SOR1020 Introduction to Probability and Statistics

Andrea Munaro

Contents

Chapter 1

Preliminaries

In this short preliminary chapter, we will recall some basic facts about sets that will be frequently used and introduce the induction principle.

1.1 Sets

A **set** is any well-defined collection of objects. Each object in a set is called an **element** (or **member**) of the set and the notation $a \in A$ reads "the element a belongs to the set A". A set A is a **subset** of a set B, denoted $A \subseteq B$, if every element of A is also an element of B. The **empty set**, denoted by $\varnothing$, is the set that contains no elements. Clearly, $\varnothing$ is a subset of any set.

There are many acceptable ways to assert the contents of a set. For example, a set can be described by an explicit list: $A = \{1, 2, 3, 4, 5, 6\}$, $B = \{H, T\}$ $C = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$. Sets can also be described in words. For instance, we can define the set E to be the collection of even natural numbers. Sometimes it is more efficient to provide a kind of rule or algorithm for determining the elements of a set. As an example, let $C = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. Clearly, C is nothing but the set of points on the circle with center the origin and radius 1.

Two sets A and B are **equal**, denoted $A = B$, if they contain exactly the same elements. In other words, $A = B$ means that $A \subseteq B$ and $A \supseteq B$. Given a set A, its **power set** is the set whose elements are the subsets of A. Sometimes the power set of A is denoted by 2^A, for a reason to be discussed in the next chapter.

Example 1.1.1. Let $A = \{*, \circ\}$. The power set of A is then the set $\{\varnothing, \{*\}, \{\circ\}, \{*, \circ\}\}$.

It is often the case that all sets we are interested in are subsets of one particular set X, which we call the **universe**. Let A and B be two subsets of X. The **complement of** B **in** A, denoted $A \setminus B$, is the set of elements of X which belong to A but not to B. When the set X is clear from the context, we write $A^c := X \setminus A$ and refer to A^c simply as the **complement** of A. The **intersection** of A and B is the set of elements of X which belong to both A and B. The sets A and B are **disjoint** if $A \cap B = \varnothing$. The **union** of A and B is set of elements of X which belong to A or to B or to both.

Remark 1.1.2. It is useful to represent graphically the relationships between sets using Venn diagrams. Each set is represented by a region of the plane enclosed by a curve. Such diagrams cannot be used to prove theorems. However, providing intuition about the possible relationships between sets, they do suggest what statements about sets might be provable.

Example 1.1.3. Let our universe be $X = \{1, 2, 3, 4, 5\}$, and let $A = \{1, 2, 3\}$ and $B = \{1, 3, 5\}$. Then $A \cap B = \{1, 3\}$, $A \cup B = \{1, 2, 3, 5\}$, $A \setminus B = \{2\}$, $A^c = \{4, 5\}$, $B^c = \{2, 4\}$.

In the following lemma, we collect some simple algebraic properties of the intersection and union operations. The reader is encouraged to prove them using the definitions above.

Lemma 1.1.4. *Let A, B, C be subsets of a set. The following holds:*

- *$A \cup B = B \cup A$ and $A \cap B = B \cap A$ (commutativity);*
- *$A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$ (associativity);*
- *$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ (distributivity).*

Proof. We show only the first distributive property and invite the reader to show the remaining statements.

In order to show that two sets are equal, we need to show that both inclusions hold. Let us first show that $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$. Given an element $x \in A \cup (B \cap C)$, we need to show that x belongs to $(A \cup B) \cap (A \cup C)$ as well. Since $x \in A \cup (B \cap C)$, either $x \in A$ or $x \in B \cap C$. In either case, x belongs to both $A \cup B$ and $A \cup C$ and so $x \in (A \cup B) \cap (A \cup C)$.

