MATH 170E
Introduction to Probability and Statistics 1: Probability

Pablo López Rivera

University of California, Los Angeles (UCLA)

Fall 2025
Last update: October 28, 2025

Contents

Chapter 1

Random experiments and probability

The goal of this first chapter is to provide an introduction to the language of probability theory, which, in the context of this course, is the field within mathematics concerned with randomness and uncertainty, providing a rigorous framework to study these phenomena.

Let us highlight the ubiquity of probability theory that goes beyond its interaction with other fields in mathematics and its applications: it has proven to be a key tool in many different domains, such as physics, statistics, computer science, economics, sociology, biology, engineering, operations research, finance, marketing, business, etc. More generally, probability can be a powerful tool whenever we deal with uncertainty, randomness, and data.

In this first chapter, we will introduce the fundamental concepts of **random experiments** and **probability**. To do so, we first start by defining what a random experiment is in Section 1.1. Then, as a recap, in Section 1.2, we recall the language of sets and their basic operations. In Section 1.3, we define a probability. In Section 1.4, we will review counting methods that will be helpful to compute basic probabilities. Finally, in Sections 1.5, 1.6, 1.7, and 1.8, we introduce the notions of conditional probability, independence, the law of total probability, and the Bayes theorem, respectively.

1.1 Random experiments

Our first goal in this course will be to describe in mathematical terms what a random experiment is.

Definition 1.1 (Deterministic and random experiments)**.** An *experiment* is a procedure that has an observable outcome. We say that it is *deterministic* if its outcome can be predicted; that is, it has only one possible outcome. On the other hand, we say that it is *random* if it has more than one possible outcome that we cannot predict in advance.

Let us observe that when we perform a random experiment, in contrast with a deterministic experiment, we do not know beforehand its outcome. However, it is reasonable to assume that we know at least all the **possible outcomes**, which motivates the following definition.

Definition 1.2 (Outcome space and events)**.** Given a random experiment, we define its *outcome space* as the collection of all its possible outcomes, and we denote it by Ω. A subset of outcomes $A \subset \Omega$ is said to be an *event* associated with the random experiment.

Before continuing, let us provide examples of random experiments, their outcome spaces, and possible events.

Example 1.3 (Tossing a coin)**.** Suppose that our experiment consists of tossing a coin. If we assume it cannot land vertically, we only have two possible outcomes: *heads* and *tails*, denoted by H and T, respectively. Then its outcome space is given by

$$\Omega = \{H, T\}.$$

Some events we can consider are "obtaining heads" and "obtaining tails", which are represented by the sets $A = \{H\}$ and $B = \{T\}$, respectively.

Example 1.4 (Tossing two coins)**.** Suppose now that we are tossing two coins in order (that is, obtaining heads then tails (HT) is not the same as tails and then heads (TH). Then our new outcome space is given by

$$\Omega = \{HH, HT, TT, TH\}.$$

Some events could be "obtaining at least once tails", which is represented by the set $A = \{HT, TT, TH\}$; or "obtaining only heads", which is represented by $B = \{HH\}$; or "obtaining tails in the first coin flipping", which can be represented by $C = \{TT, TH\}$.

Example 1.5 (Instagram posts)**.** Now our random experiment is the following: suppose that you post a nice picture on Instagram and you wonder how many likes it will get. Note that

$$\Omega = \mathbb{N} = \{0, 1, 2, \dots\}$$

since a priori, one could have as many likes as people have an account on the social network if your account is public (assuming that there are infinite people on Instagram). Then you may be interested in studying the following events: "it got more than 100 likes", represented by $A = \{101, 102, \dots\}$ (i.e., the photo was a complete success); or "it got no likes at all", represented by $B = \{0\}$ (i.e., the picture was awful). In contrast, if your account is private, and you have $N \in \mathbb{N}$ followers, then note that

$$\Omega = \{0, \dots, N\}.$$

Example 1.6 (Darts)**.** Suppose you are playing darts and are a good player, so you always hit your darts inside the dartboard. Then we may represent the outcome space by a circle:

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}.$$

A possible event is "hitting the first quadrant", represented by the set

$$A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1, x \geqslant 0, y \geqslant 0\}.$$

Exercise 1.1. In Example 1.6, how would you represent using a set the event "hitting the inner bullseye" (the red one in the middle)?

1.2 Sets and their operations

In the last section, we saw that we are modeling our theory in events, i.e., sets. Before continuing, we provide a recap on the basic set operations that will be pertinent to this course. Firstly, recall that if Ω is a set and x is an element of Ω, we say that x *belongs to* Ω and we denote it by $x \in \Omega$. Now we recall the notion of power set.

Definition 1.7 (Power set)**.** Let Ω be a set. We define the *power set* associated with Ω, which we denote by $\mathcal{P}(\Omega)$, as the set that contains all the subsets of Ω. That is,

$$\mathcal{P}(\Omega) \coloneqq \{A : A \subset \Omega\}.$$

Remark 1.8.

(i) Note that the power set is a set that has sets as its elements.

(ii) If Ω is a set, always $\varnothing \in \mathcal{P}(\Omega)$ and $\Omega \in \mathcal{P}(\Omega)$.

Let us provide an example before introducing other set operations.

Example 1.9. If $\Omega = \{0, 1, 2\}$, then

$$\mathcal{P}(\Omega) = \{\varnothing, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{1, 2\}, \{0, 2\}, \Omega\}.$$

We recall the definition of the complement, union, intersection, and difference of sets.

Definition 1.10 (Complement, union, intersection, and difference)**.** Let Ω be a set and let $A, B \subset \Omega$.

(i) We define the *complement* of A (with respect to Ω), which we denote by A', as the set that contains all the elements that **do not** belong to A (see Figure 1.1):

$$A' \coloneqq \{x \in \Omega : x \notin A\}.$$

(ii) We define the *union* between A and B, denoted by $A \cup B$, as the set that contains all the elements that belong to A **or** belong to B (see Figure 1.2):

$$A \cup B \coloneqq \{x \in \Omega : x \in A \text{ or } x \in B\}.$$

(iii) We define the *intersection* between A and B, denoted by $A \cap B$, as the set that contains all the elements that belong to A **and** belong to B (see Figure 1.3):

$$A \cap B \coloneqq \{x \in \Omega : x \in A \text{ and } x \in B\}.$$

(iv) We define the *set difference* of A and B, which we denote by $A \setminus B$, as the set that contains all the elements that belong to A but not to B (see Figure 1.4):

$$A \setminus B \coloneqq \{x \in \Omega : x \in A \text{ and } x \notin B\}.$$

Remark 1.11. In general, $A \setminus B \neq B \setminus A$ (compare Figures 1.4 and 1.5 and see Exercise 1.2 below).

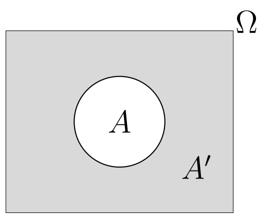

Figure 1.1: Complement of A in Ω (grey area).

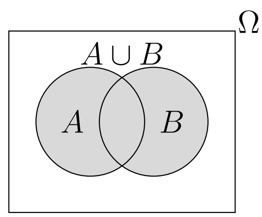

Figure 1.2: Union of A and B (grey area).

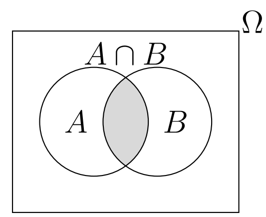

Figure 1.3: Intersection of A and B (grey area).

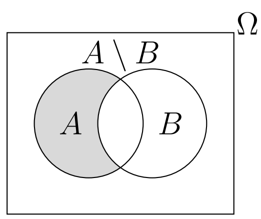

Figure 1.4: Difference of A and B (grey area).

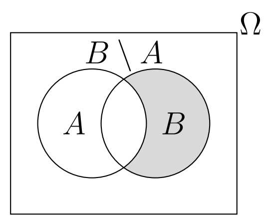

Figure 1.5: Difference of B and A (grey area).

Exercise 1.2. Let $\Omega = \{0, 1, 2, 3, 4\}$, and let $A = \{0, 2\}$ and $B = \{1, 2, 3\}$. Compute A', B', $A \cup B$, $A \cap B$, $A \setminus B$, and $B \setminus A$.

These operations verify the following algebraic properties.

Proposition 1.12 (Algebra of sets)**.** *Let Ω be a set and let $A, B, C \subset \Omega$.*

(i) Union and intersection are commutative*:*

$$A \cup B = B \cup A \quad \textit{and} \quad A \cap B = B \cap A.$$

(ii) The empty set is the neutral element *for the union: $A \cup \varnothing = A$.*

(iii) *The set* Ω *is the* neutral element *for the intersection:* $A \cap \Omega = A$.

(iv) *Union and intersection are* associative*:*

$$(A \cup B) \cup C = A \cup (B \cup C) \quad \textit{and} \quad (A \cap B) \cap C = A \cap (B \cap C).$$

(v) *Intersection* distributes *over union:*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

(vi) *Union* distributes *over intersection:*

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

(vii) *Union and intersection verify* De Morgan's laws*:*

$$(A \cup B)' = A' \cap B' \quad \textit{and} \quad (A \cap B)' = A' \cup B'.$$

(viii) *The complement satisfies*

$$\Omega = A \cup A', \quad \varnothing = A \cap A', \quad \textit{and} \quad (A')' = A.$$

(ix) *The set difference can be written as*

$$A \setminus B = A \cap B'.$$

(x) *One has* $A \cap (B \setminus A) = \varnothing$*, and*

$$A \cup B = A \cup (B \setminus A).$$

Exercise 1.3. Prove Proposition 1.12.

To end this section, we finish with some notions that will be useful later.

Definition 1.13 (Mutually exclusive sets)**.** Let Ω be a set, and let $(A_i)_{i=1}^n = (A_1, \ldots, A_n)$ be a collection of sets (that is, for each $i \geqslant 1$, $A_i \subset \Omega$). We say that $(A_i)_{i=1}^n$ is *mutually exclusive* if for every $i \neq j$, A_i and A_j are *disjoint*:

$$A_i \cap A_j = \varnothing.$$

Definition 1.14 (Exhaustive sets)**.** Let Ω be a set, and let $(A_i)_{i=1}^n = (A_1, \ldots, A_n)$ be a collection of sets.We say that $(A_i)_{i=1}^n$ is *exhaustive* if

$$\bigcup_{i=1}^{n} A_i = A_1 \cup \cdots \cup A_n = \Omega.$$

Remark 1.15. If $A \subset \Omega$, then A induces a natural collection of sets that is mutually exclusive and exhaustive, which is given by $\{A, A'\}$ since $\Omega = A \cup A'$ and $A \cap A' = \varnothing$.

