Proposition 5.9. *Let $\phi : \mathbb{R}^n \to \mathbb{R}$. If $X_1, \ldots, X_n$ have joint density f, then the expectation of the random variable $Z = \phi(X_1, \ldots, X_n)$ can be calculated by the formula*

$$\mathbb{E}[\phi(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(x_1, \ldots, x_n) \cdot f(x_1, \ldots, x_n)dx_n \ldots dx_1, \quad (5.4)$$

whenever the integral is well defined.

Proof. Admitted. $\qquad\square$

Applications: Consider the pair (X, Y) as in Example 1 above. By applying considering the function $\phi(x, y) = \mathbf{1}_{(x,y) \in R}$, we have for every rectangle $R = (a, a') \times (b, b') \subset [0, 1]^2$

$$\mathbb{P}[(X, Y) \in R] = \mathbb{E}[\phi(X, Y)] = \int_a^{a'} \int_b^{b'} dxdy = (a' - a)(b' - b) = \text{Area}(R),$$

and (X, Y) intuitively represents a uniform point in the square $[0, 1]^2$.

Equivalently, if we consider (X, Y) as in example 2, we find that for every rectangle $R = (a, a') \times (b, b') \subset D$

$$\mathbb{P}[(X, Y) \in R] = \frac{1}{\pi}(a' - a)(b' - b) = \frac{\text{Area}(R)}{\text{Area}(D)},$$

and (X, Y) intuitively represents a uniform point in the square $[0, 1]^2$.

2.3 Marginal densities

As in the discrete case, if $X_1, \ldots, X_n$ have a joint density f, then each X_i taken individually is continuous, and the density of X_i can be calculated from f by integrating over all the variables x_j, $j \neq i$.

Proposition 5.10. *Let $X_1, \cdots, X_n$ be n random variables with a joint density $f = f_{X_1, \ldots, X_n}$. Then for every i, X_i is a continuous random variable with density f_i given by*

$$f_i(z) = \int_{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathbb{R}^{n-1}} f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} dx_n.$$

for every $z \in \mathbb{R}$

Proof. We prove the result in the case two random variables X, Y (case $n = 2$). If X, Y possess a joint density $f_{X,Y}$, then we have

$$\mathbb{P}[X \leq a] = \mathbb{P}[X \in [-\infty, a], Y \in [-\infty, \infty]]$$
$$= \int_{-\infty}^{a} \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx,$$

and therefore X is continuous with density

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Equivalently Y is continuous with density

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

$\square$

Let us calculate the marginal densities in the two examples of joint densities we have seen above.

Example 1: *Uniform point in the square*

If $f_{X,Y}(x, y) = \mathbf{1}_{0 \leq x, y \leq 1}$, then X has density

$$f_X(x) = \int_{0,1} \mathbf{1}_{0 \leq x \leq 1} \mathbf{1}_{0 \leq y \leq 1} dy = \mathbf{1}_{0 \leq x \leq 1}.$$

and equivalently $f_Y(y) = \mathbf{1}_{0 \leq y \leq 1}$. In other words, both X and Y are uniform random variables in $[0, 1]$.

Example 2: *Uniform point in the disk*

If $f_{X,Y}(x, y) = \frac{1}{\pi} \mathbf{1}_{x^2 + y^2 \leq 1}$, then the density of X is

$$f_X(x) = \int_{-\infty}^{\infty} \frac{1}{\pi} \mathbf{1}_{x^2 + y^2 \leq 1} dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi} \sqrt{1 - x^2},$$

and equivalently $f_Y(y) = \frac{2}{\pi} \sqrt{1 - y^2}$.

2.4 Independence for continuous random variables

The following theorem gives a useful characterization for the independence of continuous random variables.

Theorem 5.11. *Let $X_1, \ldots, X_n$ be n continuous random variables with respective densities $f_1, \ldots, f_n$. The following are equivalent*

  (i) $X_1, \ldots, X_n$ are independent,

  (ii) $X_1, \ldots, X_n$ are jointly continuous with joint density

$$\boxed{f(x_1, \ldots, x_n) = f_1(x_1) \, \cdots \, f_n(x_n)}.$$

Remark 5.12. *An important consequence is that two independent continuous random variables are automatically jointly continuous.*

Proof. We prove the result for two random variables ($n = 2$). The more general case is proved similarly. Let X, Y be two continuous random variables with density f_X and f_Y respectively.