Let us finally show that $A \cup (B \cap C) \supseteq (A \cup B) \cap (A \cup C)$. Given an element $x \in (A \cup B) \cap (A \cup C)$, we need to show that x belongs to $A \cup (B \cap C)$ as well. By definition of intersection, x belongs to both $A \cup B$ and $A \cup C$. There are now two possibilities: either $x \in A$ or $x \notin A$. If $x \in A$, then trivially $x \in A \cup (B \cap C)$. If $x \notin A$, then it must be that $x \in B$ (since $x \in A \cup B$) and that $x \in C$ (since $x \in A \cup C$). Therefore, $x \in B \cap C$ and so again $x \in A \cup (B \cap C)$, as claimed. □

Given two objects a and b, we can form a new object, the **ordered pair** (a, b). The objects a and b are called the first and second **component** of the ordered pair (a, b), respectively. Two ordered pairs (a, b) and (a', b') are **equal**, denoted $(a, b) = (a', b')$, if $a = a'$ and $b = b'$. If X and Y are sets, then the **Cartesian product** $X \times Y$ of X and Y is the set of all ordered pairs (x, y) with $x \in X$ and $y \in Y$.

Example 1.1.5. $\mathbb{R} \times \mathbb{R}$ (also denoted as $\mathbb{R}^2$) is the real plane the reader is already familiar with.

Example 1.1.6. For $X = \{a, b\}$ and $Y = \{\circ, *, \triangleleft\}$,

$$X \times Y = \{(a, \circ), (a, *), (a, \triangleleft), (b, \circ), (b, *), (b, \triangleleft)\}.$$

Let us now introduce the important notion of countable set.

Definition 1.1.7. A set is **countable** if it is in bijection with a subset of the set of positive integers. A set is **uncountable** if it is not countable.

In other words, a set is countable if we can make a list of its elements i.e., if we can can find a first element, a second one, and so on, and eventually assign to each element an integer, perhaps going on forever.

Example 1.1.8. The set $\mathbb{N}$ of natural numbers is countable. Every finite set is countable.

Example 1.1.9. The set of rational numbers $\mathbb{Q}$ is countable. To see this, let us first list the non-negative rational numbers as follows. Take first all those whose numerator and denominator sum to 1, then 2, then 3, and so on. When several do so, order them by increasing size. We obtain the following list:

$$\frac{0}{1}, \frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{2}{2}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \ldots$$

Every positive rational number appears on this list somewhere, and actually appears often on it (this happens because $1/2$ appears as $1/2$ and also as $2/4$ and $3/6$ and so on). But all fractions eventually appear, and appear over and over again.

With a similar reasoning, we can list the negative rational numbers as follows:

$$-\frac{1}{1}, -\frac{1}{2}, -\frac{2}{1}, -\frac{1}{3}, -\frac{2}{2}, -\frac{3}{1}, \ldots$$

Consider now these two lists (of non-negative and negative rational numbers). We build a new list of all rational numbers by making the odd entries of this new list the non-negative rational numbers, and the even entries the rest, following the orders in the two previous lists. We obtain:

$$\frac{0}{1}, -\frac{1}{1}, \frac{1}{1}, -\frac{1}{2}, \frac{1}{2}, -\frac{2}{1}, \frac{2}{1}, -\frac{1}{3}, \frac{1}{3}, -\frac{2}{2}, \frac{2}{2}, \ldots$$

Keeping only the first occurrence of each rational number, we obtain a desired list of $\mathbb{Q}$ (notice that in fact there are several different ways one could list the elements of $\mathbb{Q}$).

Rational numbers are described by pairs of integers, and the arguments above generalize to imply that any collection of pairs of members of a countable set are countable.

Example 1.1.10. Not all infinite sets are countable. Indeed, the set of all infinite sequences consisting of 0's and 1's is uncountable. Other examples of uncountable sets are the set of real numbers $\mathbb{R}$. In fact, even the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ is uncountable. These results can be proved using the so-called *Cantor's diagonal argument* (we omit details).

Remark 1.1.11. Loosely speaking, the fact that $\mathbb{R}$ is uncountable whereas $\mathbb{Q}$ is countable implies that the "size" of $\mathbb{R}$ is strictly bigger than that of $\mathbb{Q}$: "there are more real numbers than rational numbers".

We can now generalize the notions of union and intersection of two sets to union and intersection of a collection of sets:

Definition 1.1.12. Let $\mathcal{C}$ be a collection of subsets of a set X. The subset of X containing all elements that belong to at least one set of $\mathcal{C}$ is the **union** of the collection $\mathcal{C}$, denoted by $\bigcup \mathcal{C}$. If $\mathcal{C} = \{A_1, A_2, \ldots, A_n\}$ is finite, we usually write $\bigcup \mathcal{C} = \bigcup_{i=1}^{n} A_i$. If $\mathcal{C} = \{A_1, A_2, \ldots\}$ is countable infinite, we usually write $\bigcup \mathcal{C} = \bigcup_{i=1}^{\infty} A_i$.