Remark 1.16. Both Definitions 1.13 and 1.14 can be extended mutatis mutandis for infinite collections of subsets.

1.3 Measuring events

Once we have written in the language of sets the possible outcomes of a random experiment, we arrive at the essential point of probability theory: we want to **measure events** in terms of how they are likely to happen.

Definition 1.17 (Probability)**.** Given a random experiment with outcome space Ω, a *probability* is a function $\mathbb{P}\colon \mathcal{P}(\Omega) \to \mathbb{R}$ (that is, it takes subsets of Ω and returns a real number) that satisfies the following three properties:

(i) It is a nonnegative function: for every $A \subset \Omega$, $\mathbb{P}(A) \geqslant 0$.

(ii) The full outcome space has probability one: $\mathbb{P}(\Omega) = 1$.

(iii) It is *countably additive*: for any countable collection of events $(A_i)_{i=1}^{+\infty}$ that is mutually exclusive, then
$$\mathbb{P}\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i).$$

Remark 1.18 (Probabilities are finitely additive)**.** Item (iii) implies, in particular, that $\mathbb{P}$ is *finitely additive*: for any finite collection of events $(A_i)_{i=1}^{n}$ that is mutually exclusive, then
$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$
For example, when $n = 2$ this means that if $A_1, A_2 \subset \Omega$ with $A_1 \cap A_2 = \varnothing$,
$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

Before introducing concrete examples of probabilities, let us state their most important properties.

Theorem 1.19. *Consider an experiment with outcome space Ω and let $\mathbb{P}\colon \mathcal{P}(\Omega) \to \mathbb{R}$ be a probability. Then $\mathbb{P}$ satisfies the following properties:*

(i) For any $A \subset \Omega$, $\mathbb{P}(A') = 1 - \mathbb{P}(A)$.

(ii) $\mathbb{P}(\varnothing) = 0$.

(iii) It is monotone*: for $A, B \subset \Omega$ with $A \subset B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.*

(iv) For any $A \subset \Omega$, $\mathbb{P}(A) \leqslant 1$.

(v) The inclusion-exclusion principle*: for any $A, B \subset \Omega$,*
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Proof.

(i) Let $A \subset \Omega$. In the light of Remark 1.15, the collection $\{A, A'\}$ is mutually exclusive, so item (iii) in Definition 1.17 (recall Remark 1.18) yields

$$\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A').$$

By item (ii) in Definition 1.17, we have that $\mathbb{P}(\Omega) = 1$. Hence, $1 = \mathbb{P}(A) + \mathbb{P}(A')$, which yields $\mathbb{P}(A') = 1 - \mathbb{P}(A)$.

(ii) Direct from item (i) in Theorem 1.19 by applying $A = \varnothing$.

(iii) Let $A, B \subset \Omega$ with $A \subset B$. Then we can see that $B = A \cup (B \setminus A)$. Since $A \cap (B \setminus A) = \varnothing$, then by item (iii) in Definition 1.17,

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A') \overset{(*)}{\geqslant} \mathbb{P}(A) + 0 = \mathbb{P}(A),$$

where the inequality (*) is justified since $\mathbb{P}(B \cap A') \geqslant 0$ (item (i) in Definition 1.17).

(iv) Let $A \subset \Omega$. Then by (iii) in Theorem 1.19, we have that $\mathbb{P}(A) \leqslant \mathbb{P}(\Omega)$, but $\mathbb{P}(\Omega) = 1$ (item (ii) in Definition 1.17).

(v) Let $A, B \subset \Omega$. Recall item (x) in Proposition 1.12. Then

$$A \cup B = A \cup (B \setminus A).$$

Since A and $B \setminus A$ are disjoint,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A). \tag{1.3.1}$$

On the other hand, by items (iii), (viii), (v), and (ix) in Proposition 1.12, we have that

$$B = B \cap \Omega = B \cap (A \cup A') = (B \cap A) \cup (B \cap A') = (B \cap A) \cup (B \setminus A)$$

Since $(B \cap A) \cap (B \setminus A) = \varnothing$, then

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \setminus A). \tag{1.3.2}$$

The conclusion follows from mixing both equations (1.3.1) and (1.3.2).

□

Exercise 1.4. For $A, B, C \subset \Omega$, compute $\mathbb{P}(A \cup B \cup C)$ in the fashion of item (v) in Theorem 1.19. Can you generalize the statement if we have a collection of n subsets?

Now we are ready to provide a first example of a probability associated with a random experiment having a finite number of possible outcomes, which is quite natural: the equally likely probability, which consists of assigning the same probability to each possible outcome.

Definition 1.20 (Equally-likely)**.** Consider a random experiment with a finite outcome space Ω with $m \in \mathbb{N}^*$ elements. The *equally-likely* or *uniform* probability is defined as the probability $\mathbb{P}$ such that

$$\forall \omega \in \Omega, \quad \mathbb{P}(\{\omega\}) = \frac{1}{m}.$$

Remark 1.21 (Laplace's rule)**.** Consider a finite outcome space Ω having m elements, equipped with the uniform probability $\mathbb{P}$. Consider an event $A \subset \Omega$ having k elements. Then

$$\mathbb{P}(A) = \frac{k}{m},$$

which is no more than Laplace's rule: under the uniform probability, the probability of any event can be computed by dividing the number of results that form the event by the number of possible outcomes. In other words, that is no more than the **ratio between favorable and total cases**.

Remark 1.22 (Law of large numbers: take 1)**.** Another interpretation of Laplace's rule is the following: imagine that you observe a random experiment with outcome space Ω and suppose it has a probability $\mathbb{P}$ that is unknown. Given an event $A \subset \Omega$, you may want to estimate how likely it is to happen, that is, you want to approximate $\mathbb{P}(A)$. If you can repeat many times the experiments (and if you assume that $\mathbb{P}$ is always the probability, for each repetition), then you can estimate $\mathbb{P}(A)$ in the following way: if you perform the experiment n times, let $m(A)$ be the number of times you observed the event A amongst the n repetitions. Then, if n is large enough,

$$\mathbb{P}(A) \approx \frac{m(A)}{n}.$$

This is no more than the *law of large numbers*, which we will see later in this course.

Example 1.23 (Rolling a dice)**.** Rolling a dice with six faces has six possible outcomes:

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

Saying that the dice is fair is equivalent to saying that each number has the same probability, equal to $\frac{1}{6}$. The probability of getting an odd number is $\frac{3}{6} = \frac{1}{2}$.

Example 1.24 (Drawing cards from a deck)**.** If you draw a card from a standard 52-card deck, the probability of drawing an ace of spades is equal to $\frac{1}{52}$. The probability of getting an ace is $\frac{4}{52} = \frac{1}{13}$. On the other hand, the probability of drawing a spade is $\frac{13}{52} = \frac{1}{4}$.

1.4 Counting principles

As we have seen at the end of the last section, a prominent example of probability is the uniform one, in the context of experiments with a finite outcome space. Recall that, in the light of Laplace's rule (Remark 1.21), if we want to compute $\mathbb{P}(A)$, where $\mathbb{P}$ is the uniform distribution, then

$$\mathbb{P}(A) = \frac{\text{size}(A)}{\text{size}(\Omega)}.$$

Hence, we are done if we can determine how many elements are in both A and Ω. Having said that, the goal of this section is to introduce some *counting principles* that will allow us to count elements from sets.

1.4.1 Rule of sum

The rule of sum states that if one can choose between two actions A_1 and A_2 that **cannot be done at the same time**, and there are n_1 ways to do A_1, and distinct from them, n_2 ways to do A_2, then the number of ways to do A_1 or A_2 is $n_1 + n_2$.

Remark 1.25. The rule of sum can be extended to the case when one has $A_1, \ldots, A_m$ different actions that cannot be done at the same time, and there are $n_1, \ldots, n_m$ different ways to do each one, respectively: the number of ways to do A_1, A_2, $\ldots$, or A_m is $n_1 + \cdots + n_m$.

Example 1.26 (Choosing a course)**.** Suppose that for this semester, you have to choose between MATH 131A Analysis or MATH 131AH Analysis (Honors), and that four instructors are teaching MATH 131A and two are teaching MATH 131AH. Then you have $4 + 2 = 6$ different classes to choose from.

1.4.2 Rule of product

The rule of product states that if an action A is divided into two actions A_1 and A_2 that are **independent** and such that there are n_1 ways to do A_1 and n_2 ways to do A_2, then there are $n_1 \times n_2$ ways to do A.

Remark 1.27. The rule of product can be extended to the case when one has an action A that is divided into $A_1, \ldots, A_m$ different actions that are perfomed independently, and there are $n_1, \ldots, n_m$ different ways to do each one, respectively: the number of ways to do A is $n_1 \times \cdots \times n_m$.

Example 1.28 (Menu of the day)**.** Suppose you go to a restaurant and choose the "menu of the day" that allows you to choose a starter amongst three options, a main dish amongst two options, and a dessert amongst five options. Then you have $3 \times 2 \times 5 = 30$ different choices.

1.4.3 Permutations of n objects

A direct consequence of the multiplication principle is the following: suppose that you have n different objects to be placed into n different sites. Then there are $n!$ different ways to do that action, where $n!$ is n *factorial*: if $n = 0$, we define $0! := 1$, and

$$\forall n \geqslant 1, \quad n! := n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1.$$

Example 1.29 (PhD advisor)**.** Suppose you are a Professor at UCLA and you have five PhD students. You must meet them each week, but on different days (excluding Saturdays and Sundays). Then you have $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ different ways to schedule weekly meetings with them.

1.4.4 Permutations of n objects taken k

Suppose you have n different objects and you have to choose k between those n **in order**. Then you have P_k^n different ways to choose them, where

$$\forall n, k \geqslant 0, \quad P_k^n := \frac{n!}{(n-k)!}.$$

Example 1.30 (Holidays)**.** Suppose that you want to travel to Europe on your holidays (3 weeks) and that you want to visit different countries, spending exactly 1 week in each country. On the other hand, since it will be your first time in Europe, you want to visit the most touristic countries: France, Spain, Italy, England, Portugal, and Germany. Then, if you consider the order in which you visit the countries, you have

$$P_3^6 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \times 5 \times 4 \times \cancel{3 \times 2 \times 1}}{\cancel{3 \times 2 \times 1}} = 6 \times 5 \times 4 = 120$$

different possible itineraries.