$\boxed{(i) \Rightarrow (ii)}$ If X and Y are independent, for every $a, b \in \mathbb{R}$ we have

$$\mathbb{P}[X \le a, Y \le b] = \mathbb{P}[X \le a] \cdot \mathbb{P}[Y \le b]$$
$$= \int_{-\infty}^{a} f_X(x)dx \cdot \int_{-\infty}^{b} f_Y(y)dy$$
$$= \int_{-\infty}^{a} \int_{-\infty}^{b} f_X(x)f_Y(y)dxdy.$$

Therefore, X and Y have joint density $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$.

$\boxed{(ii) \Rightarrow (i)}$ Applying the formula (5.4), we find, for every $a, b \in \mathbb{R}$,

$$\mathbb{E}[\phi(X)\psi(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x)\psi(y)f_{X,Y}(x,y)dxdy$$
$$\overset{(ii)}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x)\psi(y)f_X(x)f_Y(y)dxdy$$
$$= \int_{-\infty}^{\infty} \phi(x)f_X(x)dx \cdot \int_{-\infty}^{\infty} \psi(y)f_Y(y)dy$$
$$= \mathbb{E}[\phi(X)] \cdot \mathbb{E}[\psi(Y)].$$

Hence, by the characterization of independence (Theorem 4.18), we conclude that X and Y are independent random variables. $\square$

Example 1: *Uniform point in the square*
If X and Y have joint density $f_{X,Y}(x,y) = \mathbf{1}_{0 \le x,y \le 1}$, then

$$f_{X,Y}(x,y) = \mathbf{1}_{0 \le x \le 1}\mathbf{1}_{0 \le y \le 1} = f_X(x) \cdot f_Y(y).$$

In other words, the two coordinates of a uniform random point in $[0,1]^2$ are independent.

Example 2: *Uniform point in the disk*

If X and Y have joint density $f_{X,Y} = \frac{1}{\pi}\mathbf{1}_D$, then we have seen that $f_X(x) = \frac{2}{\pi}\sqrt{1-x^2}$ and $f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2}$, and therefore

$$f_{X,Y}(x,y) \neq f_X(x)f_Y(y).$$

The two coordinates X and Y of a uniform point in D are not independent! This fact can easily be understood by looking at the event that X is larger than (say) $\sqrt{3}/2$. In this case, we have some information about Y which is constrainted to belong to $[-1/2, 1/2]$.

Chapter 6

Asymptotic results

In this chapter, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$ In other words, we are given some random variables $X_i : \Omega \to \mathbb{R}$ such that

$$\forall i_1 < \cdots < i_k \ \forall x_1, \ldots x_k \in \mathbb{R} \quad \mathbb{P}[X_{i_1} \leq x_1, \ldots, X_{i_k} \leq x_k] = F(x_1) \cdots F(x_k).$$

where F is the common distribution function. For every n, we consider the partial sum

$$S_n = X_1 + \cdots + X_n,$$

and we are interested in the behavior (when n is large) of the random variable defined by

$$\frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}. \tag{6.1}$$

and sometimes called the **empirical average**.

1   Law of large numbers

Theorem 6.1. *Assume that $\mathbb{E}[|X_1|] < \infty$. Defining $m = \mathbb{E}[X_1]$ we have*

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = m \ \ a.s. \tag{6.2}$$

What does Eq. (6.2) mean?

If we fix $\omega \in \Omega$, the values $\frac{X_1(\omega)}{1}$, $\frac{X_1(\omega)+X_2(\omega)}{2}$, $\cdots$ simply define a sequence of real numbers. The properties of this sequence depends on the outcome ω. We are here interested on the ω's for which the sequence $\frac{X_1(\omega)+\cdots+X_n(\omega)}{n}$ converges to m. More precisely we consider $E \subset \Omega$ defined by

$$E = \{\omega \ : \ \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = m\} \tag{6.3}$$

One can check that E is an event and Eq. (6.2) says that $\mathbb{P}[E] = 1$.