The subset of X containing all elements that belong to all sets of $\mathcal{C}$ is the **intersection** of the collection $\mathcal{C}$, denoted by $\bigcap \mathcal{C}$. If $\mathcal{C} = \{A_1, A_2, \ldots, A_n\}$ is finite, we let $\bigcap \mathcal{C} = \bigcap_{i=1}^{n} A_i$. If $\mathcal{C} = \{A_1, A_2, \ldots\}$ is countable infinite, we let $\bigcap \mathcal{C} = \bigcap_{i=1}^{\infty} A_i$.

The following relations between unions and intersections will be used repeatedly. Let us start with a concrete observation. "It will not snow or rain" means "It will not snow and it will not rain". If S is event that it snows and R is event that it rains, then $(S \cup R)^c = S^c \cap R^c$. "It will not both snow and rain" means "Either it will not snow or it will not rain" i.e., $(S \cap R)^c = S^c \cup R^c$. More generally, the following holds:

Lemma 1.1.13 (De Morgan's laws). *let $\{A_i : i \in I\}$ be a collection of sets with I countable. Then*

$$\Big(\bigcup_{i\in I} A_i\Big)^c = \bigcap_{i\in I} A_i^c \quad \textit{and} \quad \Big(\bigcap_{i\in I} A_i\Big)^c = \bigcup_{i\in I} A_i^c.$$

Proof. Easy exercise left to the reader. □

1.2 Induction principle

One trick which arises with some frequency in mathematics is the use of induction arguments or "proof by induction". For each $n \in \mathbb{N}$, let $\mathcal{A}(n)$ be a statement depending on n. To prove by induction on n that $\mathcal{A}(n)$ is true for each $n \in \mathbb{N}$, one uses the following steps:

1. Base case: Prove that $\mathcal{A}(0)$ is true.
2. This step has two parts:
   (a) Induction hypothesis: Suppose that $\mathcal{A}(n)$ is true for some $n \in \mathbb{N}$.
   (b) Induction step $(n \to n+1)$: Prove that $\mathcal{A}(n+1)$ follows from (a) and possibly other previously proved statements.

If 1. and 2. can be done, then $\mathcal{A}(n)$ is true for all $n \in \mathbb{N}$. To see this, let

$$N = \{n \in \mathbb{N} : \mathcal{A}(n) \text{ is true}\}.$$

Then 1. implies that $0 \in N$ and from 2. we have that $n \in N$ implies $n+1 \in N$ for all $n \in \mathbb{N}$. It follows that $N = \mathbb{N}$.

In many applications it is useful to start the induction with some number other than 0. This leads to the following slight generalization of the above method:

Lemma 1.2.1 (Induction principle). *Let $n_0 \in \mathbb{N}$ and, for each $n \geq n_0$, let $\mathcal{A}(n)$ be a statement. If*

1. *$\mathcal{A}(n_0)$ is true, and*
2. *for each $n \geq n_0$, $\mathcal{A}(n+1)$ can be proved from the assumption that $\mathcal{A}(n)$ is true,*

then $\mathcal{A}(n)$ is true for all $n \geq n_0$.

Example 1.2.2. For each natural number $n \geq 1$, we have $1 + 3 + 5 + \cdots + (2n-1) = n^2$.

To show this, we proceed by induction. The base case is trivially true as $1 = 1^2$. The induction hypothesis is: Suppose that, for some $n \geq 1$, we have $1+3+5+\cdots+(2n-1) = n^2$. The induction step proceeds as follows:

$$\begin{aligned} 1 + 3 + 5 + \cdots + (2(n+1) - 1) &= 1 + 3 + 5 + \cdots + (2n+1) \\ &= 1 + 3 + 5 + \cdots + (2n-1) + (2n+1) \\ &= n^2 + 2n + 1 \\ &= (n+1)^2, \end{aligned}$$

which is exactly what we were supposed to show.