1.4.5 Combinations of n objects taken k

Suppose you have n different objects and you have to choose k between those n, but you **do not care about the ordering in which those were drawn**. Then you have C_k^n different ways to choose them, where

$$\forall n, k \geqslant 0, \quad C_k^n := \frac{n!}{k!(n-k)!}.$$

We also write $\binom{n}{k}$ for C_k^n.

Example 1.31 (Coachella)**.** Suppose you are the producer of Coachella, so you have the hard but amusing task of choosing three headliners (you do not have to decide the day they will play, just the artists). You have to choose between eight artists: Bad Bunny, Tame Impala, Dua Lipa, The Strokes, Daft Punk (let us assume that they have reunited), The Rolling Stones, Drake, or Oasis. Then you have

$$C_4^8 = \frac{8!}{3!(8-3)!} = \frac{8!}{3!5!} = \frac{8 \times 7 \times 6 \times \cancel{5 \times 4 \times 3 \times 2 \times 1}}{(3 \times 2 \times 1) \times \cancel{(5 \times 4 \times 3 \times 2 \times 1)}} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

possible choices of headliners.

1.5 Conditional probability

The object we define in this section is motivated by the following situation: suppose you observe a random experiment with outcome space Ω and probability $\mathbb{P}$. Let $B \subset \Omega$ be an event you know has already happened. If $A \subset \Omega$ is another event, you may wonder **what is the probability of A, given that B has occurred**.

Definition 1.32 (Conditional probability)**.** Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Fix $B \subset \Omega$ with $\mathbb{P}(B) > 0$. We define the *conditional probability given B* as the function $\mathbb{P}(\cdot|B) \colon \mathcal{P}(\Omega) \to \mathbb{R}$ defined by

$$\forall A \subset \Omega, \quad \mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Note that the philosophy behind Definition 1.32 is the following: given that the event B has happened, it is natural to compute, for $A \subset \Omega$, $\mathbb{P}(A \cap B)$, which is **the probability that both A and B have happened**. The sole problem of the function $A \mapsto \mathbb{P}(A \cap B)$ is that **it is not a probability** in the sense of Definition 1.17 when $\mathbb{P}(B) < 1$ because $\mathbb{P}(\Omega \cap B) < 1$. However, when we divide $\mathbb{P}(A \cap B)$ by $\mathbb{P}(B)$, we are "renormalizing", so that $\mathbb{P}(\Omega|B)$ adds up to 1, so it is indeed a probability.

Proposition 1.33. *Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Fix $B \subset \Omega$ with $\mathbb{P}(B) > 0$. Then the function $\mathbb{P}(\cdot|B) \colon \mathcal{P}(\Omega) \to \mathbb{R}$ is a probability.*

Exercise 1.5. Prove Proposition 1.33.

Remark 1.34. In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

Let us provide a concrete example of conditional probability.

Example 1.35 (Rolling a die)**.** Suppose you roll a fair six-sided (i.e., a standard) die. Let B be the event "you obtain a number strictly greater than 3" and A "obtaining an even number". If you want to compute $\mathbb{P}(A|B)$, note that, on the one hand, $B = \{4, 5, 6\}$, so $\mathbb{P}(B) = \frac{3}{6} = \frac{1}{2}$. On the other hand, remark that $A \cap B = \{4, 6\}$, so $\mathbb{P}(A \cap B) = \frac{2}{6} = \frac{1}{3}$. Then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Conditional probability satisfies the following properties.

Proposition 1.36. *Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$.*

(i) Let $A, B \subset \Omega$ with $\mathbb{P}(B) > 0$. Then

$$\mathbb{P}(A'|B) = 1 - \mathbb{P}(A|B).$$

(ii) *Let* $B \subset \Omega$ *with* $\mathbb{P}(B) > 0$. *Then* $\mathbb{P}(\cdot|B)$ *is countably additive: for any countable collection of events* $(A_i)_{i=1}^{+\infty}$ *that is mutually exclusive, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{+\infty} A_i \middle| B\right) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i|B).$$

(iii) *Let* $B \subset \Omega$ *with* $\mathbb{P}(B) > 0$. *Then* $\mathbb{P}(\cdot|B)$ *is finitely additive: for any finite collection of events* $(A_i)_{i=1}^{n}$ *that is mutually exclusive, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i \middle| B\right) = \sum_{i=1}^{n} \mathbb{P}(A_i|B).$$

(iv) *Let* $A, B \subset \Omega$ *with* $\mathbb{P}(A) > 0$. *Then*

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

(v) *Let* $A, B \subset \Omega$ *with* $\mathbb{P}(B) > 0$. *Then*

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

(vi) *Let* $A, B, C \subset \Omega$ *with* $\mathbb{P}(A) > 0$ *and* $\mathbb{P}(A \cap B)$. *Then*

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

Proof.

(i) It is a direct consequence of Proposition 1.33 and item (i) in Theorem 1.19.

(ii) Similar to (i): consequence of Proposition 1.33.

(iii) By definition, $\mathbb{P}(B|A) = \frac{\mathbb{P}(B\cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}$. The conclusion follows if we multiply the equation by $\mathbb{P}(A)$.

(iv) Similar to (iii).

(v) Let us compute the right-hand side:

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

$\square$

Let us provide another example where these properties can be applied.

Example 1.37 (Insurance)**.** An insurance company sells several types of insurance policies, including auto and homeowner policies. Let A_1 be those people with an auto policy only, A_2 those people with a homeowner policy only, A_3 those people with both an auto and homeowner policy, and A_4 those with only types of policies other than auto and homeowner policies. For a person randomly selected from the company's policy-holders, suppose that

$$\mathbb{P}(A_1) = 0.3, \quad \mathbb{P}(A_2) = 0.2, \quad \mathbb{P}(A_3) = 0.2, \quad \text{and} \quad \mathbb{P}(A_4) = 0.3.$$

Let B be the event that an auto or homeowner policy holder will renew at least one of those policies. Say from past experience that we assign the conditional probabilities

$$\mathbb{P}(B|A_1) = 0.6, \quad \mathbb{P}(B|A_2) = 0.7, \quad \text{and} \quad \mathbb{P}(B|A_3) = 0.8.$$

Given that the person selected at random has an auto or homeowner policy, what is the conditional probability that the person will renew at least one of those policies? Note that we want to compute $\mathbb{P}(B|A_1 \cup A_2 \cup A_3)$. Note that

$$\begin{aligned}\mathbb{P}(B|A_1 \cup A_2 \cup A_3) &= \frac{\mathbb{P}(B \cap (A_1 \cup A_2 \cup A_3))}{\mathbb{P}(A_1 \cup A_2 \cup A_3)} = \frac{\mathbb{P}((B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3))}{\mathbb{P}(A_1 \cup A_2 \cup A_3)} \\ &= \frac{\mathbb{P}(B \cap A_1) + \mathbb{P}(B \cap A_2) + \mathbb{P}(B \cap A_1)}{\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3)} \\ &= \frac{\mathbb{P}(A_1)\mathbb{P}(B|A_1) + \mathbb{P}(A_2)\mathbb{P}(B|A_2) + \mathbb{P}(A_3)\mathbb{P}(B|A_3)}{\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3)} \\ &= \frac{0.3 \times 0.6 + 0.2 \times 0.7 + 0.2 \times 0.8}{0.3 + 0.2 + 0.2} \approx 0.6857.\end{aligned}$$

1.6 Independence

Many random experiments have the following property: their repetitions are independent; that is, they do not depend on previous realizations nor do they affect future ones.

Example 1.38 (Tossing coins and independence)**.** To give an example, let us go back to Example 1.4, so that

$$\Omega = \{HH, HT, TT, TH\},$$

and let us endow Ω with the uniform probability Ω. Let A be the event "obtaining heads in the first tossing" and B be the event "obtaining heads in the second tossing". At least intuitively, these two events should be independent, but how? Then let us study the dependence between A and B via the conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{HH\})}{\mathbb{P}(\{HH, TH\})} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{2}{4} = \frac{1}{2}.$$

But

$$\mathbb{P}(A) = \mathbb{P}(\{HH, HT\}) = \frac{2}{4} = \frac{1}{2},$$

so

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

In the light of item (v) in Proposition 1.36, this can be rewritten as

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

What we have seen in Example 1.38 motivates the probabilistic definition of independence.

Definition 1.39 (Independence)**.** Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $A, B \subset \Omega$. We say that A and B are *independent* if

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

Now, let us observe some of the immediate consequences of the previous definition.

Proposition 1.40. *Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$.*

(i) Let $A, B \subset \Omega$ be independent events. Then the pairs A and B'; A' and B; and A' and B' are independent as well.

(ii) Let $A, B \subset \Omega$ be independent events with $\mathbb{P}(A) > 0$. Then A and B are independent if and only if

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

(iii) Let $A, B \subset \Omega$ be independent events with $\mathbb{P}(B) > 0$. Then A and B are independent if and only if

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

(iv) Let $A \subset \Omega$ with $\mathbb{P}(A) = 0$. Then, for any $B \subset \Omega$, A and B are independent.

Now we want to generalize Definition 1.39 for more than two events. However, the following exercise shows that we should do it carefully.

Exercise 1.6. Consider the following experiment: you roll a fair six-sided die two times, and consider the following events: A representing "obtaining an odd number on the first roll"; B representing "obtaining an odd number on the second roll"; and C representing "the sum of the two rolls is odd".

(i) Write the outcome space Ω.

(ii) Why should we consider here the uniform probability $\mathbb{P}$ on $\mathcal{P}(\Omega)$?

(iii) Write explicitely the events A, B and C.

(iv) Show that the the collection $\{A, B, C\}$ is *pairwise independent*: that is, the pairs A and B, B and C, and A and C are independent.

(v) Show that $\mathbb{P}(A \cap B \cap C) = 0$.

(vi) Compute $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

The last exercise shows that even if we assume that a trio of events $\{A, B, C\}$ is pairwise independent, it is not necessarily true that $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. Now we provide the proper definition.

Definition 1.41 (Mutual independence)**.** Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $A, B, C \subset \Omega$. We say that the events A, B, and C are *mutually independent* if

(i) A, B, and C are pairwise independent:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C), \quad \text{and} \quad \mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C).$$

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

We can even extend the definition for more than three events.

Definition 1.42 (Mutual independence, general case)**.** Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $(A_i)_{i=1}^n$ be collection of events. We say that the events $(A_i)_{i=1}^n$ are *mutually independent* if for every $k \in \mathbb{N}^*$, and every $1 \leqslant i_1 < \cdots < i_k \leqslant n$,

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}). \tag{1.6.1}$$

Remark 1.43. In simple words, the events of a collection $(A_i)_{i=1}^n$ are mutually independent if all the pairs, triples, quartets, etc. made of events of the collection satisfy (1.6.1).