Remark 6.2. *In the statement of the theorem, it may be surprising that the assumption and the definition of m are in terms of X_1 only. Actually, since the random variables are i.i.d. we also have $\mathbb{E}[|X_i|] < \infty$ and $m = \mathbb{E}[X_i]$ for every i.*

Example 1: *Bernoulli random variables*

If $X_1, X_2, \ldots$ is an infinite sequence of i.i.d. Bernoulli random variables with parameter p. Then we have

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = p \ \ \ a.s.$$

Example 2: *Exponential random variables*

If $T_1, T_2, \ldots$ is an infinite sequence of i.i.d. Exponential random variables with parameter λ. Then we have

$$\lim_{n\to\infty} \frac{T_1 + \cdots + T_n}{n} = \lambda \ \ \ a.s.$$

Proof. We prove the law of large numbers under a stronger moment assumption. We assume that

$$C := \mathbb{E}[X_1^4] < \infty.$$

Without loss of generality, we may assume that

$$\mathbb{E}[X_1] = 0. \tag{6.4}$$

Indeed, if we have the result for $m = 0$, we can extend it to $m \neq 0$ by considering the random variables $Y_i = X_i - m$, $i \geq 1$.

Fix $n \geq 1$ and consider the random variable

$$S_n = X_1 + \cdots + X_n.$$

By expanding Z^4 and using linearity of the expectation, we have

$$\mathbb{E}[S_n^4] = \sum_{1 \le i,j,k,\ell \le n} \mathbb{E}[X_i X_j X_k X_\ell].$$

As soon as one factor X_α appears a single time in the product $X_i X_j X_k X_\ell$, then independence and the hypothesis (6.4) imply that the expectation of the term vanishes. Hence, the only non vanishing terms are of the form $\mathbb{E}[X_i^4]$ and $\mathbb{E}[X_i^2 Y_j^2]$, for $i \ne j$. By independence and Jensen inequality, for $i \ne j$ we have $\mathbb{E}[X_i^2 Y_j^2] = \mathbb{E}[X_1^2]^2 \le C$. Since there are at most $n^2 + n$ non-vanishing terms in the sum above and each such term is smaller than C, we obtain

$$\mathbb{E}[S_n^4] \le C(n^2 + n) \le 2Cn^2.$$

For every n, consider the event

$$F_n = \left\{ \omega \in \Omega \ : \ \frac{|S_n(\omega)|}{n} < n^{-1/8} \right\}.$$

By Markov inequality, we have

$$\mathbb{P}[F_n^c] = \mathbb{P}[S_n^4 \ge n^{7/2}] \le \frac{\mathbb{E}[S_n^4]}{n^{7/2}} \le \frac{2C}{n^{3/2}}.$$

Now, for $N \ge 1$, consider the event

$$E_N = \bigcap_{n \ge N} F_n = \{\forall n \ge N \quad \frac{|S_n|}{n} \le n^{-1/8}\}.$$

By the union bound, we have

$$\mathbb{P}[E_N^c] = \mathbb{P}[\bigcup_{n \ge N} F_n^c] \le \sum_{n \ge N} \mathbb{P}[F_n^c] \le \sum_{n \ge N} \frac{C}{n^{3/2}}.$$

Hence $\lim_{N \to \infty} \mathbb{P}[E_N] = 1$. Furthermore, for every N, we have $E_N \subset E$ where E is the event defined in Eq. (6.3) (with $m = 0$). Therefore, $\mathbb{P}[E_N] \le \mathbb{P}[E]$, and the result follows by taking the limit as N tends to infinity. $\qquad \square$