Exercise 1.2.3. *Show the following statement by induction: For each $n \geq 1$,*

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Chapter 2

Probability

Uncertainty and randomness are unavoidable aspects of our experience: play cards, invest in shares, etc. Although probability has been around for several centuries, it wasn't until recently that the subject was made rigorous. In the thirties, the Russian mathematician Kolmogorov showed that probability is in fact full-fledged analysis or, more precisely, measure theory. Properly justifying this assertion goes beyond the scope of the module but we will provide some examples showing the analytical nature of probability.

This chapter is devoted to formally introducing the objects of probability. In other words, the goal is to abstract the common features arising in everyday examples in order to build a **probabilistic model** i.e., a mathematical description of an uncertain situation. The advantage of taking an abstract approach is that it allows to develop general tools that can be adapted to several specific situations. We start with the following definition.

Definition 2.0.1. Any well-defined procedure or chain of circumstances is called an **experiment**. The end result, or occurrence, is the **outcome** of the experiment, also known as **elementary event**. The set of all possible outcomes is the **sample space**, denoted by Ω.

Experiment	Possible outcomes
Roll a die	$\Omega = \{1, 2, 3, 4, 5, 6\}$
Toss a coin	$\Omega = \{$H, T$\}$
Toss a coin until heads appears	$\Omega = \{$H, TH, TTH, TTTH, $\dots\}$
Infinite sequence of coin tosses	Ω is the set of all possible infinite sequences of H and T

In the first two experiments Ω is finite, whereas in the latter two it is infinite. Typically, rather than individual outcomes of the sample space, we are interested in collections of outcomes.

Experiment	Set of outcomes of interest
Roll a die	The outcome is an even number
Toss a coin	The outcome is either H or T
Infinite sequence of coin tosses	The outcome consists of finitely many H

These collections of outcomes are associated to the intuitive notion of event, which is then a subset of the sample space. If the result of the experiment belongs to this subset, we would say that the event occured. Thinking of events as subsets of the sample space, we can then perform on them the usual set-theoretic operations.

Notation	Set jargon	Probability jargon
Ω	Universe	Sample space
ω	Element of Ω	Outcome (also called elementary event)
A	Subset of Ω	A occurs (i.e., the end result belongs to A)
A^c	Complement of A	A does not occur (i.e., the end result does not belong to A)
$A \cap B$	Intersection	Both A and B occur
$A \cup B$	Union	At least one of A and B occurs
$A \setminus B$	Complement of B in A	A occurs but not B
$A \subseteq B$	Inclusion	If A occurs then B occurs

We would ultimately like to assign a probability to an event and so the natural question is: Is there any property events should satisfy? As already observed, in general, the sample space might be infinite and the events we are interested in might contain infinitely many outcomes:

Example 2.0.2. A coin is tossed until the first head turns up and we are concerned with the number of tosses before this happens. We let $\Omega = \{\omega_1, \omega_2, \dots\}$, where ω_i denotes the outcome "the first $i-1$ tosses are tails and the i-th is head". We might be interested in the following event A: "the first head occurs after an even number of tosses". Clearly, $A = \{\omega_2, \omega_4, \dots\} = \bigcup_{i=1}^{\infty} \{\omega_{2i}\}$ is a countable union of members of Ω i.e., elementary events.

2.1 Sigma-fields

Back to our question: which subsets of the sample space Ω are events? There are certain requirements that we wish the collection of events to satisfy:

- Ω is an event: this is the trivial event that something happened.
- If $A \subseteq \Omega$ is an event, so is A^c: if we are allowed to ask whether A has occurred, we should also be allowed to ask whether A has not occurred.
- If $A_1, A_2, \dots \subseteq \Omega$ are events, so is their union $\bigcup_{i=1}^{\infty} A_i$: if we are allowed to ask whether each A_i has occurred, we should also be allowed to ask whether at least one of the A_i's has occurred, as seen in Example 2.0.2.

Definition 2.1.1. A collection $\mathcal{F}$ of subsets of Ω is called a σ**-field** (or σ**-algebra**) of Ω if it satisfies the following conditions:

1. $\Omega \in \mathcal{F}$;
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
3. If $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Our events will form a σ-field of the sample space Ω. Why don't we go further and allow uncountable unions? Well, our unions here will be closely tied to sums of probabilities and uncountable sums can be extremely messy.