1.7 Law of total probability

In this section, we provide a powerful principle, the law of total probability, which is helpful when the outcome space can be divided into smaller pieces.

Theorem 1.44 (Law of total probability)**.** *Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $(B_i)_{i=1}^n$ be a collection of events that is mutually exclusive and exhaustive (cf. Definitions 1.13 and 1.14). Let $A \subset \Omega$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \cap B_i).$$

Before proceeding with the proof, note that Theorem 1.44 says that if Ω can be partitioned into smaller pieces $(B_i)_{i=1}^n$ and we can compute the probability of the events $A \cap B_i$ for all $1 \leqslant i \leqslant n$, then we can calculate $\mathbb{P}(A)$.

Proof of Theorem 1.44. Since $(B_i)_{i=1}^n$ is exhaustive, then $\Omega = \bigcup_{i=1}^n B_i$, so

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^n B_i \right) = \bigcup_{i=1}^n (A \cap B_i).$$

The events $(B_i)_{i=1}^n$ are mutually exclusive, and thus so are $(A \cap B_i)_{i=1}^n$. Then

$$\mathbb{P}\left(\bigcup_{i=1}^n (A \cap B_i) \right) = \sum_{i=1}^n \mathbb{P}(A \cap B_i),$$

and the conclusion follows. $\square$

It is also possible to rewrite the law of total probability in the language of conditional probability.

Corollary 1.45 (Law of total probability, conditional version)**.** *Consider an experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $(B_i)_{i=1}^n$ be a collection of events that is mutually exclusive and exhaustive. Furthermore, assume that for each $1 \leqslant i \leqslant n$, $\mathbb{P}(B_i) > 0$. Let $A \subset \Omega$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^k \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

Proof. From the law of total probability, we know that

$$\mathbb{P}(A) = \sum_{i=1}^k \mathbb{P}(A \cap B_i).$$

But note that for each $1 \leqslant i \leqslant n$,

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

$\square$

We finish this section with an example.

Example 1.46 (Aces)**.** The experiment here is drawing two cards from a standard deck in order and without replacing. We want to compute the probability of the event A given by "the second card drawn is an ace". We will partition Ω into two disjoint pieces: let B_1

be the event "the first card drawn is an ace" and let B_2 represent "the first card drawn is not an ace", so that $\Omega = B_1 \cup B_2$ and $B_1 \cap B_2 = \varnothing$. Then, by Corollary 1.45,

$$\mathbb{P}(A) = \mathbb{P}(B_1)\mathbb{P}(A|B_1) + \mathbb{P}(B_2)\mathbb{P}(A|B_2). \tag{1.7.1}$$

Note that

$$\mathbb{P}(B_1) = \frac{4}{52} = \frac{1}{13} \quad \text{and} \quad \mathbb{P}(B_2) = 1 - \mathbb{P}(B_1) = \frac{12}{13}.$$

For the conditional probabilities, we can directly compute, using Laplace's rule,

$$\mathbb{P}(A|B_1) = \frac{3}{51} \quad \text{and} \quad \mathbb{P}(A|B_1) = \frac{4}{51}.$$

If we put all these probabilities into (1.7.1), we obtain

$$\mathbb{P}(A) = \frac{1}{13} \times \frac{3}{51} + \frac{12}{13} \times \frac{4}{51} = \frac{1}{13}.$$

1.8 Bayes' theorem

To end this first chapter, we introduce one of the most useful properties of probabilities: Bayes' theorem.

Theorem 1.47 (Bayes)**.** *Consider an experiment with outcome space* Ω *and let* $\mathbb{P}$ *be a probability defined on* $\mathcal{P}(\Omega)$*. Let* $A, B \subset \Omega$ *with* $\mathbb{P}(A) > 0$ *and* $\mathbb{P}(B) > 0$*. Then*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{1.8.1}$$

Proof. If we start from the right-hand side,

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

□

We end with an example to prove its power.

Example 1.48 (COVID-19: the bad old days)**.** Suppose a screening test for COVID-19 has the following parameters: the probability that the test is positive given that you are sick (i.e., the probability of being a true positive) is

$$\mathbb{P}(\text{positive}|\text{sick}) = 0.95.$$

On the other hand, the probability that the test is positive given that you are not sick (i.e., the probability of being a false positive) is

$$\mathbb{P}(\text{positive}|\text{not sick}) = 0.01.$$

If we want to compute the probability of being sick given that you test positive, we may use Bayes' theorem:

$$\mathbb{P}(\text{sick}|\text{positive}) = \frac{\mathbb{P}(\text{positive}|\text{sick})\mathbb{P}(\text{sick})}{\mathbb{P}(\text{positive})}. \tag{1.8.2}$$

It is known that the probability you have the disease is 0.01, as 10 in every 1000 people who have tested have the disease, so $\mathbb{P}(\text{sick}) = 0.01$. Hence, the only missing quantity in the right-hand side of (1.8.2) is $\mathbb{P}(\text{positive test})$. However, using the law of total probability (Corollary 1.45), we have

$$\begin{aligned}\mathbb{P}(\text{positive}) &= \mathbb{P}(\text{sick})\mathbb{P}(\text{positive}|\text{sick}) + \mathbb{P}(\text{not sick})\mathbb{P}(\text{positive}|\text{not sick})\\ &= \mathbb{P}(\text{sick})\mathbb{P}(\text{positive}|\text{sick}) + (1 - \mathbb{P}(\text{sick}))\mathbb{P}(\text{positive}|\text{not sick}).\end{aligned}$$

That is, we can rewrite (1.8.2) as

$$\begin{aligned}\mathbb{P}(\text{sick}|\text{positive}) &= \frac{\mathbb{P}(\text{positive}|\text{sick})\mathbb{P}(\text{sick})}{\mathbb{P}(\text{sick})\mathbb{P}(\text{positive}|\text{sick}) + (1 - \mathbb{P}(\text{sick}))\mathbb{P}(\text{positive}|\text{not sick})}\\ &= \frac{0.95 \times 0.01}{0.01 \times 0.95 + (1 - 0.01) \times 0.01}\\ &\approx 0.49.\end{aligned}$$

That is, given that you tested positive, with probability 0.49 you are sick.

Chapter 2

Discrete probability distributions

In this chapter, we will study discrete random variables; that is, functions or observables that are defined on finite or at most countably infinite outcome spaces. First, we will introduce the concept of (discrete) random variables in Section 2.1. Then, in Sections 2.2 and 2.3, we will define the expectation of a random variable and see some remarkable expectations that one may compute. Finally, in Section 2.4, we provide examples of discrete random variables beyond the uniform case.

2.1 Discrete random variables

In the setting of Chapter 1, we worked with random experiments. In reality, most of the time we will only have access to some *observables* of the whole experiment. In the context of probability theory, the objects that play the role of those are random variables, which we proceed to define in the context of *discrete* outcome spaces (that is, the ones that are finite or at most countably infinite[1]).

Definition 2.1 (Random variable)**.** Consider a random experiment with a finite or countably infinite outcome space Ω. A function $X\colon \Omega \to \mathbb{R}$ is said to be a *random variable.* We define the *support of* X, denoted by $\operatorname{supp}(X)$, as the set of all values that the random variable X takes:

$$\operatorname{supp}(X) \coloneqq \{X(\omega) : \omega \in \Omega\};$$

that is, $\operatorname{supp}(X)$ is the image of the set Ω via the map X:

$$\operatorname{supp}(X) = X(\Omega).$$

Essentially, random variables are observables of our outcome space in the following sense: if $X\colon \Omega \to \mathbb{R}$ is a random variable, then for each possible state $\omega \in \Omega$, we get a real number $X(\omega) \in \mathbb{R}$. Let us provide some examples of random variables.

[1]That is, sets having the same cardinality as $\mathbb{N}$, the set of natural numbers.

Example 2.2 (Tossing a coin)**.** Suppose we are tossing a coin. Then our outcome space is

$$\Omega = \{H, T\}.$$

We can encode the outcome of the experiment using numbers in the following way: let $X\colon \Omega \to \mathbb{R}$ be the random variable defined by

$$\forall \omega \in \Omega, \quad X(\omega) = \begin{cases} 0, & \text{if } \omega = H \\ 1, & \text{if } \omega = T. \end{cases}$$

Note that $\operatorname{supp}(X) = \{0, 1\}$.

Example 2.3 (Tossing coins)**.** Suppose we are tossing two coins, so our outcome space is

$$\Omega = \{HH, HT, TT, TH\}.$$

We define $X\colon \Omega \to \mathbb{R}$ by

$$\forall \omega \in \Omega, \quad X(\omega) \coloneqq \text{number of heads obtained in } \omega.$$

That is,

$$X(\omega) = \begin{cases} 0, & \text{if } \omega = TT \\ 1, & \text{if } \omega = HT \text{ or } \omega = TH \\ 2, & \text{if } \omega = HH. \end{cases}$$

In this case, $\operatorname{supp}(X) = \{0, 1, 2\}$.

Example 2.4 (Rolling dice)**.** Suppose you roll two fair standard dice. Hence, the sample space can be written as

$$\Omega = \{(n_1, n_2) : n_1, n_2 \in \{1, 2, 3, 4, 5, 6\}\}.$$

Then we can define $X_1\colon \Omega \to \mathbb{R}$ as the sum of both numbers appearing in the two rolls:

$$\forall (n_1, n_2) \in \Omega, \quad X_1(n_1, n_2) \coloneqq n_1 + n_2.$$

Another example $X_2\colon \Omega \to \mathbb{R}$ is the mean between both numbers:

$$\forall (n_1, n_2) \in \Omega, \quad X_2(n_1, n_2) \coloneqq \frac{n_1 + n_2}{2}.$$

A final example can be just giving the number of the first die: we define $X_3\colon \Omega \to \mathbb{R}$ by

$$\forall (n_1, n_2) \in \Omega, \quad X_3(n_1, n_2) \coloneqq n_1.$$

Note that the supports are different:

$$\operatorname{supp}(X_1) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\},$$

$$\operatorname{supp}(X_2) = \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6\},$$
$$\operatorname{supp}(X_3) = \{1, 2, 3, 4, 5, 6\}.$$

Remark 2.5. The support of a discrete random variable is always a discrete set: that is, it is finite or countably infinite.