2 Application: Monte-Carlo integration

The law of large number can be useful to approximate integrals, that may be difficult to compute exactly. Let $d \ge 1$ be an integer. Let $g : [0,1] \to \mathbb{R}$ such that

$$\int_0^1 |g(x)| dx < \infty.$$

How goal is to calculate

$$I = \int_0^1 g(x) dx.$$

Such an integral may be delicate to compute exactly, and we give a general method to obtain approximations of I. The key idea is to interpret I as an expectation. Let U be a uniform random variable in $[0,1]$. Then,

$$\mathbb{E}[g(U)] = \int_0^1 g(x)dx = I.$$

Hence, approximating I is equivalent to approximating the expectation of $g(U)$, which can be achieved by the law of large numbers. Let $U_1, U_2, \ldots$ be an i.i.d. sequence of uniform random variables in $[0,1]$, and consider $X_n = g(U_n)$ for every n. The sequence $X_1, X_2, \ldots$ is i.i.d. and we have

$$\mathbb{E}[|X_1|] = \int_0^1 |g(x)|dx < \infty,$$

and $\mathbb{E}[X_1] = I$. Hence, by the law of large numbers, we have

$$\lim_{n\to\infty} \frac{g(U_1) + \cdots + g(U_n)}{n} = I.$$

Hence, we obtain an approximation of I by calculating $\frac{g(U_1)+\cdots+g(U_n)}{n}$ for some large n. Notice that this quantity is easy to compute in practice, once we have simulated some uniform random variables $U_1, \ldots U_n$.

This method generalizes in several ways, one can use different densities to compute integrals over $\mathbb{R}$, and we can use joint densities to approximate d-dimensional integrals, $d \geq 2$.

3 Convergence in distribution

When we have deterministic numbers in $\mathbb{R}$, we can measure the distance between them: the distance between x and y is given by $|x - y|$. This gives rise to a natural notion of convergence. A sequence of real numbers $(x_n)_{n\in\mathbb{N}}$ converges to x if

$$\lim_{n\to\infty} |x_n - x| = 0.$$

For two random variables X and Y, one way to measure the "distance" between them is to look at their distribution functions. X and Y have similar probabilistic properties if their respective distribution functions F_X and F_Y are close to each other. This gives rise to the following notion of convergence for random variables, called "convergence in distribution".

Definition 6.3. *Let $(X_n)_{n\in\mathbb{N}}$ and X be some random variables. We write*

$$X_n \overset{\text{Approx}}{\approx} X \ \text{as } n \to \infty$$

if for every $x \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq x] = \mathbb{P}[X \leq x].$$

Example 1: *Bernoulli random variables*

For every n, let X_n be a Bernoulli random variable with parameter $p_n \in [0,1]$. If $\lim_{n \to \infty} p_n = p$. Then we have

$$X_n \overset{\text{Approx}}{\approx} X \text{ as } n \to \infty,$$

where X is a Bernoulli random variable with parameter p.

Example 2: *Approximation of the uniform*

In the first example, we have discrete random variables converging towards another discrete random variable. It is also possible that a sequence of discrete random variables converge towards a continuous random variable, as in the following example.

For every n, let X_n be a uniform random variable in $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}$ (i.e. $\mathbb{P}[X_n = \frac{k}{n}] = \frac{1}{n+1}$ for $k = 0, 1, 2. \ldots, n$). Then we have

$$X_n \overset{\text{Approx}}{\approx} X \text{ as } n \to \infty,$$

where X is a uniform random variable in $[0,1]$.

Indeed, for every $x \in [0,1]$, we have

$$\mathbb{P}[X_n \leq x] = \frac{\lfloor xn \rfloor}{n} \xrightarrow[n \to \infty]{} x = \mathbb{P}[X \leq x],$$

and the convergence for $x \notin [0,1]$ is trivial.

4   Central limit theorem

A question of fluctuation?

The law of large numbers tells us that for large n, the empirical average (6.1) is closed to the expectation $m = \mathbb{E}[X_1]$. A second very natural question to ask is:

$$\text{How far is } \frac{X_1 + \cdots + X_n}{n} \text{ from } m \text{ typically?}$$

The Gaussian case

Let us first look at the very instructive case when $X_1, X_2, \ldots$ is a sequence of i.i.d. normal random variables with parameters m and σ^2. Then the results we have seen on normal random variables tell us that

$$Z = \frac{X_1 + \cdots + X_n}{n} - m$$

is again a normal random variable with parameters $\bar{m} = 0$ and $\bar{\sigma}^2 = \frac{1}{n}\sigma^2$. The standard deviation $\bar{\sigma} = \frac{1}{\sqrt{n}}\sigma$ represents the typical fluctuations of Z. Roughly one can say that the typical distance between $\frac{X_1 + \cdots + X_n}{n}$ and m is of order $\frac{\sigma}{\sqrt{n}}$.