Remark 2.1.2. Notice that 1. and 2. imply that $\varnothing \in \mathcal{F}$. Moreover, if $A_1, A_2, \dots \in \mathcal{F}$ is a countable subfamily of $\mathcal{F}$ then, by De Morgan's laws, $\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c \in \mathcal{F}$. Notice also that, since $\varnothing \in \mathcal{F}$, we can extend any finite subfamily $A_1, \dots, A_n$ of $\mathcal{F}$ to a countable family by setting $A_j = \varnothing$ for each $j > n$. Therefore, finite unions and intersections of members of $\mathcal{F}$ are still in $\mathcal{F}$. Moreover, if $A, B \in \mathcal{F}$, then $A \setminus B = A \cap B^c \in \mathcal{F}$.

Example 2.1.3.

- $\{\varnothing, \Omega\}$ is a σ-field of Ω.
- The power set of Ω is a σ-field of Ω.
- For any $A \subseteq \Omega$, $\{\varnothing, A, A^c, \Omega\}$ is a σ-field of Ω.
- The collection of all open intervals of $\mathbb{R}$ is not a σ-field of $\mathbb{R}$. Indeed, $\bigcap_{n=1}^{\infty}(-\frac{1}{n}, \frac{1}{n}) = \{0\}$ is not an open interval.

Exercise 2.1.4. *Write down all σ-fields on $\{a, b\}$.*

Exercise 2.1.5. *Let Ω be a countable infinite set and let $\mathcal{A} = \{A \subseteq \Omega : A$ is finite or A^c is finite$\}$. Show that $\mathcal{A}$ is not a σ-field.*

A natural question might arise: Why not simply taking the power set of Ω all the time for our probabilistic interests? The reason is that, if Ω is uncountable, its power set is too rich and it turns out to be impossible to assign probabilities in a consistent fashion to all possible subsets.

> Luckily, in many situations, for example when Ω is countable, we can indeed simply consider the power set of Ω.

But let us deal with the general situation for a moment and see what is typically done. If $\mathcal{C}$ is a collection of some basic events that we want to be able to discuss, we have seen in Example 2.1.3 that it is not necessarily a σ-field. What is typically done is to enlarge such a collection so that it in fact becomes a σ-field. This is done by considering the following notion.

Definition 2.1.6. Let $\mathcal{C}$ be a collection of subsets of Ω. The σ**-field generated by** $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the smallest σ-field on Ω containing the collection $\mathcal{C}$.

Remark 2.1.7. Some cautionary words on language. We say that a σ-field $\mathcal{F}$ contains the collection $\mathcal{C}$ if each member of $\mathcal{C}$ belongs to $\mathcal{F}$ (i.e., every set in $\mathcal{C}$ is a set in $\mathcal{F}$). For two σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$, we say that $\mathcal{F}_1$ is smaller than $\mathcal{F}_2$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$.

Notice that we need to verify that this notion is well-defined! Indeed, a priori, it is not even clear why such a thing should exist.

Lemma 2.1.8. *Let $\mathcal{C}$ be a collection of subsets of Ω. Then the σ-field $\sigma(\mathcal{C})$ generated by $\mathcal{C}$ exists and is unique.*

Proof. Uniqueness is trivial: If we have two smallest σ-fields containing $\mathcal{C}$, say $\mathcal{F}_1$ and $\mathcal{F}_2$, then $\mathcal{F}_1 \subseteq \mathcal{F}_2$ and $\mathcal{F}_2 \subseteq \mathcal{F}_1$, implying that $\mathcal{F}_1 = \mathcal{F}_2$.

Consider now existence. Let $\mathscr{S}$ be the collection of all σ-fields on Ω containing $\mathcal{C}$. Notice that $\mathscr{S}$ is non-empty, as the power set of Ω belongs to $\mathscr{S}$. We claim that $\bigcap_{\mathcal{F} \in \mathscr{S}} \mathcal{F}$ is the smallest σ-field on Ω containing $\mathcal{C}$. There are three things to be checked:

- $\bigcap_{\mathcal{F} \in \mathscr{S}} \mathcal{F}$ is a σ-field.