Naturally, given a random variable $X\colon \Omega \to \mathbb{R}$, we are interested in determining, for $x \in \operatorname{supp}(X)$, the quantity $\mathbb{P}(X = x)$, which we define as the probability of the event

$$\{\omega \in \Omega : X(\omega) = x\} \subset \Omega.$$

An object that encodes those probabilities is the probability mass function associated with X.

Definition 2.6 (Probability mass function)**.** Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. For a random variable $X\colon \Omega \to \mathbb{R}$ with support $\operatorname{supp}(X)$, we define the *probability mass function* or *PMF* associated with X as the function $f_X\colon \operatorname{supp}(X) \to \mathbb{R}$ given by

$$\forall x \in \operatorname{supp}(X), \quad f_X(x) = \mathbb{P}(X = x).$$

When $x \notin \operatorname{supp}(X)$, we define $f_X(x) \coloneqq 0$.

We note some immediate properties associated with a PMF.

Proposition 2.7. *Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with support* $\operatorname{supp}(X)$ *and PMF f_X. Then:*

(i) For every $x \in \operatorname{supp}(X)$, $f_X(x) \in [0, 1]$.

(ii) For every $A \subset \operatorname{supp}(X)$, $\mathbb{P}(X \in A) = \sum_{x\in A} f_X(x)$.

(iii) $\sum_{x\in\operatorname{supp}(X)} f_X(x) = 1$.

Proof.

(i) Direct from the fact that $\mathbb{P}$ takes values in $[0, 1]$.

(ii) Note that for $x \neq y$, we have $\{X = x\} \cap \{X = y\} = \varnothing$. Therefore the collection $(\{X = x\})_{x\in A}$ is exhaustive. On the other hand, note that

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

Hence

$$\mathbb{P}(X \in A) = \mathbb{P}\left(\bigcup_{x\in A} \{X = x\}\right) = \sum_{x\in A} \mathbb{P}(X = x) = \sum_{x\in A} f_X(x).$$

(iii) It is a direct consequence of (iv) by taking $A = \operatorname{supp}(X)$ and the fact that $\mathbb{P}(X \in \operatorname{supp}(X)) = 1$.

$\square$

In the previous paragraph, we started from a discrete random variable $X\colon \Omega \to \mathbb{R}$ with support $\operatorname{supp}(X)$ and constructed its PMF as a function $f_X\colon \operatorname{supp}(X) \to \mathbb{R}$. We can proceed reversely, starting from a discrete set $A \subset \mathbb{R}$ and a function $f\colon A \to \mathbb{R}$. The question is, under what conditions can such a function be the PMF of a random variable? This discussion motivates the following definition.

Definition 2.8 (Probability density function)**.** Let $A \subset \mathbb{R}$ be a discrete set and let $f\colon A \to \mathbb{R}$ be a function. We say that f is a *probability density function* if:

(i) For every $x \in A$, $f(x) \in [0,1]$.

(ii) $\sum_{x\in A} f(x) = 1$.

Remark 2.9. For every discrete random variable $X\colon \Omega \to \mathbb{R}$, its PMF $f_X\colon \operatorname{supp}(X) \to \mathbb{R}$ is a probability density function in the sense of Definition 2.8.

Another important object is the cumulative distribution function.

Definition 2.10 (Cumulative distribution function)**.** Consider a random experiment with a finite or countably infinite outcome space Ω, and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable. We define the *cumulative distribution function* or *CDF* as the function $F_X\colon \mathbb{R} \to [0,1]$ defined by

$$\forall x \in \mathbb{R}, \quad F_X(x) \coloneqq \mathbb{P}(X \leqslant x).$$

The CDF of a random variable satisfies the following properties.

Proposition 2.11. *Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with support* $\operatorname{supp}(X)$*, PMF f_X, and CDF F_X. Then:*

(i) For every $x \in \mathbb{R}$,

$$F_X(x) = \sum_{\substack{y\in \operatorname{supp}(X)\\ y\leqslant x}} f_X(y).$$

(ii) F_X is a non-decreasing function.

(iii) $\lim_{x\to-\infty} F_X(x) = 0.$

(iv) $\lim_{x\to+\infty} F_X(x) = 1.$

Our first example of a distribution, which we have already previously encountered, is the uniform one.

Definition 2.12 (Uniform distribution)**.** Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with PMF f_X. We say that X is *uniformly distributed* if f_X is a constant function.

Remark 2.13. If X is uniformly distributed, and if $\operatorname{supp}(X) = \{1, 2, \dots, m\}$ for some $m \in \mathbb{N}^*$, then

$$\forall x \in \operatorname{supp}(X), \quad f_X(x) = \frac{1}{m}.$$

In this case,

$$F_X(x) = \begin{cases} 0, & \text{if } x < 1 \\ \frac{k}{m}, & \text{if } k \leqslant x < k+1 \text{ for } k \in \{1, \dots, m-1\} \\ 1, & \text{if } x \geqslant m; \end{cases}$$

that is, F_X is a step function.

Example 2.14. Let X be a random variable uniformly distributed on $\{1, 2, 3, 4, 5, 6\}$ (it could represent the number obtained when tossing a fair die, for example). Then

$$\forall x \in \{1, 2, 3, 4, 5, 6\}, \quad f_X(x) = \frac{1}{6}$$

and

$$F_X(x) = \begin{cases} 0, & \text{if } x < 1 \\ \frac{k}{6}, & \text{if } k \leqslant x < k+1 \text{ for } k \in \{1, 2, 3, 4, 5\} \\ 1, & \text{if } x \geqslant 6. \end{cases}$$

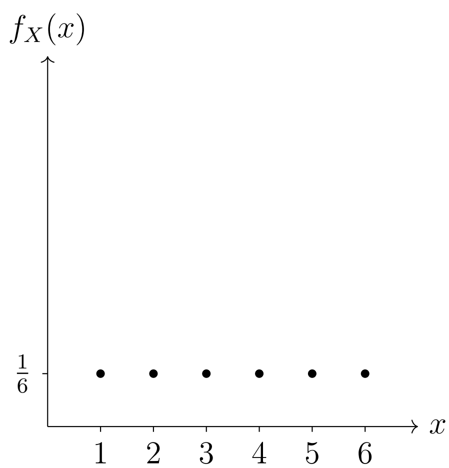

Figure 2.1: PMF of X.

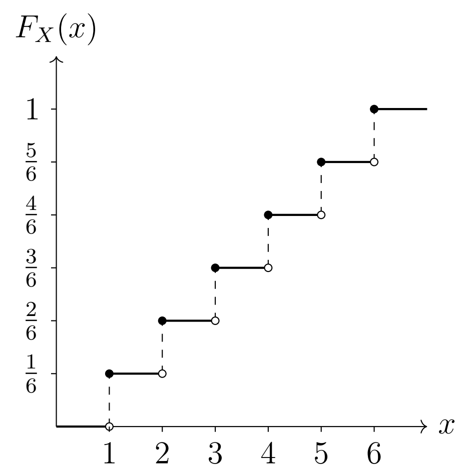

Figure 2.2: CDF of X.

2.2 Expectation of a discrete random variable

Given a random experiment $(\Omega, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, one may wonder about the "most probable value" taken by X according to the probability measure $\mathbb{P}$, which should be a deterministic quantity; i.e., not depending on $\omega \in \Omega$. This motivates the definition of mathematical expectation.

Definition 2.15 (Expectation)**.** Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with PMF f_X. We define the *expectation* or *expected value* of X, which we denote by $\mathbb{E}[X]$, as

$$\mathbb{E}[X] \coloneqq \sum_{x\in\operatorname{supp}(X)} x\,\mathbb{P}(X=x) = \sum_{x\in\operatorname{supp}(X)} x\,f_X(x).$$

That is, **the expected value of a random value is no more than the weighted mean of all the values that X takes according to their probability**.

Example 2.16. Recall Example 2.3. If we suppose that both coins are fair, then the expectation of X can be computed as

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\operatorname{supp}(X)} x\,\mathbb{P}(X=x) = \sum_{x=0}^{2} x\,\mathbb{P}(X=x)\\ &= 0\times\mathbb{P}(X=0)+1\times\mathbb{P}(X=1)+2\times\mathbb{P}(X=2)\\ &= 0\times\frac{1}{4}+1\times\frac{2}{4}+2\times\frac{1}{4}\\ &= 1.\end{aligned}$$

Remark 2.17. If $X\colon\Omega\to\mathbb{R}$ is a discrete random variable with support $\operatorname{supp}(X)$ and PMF f_X, and $u\colon\mathbb{R}\to\mathbb{R}$ is a function, then their composition $u(X)$ is a random variable $u(X)\colon\Omega\to\mathbb{R}$. We can compute its expectation as

$$\mathbb{E}[u(X)] = \sum_{x\in\operatorname{supp}(X)} u(x)\,\mathbb{P}(X=x) = \sum_{x\,\operatorname{supp}(X)} u(x)\,f_X(x).$$

Remark 2.18. Note that when $\operatorname{supp}(X)$ is finite, $\mathbb{E}[u(X)]$ is just a sum, so it is always well-defined and finite. However, when $\operatorname{supp}(X)$ is countably infinite, we have to require that

$$\sum_{x\in\operatorname{supp}(X)} |u(x)|\,f_X(x) < +\infty$$

for $\mathbb{E}[u(X)]$ to be finite.

Example 2.19. Let $X\colon\Omega\to\mathbb{R}$ be a uniform random variable with $\operatorname{supp}(X)=\{1,2,3\}$ and let $u\colon\mathbb{R}\to\mathbb{R}$ be the function $u(x)=x^2$. Then

$$\mathbb{E}[u(X)] = \sum_{x\in\operatorname{supp}(X)} u(x)\,\mathbb{P}(X=x) = \sum_{x=1}^{3} x^2\,\mathbb{P}(X=x) = 1^2\times\frac{1}{3}+2^2\times\frac{1}{3}+3^2\times\frac{1}{3}=\frac{14}{3}.$$

We state some useful properties of expectation.