   In this context, a more natural random variable to consider is to rescale Z by a factor $\frac{\sqrt{n}}{\sigma}$ in order to get fluctuations of order 1: using again the properties of normal distributions we see that

$$\frac{\sqrt{n}}{\sigma}Z = \frac{X_1 + \cdots + X_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

is a standard normal.

   As a conclusion, if we consider i.i.d. normal distributions with expectation m and variance σ^2, then the random variable

$$\frac{X_1 + \cdots + X_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

corresponds to a rescaled version of the fluctuations of $\frac{X_1 + \cdots + X_n}{n}$ and is a standard normal.

General case: the central limit theorem

If $X_1, X_2, \ldots$ are not normal, it is in general not easy to compute the law of

$$\frac{X_1 + \cdots + X_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

Nevertheless, the central limit theorem tells us that this random variable always get close to a standard normal if n is large.

Theorem 6.4 (Central limit theorem)**.** *Assume that the expectation* $\mathbb{E}[X_1^2]$ *is well defined and finite. Defining* $m = \mathbb{E}[X_1]$ *and* $\sigma^2 = \mathrm{Var}(X_1)$, *we have*

$$\mathbb{P}\Big[\frac{S_n - n \cdot m}{\sqrt{\sigma^2 n}} \le a\Big] \xrightarrow[n \to \infty]{} \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx. \tag{6.5}$$

for every $a \in \mathbb{R}$

What does Eq. (6.5) **mean?** In words, the theorem above asserts that for n large, the distribution of the random variable

$$Z_n = \frac{S_n - n \cdot m}{\sqrt{\sigma^2 n}}$$

"looks like" the distribution of a normal random variables $\mathcal{N}(0,1)$. With the notation of Section 3, we have

$$Z_n \overset{\text{Approx}}{\approx} Z \text{ as } n \to \infty,$$

where $Z \sim \mathcal{N}(0,1)$.

Remark 1:

For every n, we can use linearity properties of the expectation and variance to show that

$$\mathbb{E}[Z_n] = 0 \quad \text{and} \quad \mathrm{Var}(Z_n) = 1.$$

Remark 2:

The central limit theorem helps us predicting the behaviour of S_n for n large. For example consider $p := \mathbb{P}[Z \in [-2, 2]]$, where Z is a standard normal random variable.

It is known that $p \simeq 0.95$ (this correspond to the blue and brown area in the picture below).
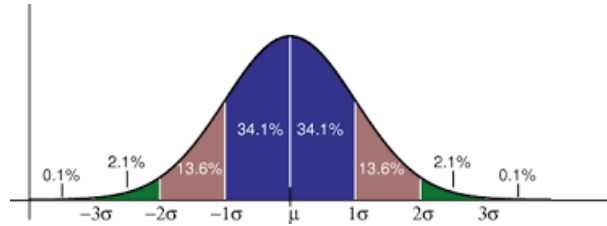

Figure 6.1: Quantiles of a normal random variable with parameters μ and σ^2

By the central limit theorem, we know that

$$\lim_{n\to\infty} \mathbb{P}\left[mn - 2\sqrt{\sigma^2 n} \leq S_n \leq mn + 2\sqrt{\sigma^2 n}\right] = p \simeq 95\%.$$

Bibliography

[LSW21] J. Lengler, A. Steger, and E. Welzl, **Algorithmen und Wahrscheinlichkeit**, 2021.

[Sch10] M. Schweizer, **Wahrscheinlichkeit und Statistik**, 2010.

[Wil01] David Williams, **Weighing the odds**, Cambridge University Press, Cambridge, 2001, A course in probability and statistics. MR 1854128