  Here we need to check the three properties in Definition 2.1.6:

  1. Since each $\mathcal{F} \in \mathscr{S}$ is a σ-field on Ω, $\Omega \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $\Omega \in \bigcap_{\mathcal{F} \in \mathscr{S}} \mathcal{F}$.
  2. Let $A \in \bigcap_{\mathcal{F} \in \mathscr{S}} \mathcal{F}$. Then $A \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $A^c \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$. Therefore, $A^c \in \bigcap_{\mathcal{F} \in \mathscr{S}} \mathcal{F}$.

3. Let $A_1, A_2, \ldots \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$. Then $A_1, A_2, \ldots \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$. But $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ for each $\mathcal{F} \in \mathscr{S}$ and so $\bigcup_{i=1}^{\infty} A_i \in \bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$.

- $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ contains $\mathcal{C}$.

  This follows from the fact that, for each $\mathcal{F} \in \mathscr{S}$, $\mathcal{F}$ contains $\mathcal{C}$.

- $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F}$ is the smallest σ-field containing $\mathcal{C}$.

  Let $\mathcal{G}$ be any σ-field containing $\mathcal{C}$. Then $\mathcal{G} \in \mathscr{S}$ and so $\bigcap_{\mathcal{F}\in\mathscr{S}} \mathcal{F} \subseteq \mathcal{G}$. □

In order to define what is arguably the most important σ-field on $\mathbb{R}$, we need a little bit of topology.

Definition 2.1.9. A subset $U \subseteq \mathbb{R}$ is **open** if, for each $x \in U$, there exists $\varepsilon > 0$ such that the open interval centered at x and with radius ε is contained in U. In other words, $(x-\varepsilon, x+\varepsilon) \subseteq U$.

Similarly, a subset $U \subseteq \mathbb{R}^2$ is **open** if, for each $x \in U$, there exists $\varepsilon > 0$ such that the open ball centered at x and with radius ε is contained in U. In other words, $B_\varepsilon(x) = \{y \in \mathbb{R}^2 : |x-y| < \varepsilon\} \subseteq U$, where $|x-y|$ denotes the Euclidean distance in $\mathbb{R}^2$ between x and y.

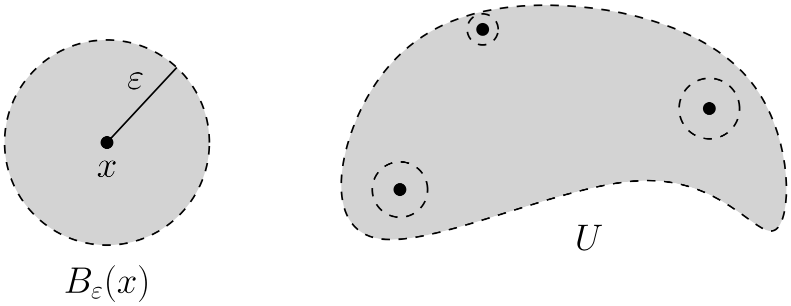


Figure 2.1: The open ball $B_\varepsilon(x)$ in $\mathbb{R}^2$ and an example of an arbitrary open set $U \subseteq \mathbb{R}^2$: for each of its points x there exists a sufficiently small open ball centered at x and contained in U.

Example 2.1.10. Every open interval in $\mathbb{R}$ is open. For each $x \in \mathbb{R}$, $\{x\}^c$ is an open set in $\mathbb{R}$ but $\{x\}$ is obviously not open. More generally, the closed interval $[x, y]$ is not open but $[x, y]^c$ is.

Exercise 2.1.11. *Let U_1 and U_2 be two open sets in $\mathbb{R}^2$. Is $U_1 \cap U_2$ open?*

Definition 2.1.12. For $n = 1, 2$, the **Borel σ-field** $\mathcal{B}$ on $\mathbb{R}^n$ is the σ-field generated by the open sets in $\mathbb{R}^n$. The sets in $\mathcal{B}$ are called **Borel sets**.

The Borel σ-field on the real line is extremely important in probability theory and will appear again when we will talk about random variables. Rather than taking the whole family of open sets, it can in fact be equivalently generated by open intervals, closed intervals, half-lines, etc.