Proposition 2.20. *Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Then:*

(i) The expectation of a constant is the constant itself: if $c \in \mathbb{R}$ is a deterministic value, then

$$\mathbb{E}[c] = c.$$

(ii) The expectation is a linear operator*:*

(a) If $X\colon \Omega \to \mathbb{R}$ is a random variable and $c \in \mathbb{R}$ a deterministic constant, then

$$\mathbb{E}[cX] = c\,\mathbb{E}[X].$$

(b) If $X, Y\colon \Omega \to \mathbb{R}$ are random variables, then

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

Equivalently, we have that for any random variables $X, Y\colon \Omega \to \mathbb{R}$ and any deterministic constants $c_1, c_2 \in \mathbb{R}$,

$$\mathbb{E}[c_1 X + c_2 Y] = c_1\,\mathbb{E}[X] + c_2\,\mathbb{E}[Y].$$

(iii) Let $X\colon \Omega \to \mathbb{R}$ be a random variable, let $u_1, u_2\colon \mathbb{R} \to \mathbb{R}$, and let $c_1, c_2 \in \mathbb{R}$. Then

$$\mathbb{E}[c_1 u_1(X) + c_2 u_2(X)] = c_1\,\mathbb{E}[u_1(X)] + c_2\,\mathbb{E}[u_2(X)].$$

Exercise 2.1. Let X be a random variable having support $\operatorname{supp}(X) = \{1, 2, 3, 4\}$ and PMF given by $f_X(x) = \frac{x}{10}$ for $x \in \operatorname{supp}(X)$.

(i) Prove that f_X is well defined, that is, show that

$$\sum_{x \in \operatorname{supp}(X)} f_X(x) = 1.$$

(ii) Determine its CDF F_X.

(iii) Compute $\mathbb{E}[X]$.

(iv) Compute $\mathbb{E}[X^2]$.

(v) Compute $\mathbb{E}[X(5 - X)]$.

2.3 Remarkable expectations

In the last section, in Remark 2.17, we defined the expectation of a function $u\colon \mathbb{R} \to \mathbb{R}$ of a discrete random variable $X\colon \Omega \to \mathbb{R}$. In this section, we emphasize some particular choices of u that will yield relevant parameters of the random variable X.

2.3.1 Mean

When $u(x) = x$, it corresponds to μ_X, the *mean* of X:

$$\mu_X \coloneqq \mathbb{E}[u(X)] = \mathbb{E}[X] = \sum_{x \in \operatorname{supp}(X)} x\, f_X(x).$$

2.3.2 Variance

When $u(x) = (x - \mu_X)^2$, it corresponds to σ_X^2, the *variance* of X:

$$\sigma_X^2 \coloneqq \mathbb{E}[u(X)] = \mathbb{E}[(X - \mu_X)^2] = \sum_{x \in \operatorname{supp}(X)} (x - \mu_X)^2 f_X(x).$$

We remark that also,

$$\sigma_X^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \tag{2.3.1}$$

We define the *standard deviation* of X as $\sigma_X \coloneqq \sqrt{\sigma_X^2}$.

2.3.3 Moments

When $u(x) = x^r$ for $r \in \mathbb{N}^*$, it corresponds to the *r-th moment* of X:

$$\mathbb{E}[u(X)] = \mathbb{E}[X^r] = \sum_{x \in \operatorname{supp}(X)} x^r f_X(x).$$

2.3.4 Moment-generating function

Now we consider a family of such functions: for $t \in \mathbb{R}$, define the function $u_t(x) = e^{tx}$ and let

$$M_X(t) \coloneqq \mathbb{E}[u_t(X)] = \mathbb{E}[e^{tX}] = \sum_{x \in \operatorname{supp}(X)} e^{tx} f_X(x).$$

We say that the function $M_X \colon \mathbb{R} \to \mathbb{R}$ given by $t \mapsto M_X(t)$ is the *moment-generating function* or the *Laplace transform* of X. Its importance resides in the following fact, which is a corollary of the theory of Laplace's transform: the moment-generating function characterizes the distribution of a random variable.

Theorem 2.21. *Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X, Y \colon \Omega \to \mathbb{R}$ with $\operatorname{supp}(X) = \operatorname{supp}(Y)$. Then $f_X = f_Y$ if and only if $M_X = M_Y$.*

That is, **two random variables with the same support and moment-generating function have the same distribution**.

Another remarkable property of the moment-generating function is that it helps compute the moments.

Proposition 2.22. *Consider a random experiment with a finite or countably infinite outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with moment-generating function M_X. Then*

$$\mathbb{E}[X^r] = M_X^{(r)}(0),$$

where $M_X^{(r)}$ denotes the r-th derivative of the function M_X. In particular,

$$\mu_X = M_X'(0) \quad \text{and} \quad \sigma_X^2 = M_X''(0) - (M_X'(0))^2.$$

2.4 More examples of discrete distributions

The purpose of this section is twofold: first, we will introduce many discrete distributions, beyond the uniform one. They appear naturally when we model some random experiments. Our second goal will be to apply the concepts defined in the last sections. Hence, we will compute those distributions' expectation, variance, and moment-generating function.

Along this section, we consider a discrete outcome space Ω endowed with a probability $\mathbb{P}$.

2.4.1 Uniform distribution

We defined the uniform distribution previously in Definition 2.12. However, for the sake of completeness, we recall its definition here.

Definition 2.23 (Uniform distribution)**.** Let $m \in \mathbb{N}^*$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to be *uniform with support* $\{1, \dots, m\}$, which we denote by $X \sim \mathrm{U}(\{1, \dots, m\})$, if its support is

$$\operatorname{supp}(X) = \{1, \dots, m\}$$

and its PMF is given by

$$\forall k \in \{1, \dots, m\}, \quad f_X(k) = \frac{1}{m}.$$

Proposition 2.24. *Let $m \in \mathbb{N}^*$ and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \mathrm{U}(\{1, \dots, m\})$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = \frac{m+1}{2} \quad \text{and} \quad \sigma_X^2 = \frac{m^2-1}{12}.$$

(ii) Its moment-generating function is given by

$$\forall t \in \mathbb{R}, \quad M_X(t) = \begin{cases} \frac{e^t(1-e^{mt})}{m(1-e^t)}, & \text{if } t \neq 0 \\ 1, & \text{if } t = 0. \end{cases}$$

Exercise 2.2. One may also consider uniform random variables on more general intervals as $\{a, a+1, \ldots, b-1, b\}$ for some $a, b \in \mathbb{Z}$ with $b \geqslant a$.

(i) Determine the associated PMF.

(ii) Compute their mean and variance.

(iii) Determine their moment-generating function.

2.4.2 Bernoulli distribution

The Bernoulli distribution is one of the simplest we will encounter. Suppose we observe a random experiment with only two possible outcomes: *success* or *failure*. Moreover, let us suppose that there exists $p \in (0,1)$ such that

$$\mathbb{P}(\text{success}) = p \quad \text{and} \quad \mathbb{P}(\text{failure}) = 1 - p.$$

We may encode the outcome of the experiment via the following random variable:

$$X := \begin{cases} 0, & \text{if failure} \\ 1, & \text{if success.} \end{cases}$$

Then we say that X has the Bernoulli distribution with parameter p.

Definition 2.25 (Bernoulli distribution)**.** Let $p \in (0,1)$. A random variable $X \colon \Omega \to \mathbb{R}$ is said to have the *Bernoulli distribution with parameter* p, which we denote by $X \sim \mathrm{Ber}(p)$, if its support is

$$\mathrm{supp}(X) = \{0, 1\}$$

and its PMF is given by

$$f_X(0) = 1 - p \quad \text{and} \quad f_X(1) = p.$$

Proposition 2.26. *Let $p \in (0,1)$ and let $X \colon \Omega \to \mathbb{R}$ be a random variable with $X \sim$ $\mathrm{Ber}(p)$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = p \quad \textit{and} \quad \sigma_X^2 = p(1-p).$$

(ii) Its moment-generating function is given by

$$\forall t \in \mathbb{R}, \quad M_X(t) = 1 - p + pe^t.$$

2.4.3 Binomial distribution

Recall the setting of Section 2.4.2: we have an experiment such that there exists $p \in (0,1)$ such that

$$\mathbb{P}(\text{success}) = p \quad \text{and} \quad \mathbb{P}(\text{failure}) = 1 - p.$$

Now, suppose that we can independently repeat it n times. Define X as the random variable that counts the total number of successes among the n repetitions, which will be an integer between 0 and n. For each $1 \leqslant i \leqslant n$, let X_i be the random variable defined as

$$X_i := \begin{cases} 0, & \text{if the } i\text{-th repetition is a failure} \\ 1, & \text{if the } i\text{-th repetition is a success.} \end{cases}$$

Then

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

so we know that X is characterized as the sum of n independent Bernoulli random variables with parameter p.

Now, for $k \in \{0, \dots, n\}$, let us compute $\mathbb{P}(X = k)$, which is the probability of obtaining exactly k successes. We have $\binom{n}{k}$ different ways to choose the k individual successes amongst the n trials. Additionally, note that the probability of obtaining exactly k successes amongst n trials is

$$p^k (1-p)^{n-k}$$

since we have to succeed k times and fail $(n-k)$ times. Then

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

Definition 2.27 (Binomial distribution)**.** Let $p \in (0,1)$ and $n \in \mathbb{N}^*$. A random variable $X \colon \Omega \to \mathbb{R}$ is said to have the *Binomial distribution with parameters n and p*, which we denote by $X \sim \operatorname{Bin}(n, p)$, if its support is

$$\operatorname{supp}(X) = \{0, \dots, n\}$$

and its PMF is given by

$$\forall k \in \{0, \dots, n\}, \quad f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Remark 2.28. Note that f_X is well-defined since

$$\sum_{k \in \operatorname{supp}(X)} f_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1$$

thanks to Newton's binomial theorem.

Theorem 2.29 (Newton's binomial theorem)**.** *Let $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$. Then*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}.$$

Remark 2.30. We have $\mathrm{Bin}(1, p) = \mathrm{Ber}(p)$.

Proposition 2.31. *Let $p \in (0,1)$ and $n \in \mathbb{N}^*$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \mathrm{Bin}(n,p)$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = np \quad \textit{and} \quad \sigma_X^2 = np(1-p).$$

(ii) Its moment-generating function is given by

$$\forall t \in \mathbb{R}, \quad M_X(t) = (1 - p + pe^t)^n.$$

2.4.4 Geometric distribution

As before, suppose we observe an experiment such that there exists $p \in (0,1)$ such that

$$\mathbb{P}(\text{success}) = p \quad \text{and} \quad \mathbb{P}(\text{failure}) = 1 - p.$$

We repeat the experiment independently until we get a success, and let X be the random variable counting the number of repetitions needed, considering the final success. Note that X takes values in $\mathbb{N}^*$ and that for any $k \in \mathbb{N}^*$,

$$\mathbb{P}(X = k) = (1-p)^{k-1} p$$

since we have to get exactly $k-1$ failures (each having probability $1-p$) before the succesful repetition that has probability p.

Definition 2.32 (Geometric distribution)**.** Let $p \in (0,1)$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to have the *Geometric distribution with parameter* p, which we denote by $X \sim \mathrm{Geom}(p)$, if its support is

$$\mathrm{supp}(X) = \mathbb{N}^*$$

and its PMF is given by

$$\forall k \in \mathbb{N}^*, \quad f_X(k) = (1-p)^{k-1} p.$$

Remark 2.33. Note that f_X is well-defined since

$$\sum_{k \in \mathrm{supp}(X)} f_X(k) = \sum_{k=1}^{+\infty} (1-p)^{k-1} p = p \sum_{k=1}^{+\infty} (1-p)^{k-1} = p \sum_{k=0}^{+\infty} (1-p)^k = p \left(\frac{1}{1-(1-p)} \right) = 1.$$

Proposition 2.34. *Let $p \in (0,1)$ and let $X \colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \mathrm{Geom}(p)$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = \frac{1}{p} \quad \text{and} \quad \sigma_X^2 = \frac{1-p}{p^2}.$$

(ii) Its moment-generating function is given by

$$\forall t < -\log(1-p), \quad M_X(t) = \frac{pe^t}{1-(1-p)e^t}.$$

2.4.5 Poisson distribution

A typical phenomenon we would like to model is the number of arrivals or occurrences of an event at a given interval of time, such as the number of customers arriving at a store during a day or the number of texts we receive each day. Our goal will be to define a probability density that fits those models under the following assumptions:

(i) The numbers of occurrences in two nonoverlapping time intervals are independent.

(ii) There exists $\lambda > 0$ such that if the time interval has length $t > 0$, then

$$\mathbb{P}(\text{observe exactly one occurrence of the event}) \approx \lambda t$$

for small values of t.

(iii) If $t > 0$ is sufficiently small, then the probability of observing two or more occurrences in an interval of length t is zero.

Let X be the random variable that counts the number of occurrences in the interval $[0,1]$, say. Let $k \in \mathbb{N}$ and let us approximate $\mathbb{P}(X = k)$. We can divide the interval $[0,1]$ into n nonoverlapping intervals of length $1/n$:

$$I_1 = [0, 1/n), I_2 = [1/n, 2/n), \ldots, I_n = [1 - 1/n, 1],$$

and denote by X_i the number of occurrences of the event in the interval I_i. If n is sufficiently big, then, in light of assumption (iii), we can think of X_i as a Bernoulli random variable, which, due to assumption (ii), should have parameter $\frac{\lambda}{n}$. According to assumption (i), the random variables $(X_i)_{i=1}^n$ are independent. Since

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

we can think of X as a Binomial random variable of parameters n and $\frac{\lambda}{n}$. That is,

$$\mathbb{P}(X = k) \approx \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

When $n \to +\infty$, we obtain

$$\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \xrightarrow[n\to+\infty]{} e^{-\lambda}\frac{\lambda^k}{k!}.$$

Definition 2.35 (Poisson distribution)**.** Let $\lambda > 0$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to have the *Poisson distribution with parameter* λ, which we denote by $X \sim \mathrm{Poi}(\lambda)$, if its support is

$$\mathrm{supp}(X) = \mathbb{N}$$

and its PMF is given by

$$\forall k \in \mathbb{N}, \quad f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Remark 2.36. Note that f_X is well-defined since

$$\sum_{k\in\mathrm{supp}(X)} f_X(k) = \sum_{k=0}^{+\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1.$$

Proposition 2.37. *Let $\lambda > 0$ and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \mathrm{Poi}(\lambda)$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = \lambda \quad \textit{and} \quad \sigma_X^2 = \lambda.$$

(ii) Its moment-generating function is given by

$$\forall t \in \mathbb{R}, \quad M_X(t) = \exp(\lambda(e^t - 1)).$$

Chapter 3

Continuous probability distributions

In this chapter, we will extend some of the notions we have introduced in Chapter 2 towards the continuous case. That is, we will consider random variables that have a continuum of possible outcomes. We begin in Sections 3.1 and 3.2 by extending the notions of random variable, probability density function, CDF, expectation, variance, etc., towards the continuous case. In Section 3.3 we provide further examples of continuous distributions. Finally, in Section 3.4, we discuss the distribution of a function of a random variable.

3.1 Continuous random variables

We start by defining random variables, objects that (again) capture the idea of the observables associated with a random experiment. Note that we do not impose any restriction on the cardinality of the outcome space here.

Definition 3.1 (Random variable)**.** Consider a random experiment with outcome space Ω. A function $X\colon \Omega \to \mathbb{R}$ is said to be a *random variable*. We define the *support of* X, denoted by $\operatorname{supp}(X)$, as the set of all values that the random variable X takes:

$$\operatorname{supp}(X) \coloneqq \{X(\omega) : \omega \in \Omega\};$$

that is, $\operatorname{supp}(X)$ is the image of the set Ω via the map X:

$$\operatorname{supp}(X) = X(\Omega).$$

Let us illustrate the previous definition with a simple example in the continuous setting.

Example 3.2 (Sampling from an interval)**.** Suppose we are sampling a single point from the interval $[a, b]$ with $a < b$, so $\Omega = [a, b]$. We can define the random variable $X\colon \Omega \to \mathbb{R}$

that returns the sampled point:

$$\forall \omega \in \Omega, \quad X(\omega) = \omega.$$

Its support is $\operatorname{supp}(X) = [a, b]$.

Given a random variable $X \colon \Omega \to \mathbb{R}$, suppose we endow the outcome space with a probability $\mathbb{P}$. As we did in the previous chapter, we would be interested in computing the probability of X belonging to a determined event. However, in the continuous setting, for a random variable $X \colon \Omega \to \mathbb{R}$, for any point $x \in \operatorname{supp}(X)$, we have that

$$\mathbb{P}(X = x) = 0.$$

Thus, we must follow a different approach to extend Definition 2.8 towards the continuous setting.

Definition 3.3 (Probability density function)**.** Consider a random experiment with outcome space Ω, and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X \colon \Omega \to \mathbb{R}$ be a random variable with support $\operatorname{supp}(X)$. We say that a function $f_X \colon \operatorname{supp}(X) \to \mathbb{R}$ is the *probability density function* or just the *density* associated with X if:

(i) For every $x \in \operatorname{supp}(X)$, $f_X(x) \geqslant 0$.

(ii) For every $x \notin \operatorname{supp}(X)$, $f_X(x) = 0$

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

(iv) For every $A \subset \operatorname{supp}(X)$,

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,\mathrm{d}x.$$

Equivalently, for every $a < b$ such that $(a, b) \subset \operatorname{supp}(X)$,

$$\mathbb{P}(X \in (a, b)) = \int_a^b f_X(x)\,\mathrm{d}x.$$

We also define the cumulative distribution function associated with a random variable. Note that the following definition does not differ from Definition 2.10.

Definition 3.4 (Cumulative distribution function)**.** Consider a random experiment with outcome space Ω, and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X \colon \Omega \to \mathbb{R}$ be a random variable. We define the *cumulative distribution function* or *CDF* as the function $F_X \colon \mathbb{R} \to [0, 1]$ defined by

$$\forall x \in \mathbb{R}, \quad F_X(x) \coloneqq \mathbb{P}(X \leqslant x).$$

The CDF of a random variable satisfies the following properties.

Proposition 3.5. *Consider a random experiment with outcome space* Ω *and let* $\mathbb{P}$ *be a probability defined on* $\mathcal{P}(\Omega)$*. Let* $X\colon \Omega \to \mathbb{R}$ *be a random variable with support* $\operatorname{supp}(X)$*, PMF* f_X*, and CDF* F_X*. Then:*

(i) *For every* $x \in \mathbb{R}$,
$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y.$$

(ii) *For every* $x \in \mathbb{R}$,
$$F'_X(x) = f_X(x).$$

(iii) F_X *is a non-decreasing function.*

(iv) $\lim_{x\to-\infty} F_X(x) = 0$.

(v) $\lim_{x\to+\infty} F_X(x) = 1$.

Example 3.6 (Uniform random variable)**.** Recall Example 3.2. If we sample the point uniformly (i.e., for every two intervals $A_1, A_2 \subset [a,b]$ of length $r > 0$, we have that $\mathbb{P}(X \in A_1) = \mathbb{P}(X \in A_2)$), then a probability density function for X is the uniform one:
$$\forall x \in [a,b], \quad f_X(x) = \frac{1}{b-a},$$
and we note this by $X \sim \mathrm{U}([a,b])$. Then its CDF is given by
$$\forall x \in \mathbb{R}, \quad F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leqslant x < b \\ 1, & \text{if } x \geqslant b. \end{cases}$$

We define the percentile values associated with a random variable.

Definition 3.7 (Percentile values)**.** Consider a random experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with density f_X and CDF $F_X\colon \mathbb{R} \to [0,1]$. For $p \in (0,1)$, we define the *p-th percentile* as a number $x_p \in \mathbb{R}$ such that
$$p = F_X(x_p) = \int_{-\infty}^{x_p} f_X(y)\,\mathrm{d}y.$$
When $p = 0.5$, we call it a *median*. When $p = 0.25$, $p = 0.5$, and $p = 0.75$, we call them the first, second, and third quartiles, respectively.

3.2 Expectation of a continuous random variable

Similarly to what we did in Chapter 2, we define a continuous random variable's expectation, variance, and moment-generating function by just replacing summation with integration.

Definition 3.8 (Expectation, variance, standard deviation, moments, and moment-generating function)**.** Consider a random experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with density f_X.

(i) We define the *expectation*, *expected value*, or *mean* of X, which we denote by $\mathbb{E}[X]$ or μ_X, as

$$\mathbb{E}[X] \coloneqq \int_{-\infty}^{+\infty} x\, f_X(x)\,\mathrm{d}x.$$

(ii) We define the *variance* of X, which we denote by σ_X^2, as

$$\sigma_X^2 \coloneqq \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x.$$

Note that, as in the discrete case,

$$\sigma_X^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)\,\mathrm{d}x - \left(\int_{-\infty}^{+\infty} x\, f_X(x)\,\mathrm{d}x\right)^2.$$

(iii) We define the *standard deviation* of X as

$$\sigma_X \coloneqq \sqrt{\sigma_X^2}.$$

(iv) For $r \in \mathbb{N}^*$, we define the *r-th moment* of X as

$$\mathbb{E}[X^r] \coloneqq \int_{-\infty}^{+\infty} x^r\, f_X(x)\,\mathrm{d}x.$$

(v) We define the *moment-generating function* of X as the function $M_X(t)\colon \mathbb{R} \to \mathbb{R}$ given by

$$\forall t \in \mathbb{R}, \quad M_X(t) \coloneqq \int_{-\infty}^{+\infty} e^{tx} f_X(x)\,\mathrm{d}x.$$

Remark 3.9. More generally, if $u\colon \mathbb{R} \to \mathbb{R}$ is a function, we may compute the expectation of the composition $u(X)\colon \Omega \to \mathbb{R}$ similarly to the discrete case (cf. Remark 2.17). That is,

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

Remark 3.10. The properties related to the objects we have just defined that were valid in the discrete case are still valid mutatis mutandis in the continuous setting. Namely, Proposition 2.20, equation (2.3.1), Theorem 2.21, and Proposition 2.22 are still true with the obvious modifications.

Exercise 3.1. Compute the mean, median, quartiles, variance, and moment-generating function of a uniform random variable X on the interval $[a, b]$ for $a < b$ (recall Example 3.6).

3.3 Examples of continuous distributions

Here we exhibit some of the most important examples of continuous distributions beyond the uniform one (cf. Example 3.6 and Exercise 3.1).

3.3.1 Exponential distribution

An important example is the exponential distribution.

Definition 3.11 (Exponential distribution)**.** Let $\lambda > 0$. A random variable $X : \Omega \to \mathbb{R}$ is said to have the *Exponential distribution with parameter* λ, which we denote by $X \sim \operatorname{Exp}(\lambda)$, if its support is

$$\operatorname{supp}(X) = \mathbb{R}_+$$

and its density is given by

$$\forall x \in \mathbb{R}_+, \quad f_X(x) = \lambda e^{-\lambda x}.$$

Remark 3.12. Note that f_X is well-defined since

$$\int_0^{+\infty} f_X(x)\,\mathrm{d}x = \lambda \int_0^{+\infty} e^{-\lambda x}\,\mathrm{d}x = \lambda \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^{+\infty} \mathrm{d}x = \lambda \left[\frac{1}{\lambda} - 0\right] = 1.$$

Proposition 3.13. *Let $\lambda > 0$ and let $X : \Omega \to \mathbb{R}$ be a random variable with $X \sim \operatorname{Exp}(\lambda)$. Then:*

(i) Its CDF is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0. \end{cases}$$

(ii) Its expectation and variance are given by

$$\mu_X = \frac{1}{\lambda} \quad \textit{and} \quad \sigma_X^2 = \frac{1}{\lambda^2}.$$

(iii) Its moment-generating function is given by

$$\forall t < \lambda, \quad M_X(t) = \frac{\lambda}{\lambda - t}.$$

(iv) It has the memoryless property:

$$\forall t, s \geqslant 0, \quad \mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s).$$

3.3.2 Gamma distribution

We continue with the Gamma distribution. First of all, we recall the definition of the *Gamma funtion*:

$$\forall \alpha > 0, \quad \Gamma(\alpha) \coloneqq \int_0^{+\infty} y^{\alpha-1} e^{-y} \,\mathrm{d}y.$$

Definition 3.14 (Gamma distribution)**.** Let $\alpha, \theta > 0$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to have the *Gamma distribution with parameters* α *and* θ, which we denote by $X \sim \Gamma(\alpha, \theta)$, if its support is

$$\operatorname{supp}(X) = \mathbb{R}_+$$

and its density is given by

$$\forall x \in \mathbb{R}_+, \quad f_X(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}.$$

Exercise 3.2. Prove that the Gamma distribution is well-defined.

Remark 3.15. Note that the exponential is a particular case of the Gamma distribution: for all $\lambda > 0$,

$$\operatorname{Exp}(\lambda) = \Gamma(1, 1/\lambda).$$

Proposition 3.16. *Let $\alpha, \theta > 0$ and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \Gamma(\alpha, \theta)$. Then:*

(i) Its expectation and variance are given by

$$\mu_X = \alpha\theta \quad \textit{and} \quad \sigma_X^2 = \alpha\theta^2.$$

(ii) Its moment-generating function is given by

$$\forall t < \frac{1}{\theta}, \quad M_X(t) = (1 - \theta t)^{-\alpha}.$$

A particular case of the Gamma distribution that plays an important role in statistics is the chi-squared distribution.

Definition 3.17 (Chi-squared distribution)**.** Let $r \in \mathbb{N}^*$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to have the *chi-squared distribution with parameter* r, which we denote by $X \sim \chi^2(r)$, if it has the Gamma distribution with parameters 2 and $\frac{r}{2}$:

$$\chi^2(r) = \Gamma(2, r/2).$$

Exercise 3.3. Let $r \in \mathbb{N}^*$ and let $X \sim \chi^2(r)$. Determine the density, mean, variance, and moment-generating function of X.

3.3.3 Normal distribution

We arrive to one of the most important and ubiquitous[1] distributions in probability theory: the Normal distribution.

Definition 3.18 (Normal distribution)**.** Let $\mu, \sigma > 0$. A random variable $X\colon \Omega \to \mathbb{R}$ is said to have the *Normal* or *Gaussian distribution with parameters μ and σ*, which we denote by $X \sim \mathcal{N}(\mu, \sigma^2)$, if its support is

$$\operatorname{supp}(X) = \mathbb{R}$$

and its density is given by

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

If $\mu = 0$ and $\sigma = 1$, we say that it has a *Normal standard distribution.*

Remark 3.19. The Gaussian distribution is well-defined. Let

$$I := \int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x.$$

If we use the change of variables

$$z = \frac{x-\mu}{\sigma},$$

we obtain

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\,\mathrm{d}z.$$

On the other hand, note that

$$\begin{aligned} I^2 &= \frac{1}{2\pi}\left(\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\,\mathrm{d}z\right)^2 = \frac{1}{2\pi}\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x\right)\left(\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}}\,\mathrm{d}y\right) \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

If we put the change of variables $x = r\cos(\theta)$ and $y = r\sin(\theta)$, we get

$$I^2 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{+\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r\,\mathrm{d}\theta = \int_0^{+\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r = \int_0^{+\infty} e^{-s}\,\mathrm{d}s = [-e^{-s}]_0^{+\infty} = 1,$$

where we used the change of variable $s = \frac{r^2}{2}$. That is, $I^2 = 1$ hence $I = 1$.

Proposition 3.20. *Let $\mu, \sigma > 0$ and let $X\colon \Omega \to \mathbb{R}$ be a random variable with $X \sim \mathcal{N}(\mu, \sigma^2)$. Then:*

[1]For reasons that will become clear in Chapter 5.

(i) Its expectation and variance are given by

$$\mu_X = \mu \quad \textit{and} \quad \sigma_X^2 = \sigma^2.$$

(ii) Its moment-generating function is given by

$$\forall t \in \mathbb{R}, \quad M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

3.4 Functions of a continuous random variable

Recall Remark 3.9: given a random variable $X\colon \Omega \to \mathbb{R}$ and a function $u\colon \mathbb{R} \to \mathbb{R}$, we can define their composition $u(X)\colon \Omega \to \mathbb{R}$ and easily compute its mean via the formula

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

Let us go a step further. What is the distribution of the random variable Y?

3.4.1 The change of variables theorem

Now we will determine the distribution of the random variable $Y \coloneqq u(X)$ under some assumptions on the function u. More precisely, let $A, B \subset \mathbb{R}$ and let $u\colon A \to B$ be a bijective function[2], so its inverse $u^{-1}\colon B \to A$ exists. Moreover, let us assume that u is differentiable and increasing, that is

$$x < y \implies u(x) < u(y),$$

or equivalently,

$$\forall x \in A, \quad u'(x) > 0.$$

In light of Proposition 3.5, **if we want to know f_Y, it just suffices to compute F_Y** because $f_Y = F_Y'$. Let $y \in B$:

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(u(X) \leqslant y) = \mathbb{P}(X \leqslant u^{-1}(y)) = F_X(u^{-1}(y)).$$

Then, for $y \in A$,

$$f_Y(y) = F_Y'(y) = (F_X(u^{-1}(y)))' = F_X'(u^{-1}(y)) \cdot (u^{-1})'(y) = f_X(u^{-1}(y)) \cdot (u^{-1})'(y),$$

where we used the chain rule and the fact that $F_X' = f_X$. That is,

$$\forall y \in B, \quad f_Y(y) = f_X(u^{-1}(y)) \cdot (u^{-1})'(y).$$

[2] That is, injective (one-to-one) and surjective (onto).

Exercise 3.4. Now assume that u is decreasing:

$$x < y \implies u(x) > u(y),$$

or equivalently,

$$\forall x \in A, \quad u'(x) < 0.$$

Prove that in this case,

$$\forall y \in A, \quad f_Y(y) = -f_X(u^{-1}(y)) \cdot (u^{-1})'(y).$$

Joining both the previous conclusion and Exercise 3.4, we have proven the following theorem.

Theorem 3.21 (Change of variables)**.** *Consider a random experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with density f_X. Let $A, B \subset \mathbb{R}$ and let $u\colon A \to B$ be a bijective and differentiable function with inverse $u^{-1}\colon B \to A$. Define $Y := u(X)$. Then the density of Y is given by*

$$\forall y \in A, \quad f_Y(y) = f_X(u^{-1}(y)) \cdot \left|(u^{-1})'(y)\right|.$$

Example 3.22. Let $X \sim \mathcal{N}(0,1)$ and let $u\colon \mathbb{R} \to (0,+\infty)$ be the exponential function. The function u is bijective (with inverse $u^{-1} = \log$, the natural logarithm) and differentiable. Then, if $Y = u(X)$, we have that, by applying Theorem 3.21, for every $y \in (0,+\infty)$,

$$f_Y(y) = f_X(u^{-1}(y)) \cdot \left|(u^{-1})'(y)\right| = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\log(y)^2} \frac{1}{y}.$$

We say that Y has a *log-normal distribution.*

3.4.2 Sampling from uniform random variables

In applications, we would like to simulate a specific random variable X with a known CDF F_X. It is known that uniform random variables are quite easy to simulate. the following theorem provides a procedure to simulate a random variable Y having the same distribution as X, that is, having the same CDF (and thus the same density): simulate $U \sim \mathrm{U}([0,1])$ and define $Y := F_X^{-1}(U)$.

Theorem 3.23. *Consider a random experiment with outcome space Ω and let $\mathbb{P}$ be a probability defined on $\mathcal{P}(\Omega)$. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with an invertible CDF F_X. Let $U\colon \Omega \to \mathbb{R}$ be a random variable with $U \sim \mathrm{U}([0,1])$. Then the random variable $Y = F_X^{-1}(U)$ has CDF $F_Y = F_X$.*

Exercise 3.5. Prove Theorem 3.23.

Chapter 4

Bivariate random variables

Chapter 5

Concentration and limit theorems