Proposition 2.1.13. *The Borel σ-field on the real line $\mathbb{R}$ is generated by any of the following collections of subsets of $\mathbb{R}$:*

- $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$;
- $\mathcal{C} = \{(x, y) : x, y \in \mathbb{R},\ x < y\}$;

- $\mathcal{C} = \{[x, y] : x, y \in \mathbb{R},\ x \leq y\}$;
- $\mathcal{C} = \{(x, y] : x, y \in \mathbb{R},\ x < y\}$.

Remark 2.1.14. Proposition 2.1.13 says that the smallest σ-field on $\mathbb{R}$ containing all open sets can be generated by the family of closed intervals. This might appear odd, as closed intervals are not open. But recall that σ-fields are closed under complementation, and it is by taking complements of closed intervals (and their intersections and unions) that we manage to get the open sets.

Example 2.1.15. Since $\mathcal{B}$ is a σ-field containing all open sets, it contains all singletons $\{x\}$, as $\{x\}^c$ is an open set in $\mathbb{R}$.

2.2 Probability measures

To any experiment we will associate the pair $(\Omega, \mathcal{F})$, where Ω is the set of all possible outcomes (elementary events) and $\mathcal{F}$ is a σ-field of subsets of Ω. We will try to assign a probability to each set in $\mathcal{F}$ and in order to do so, we will be guided by intuition.

Suppose that an experiment has several possible outcomes that are not necessarily equally likely. How can we define the probability of a certain event A? One intuitive way is the following. We run the experiment a large number N of times, keeping the initial conditions as equal as possible. Denoting by $N(A)$ the number of occurences of A after the first N trials, we would expect that when N becomes larger and larger, the ratio $N(A)/N$ converges to some finite limit. We may then define the probability $\mathbb{P}(A)$ that A occurs on a particular trial as this limit. In any case, for large N, $N(A)/N$ should be an approximation of $\mathbb{P}(A)$. Notice that

- $0 \leq N(A)/N \leq 1$;
- If $A = \varnothing$, then $N(A)/N = 0$. If $A = \Omega$, then $N(A)/N = 1$;
- If A and B are disjoint events, then $N(A \cup B) = N(A) + N(B)$ and so $N(A \cup B)/N = N(A)/N + N(B)/N$.

With the observations above in mind and recalling Example 2.0.2, we state the following:

Definition 2.2.1. A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P}\colon \mathcal{F} \to \mathbb{R}$ satisfying the following:

1. For each $A \in \mathcal{F}$, we have $0 \leq \mathbb{P}(A) \leq 1$;
2. $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$;
3. For every countable infinite collection $A_1, A_2, \ldots$ of mutually disjoint members of $\mathcal{F}$ (i.e., $A_i \cap A_j = \varnothing$ for each $i \neq j$), we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is called a **probability space**. Any set in $\mathcal{F}$ is called **event**.

Remark 2.2.2. For the time being, the reader can think of $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ as the infinite sum $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \cdots$. The proper mathematical definition will be given in MTH1011 and is not fundamental for understanding the rest of the notes.

Observation 2.2.3. *The axioms $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(A) \leq 1$ in Definition 2.2.1 are in fact redundant i.e., they can be deduced from the others. Check it!*

Remark 2.2.4. The event Ω is the **sure event**: it contains all possible outcomes and $\mathbb{P}(\Omega) = 1$. It is worth noting that there may be also other events $E \in \mathcal{F}$ such that $\mathbb{P}(E) = 1$. Such events are called **almost sure**.

Countable additivity (the last condition in the definition) readily implies the following result:

Lemma 2.2.5 (Finite additivity). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For every finite collection $A_1, A_2, \ldots, A_n$ of mutually disjoint members of $\mathcal{F}$, we have*

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Proof. Define $A_m = \varnothing$ for each $m > n$. Since $\varnothing \in \mathcal{F}$ and the newly defined countable infinite collection $A_1, A_2, \ldots$ consists of mutually disjoint members of $\mathcal{F}$, we use countable additivity to conclude. □

The conceptual construction of a probability space has no absolute physical meaning, it is just guided by some intuitive physical interpretation. The properties which the measure $\mathbb{P}$ is required to satisfy are called the **probability axioms** and were introduced by Kolmogorov, though not exactly in the form above (see Observation 2.2.3). The first two axioms are just a matter of convention. The key one is countable additivity.

Think of a probability space as the mathematical description of an experiment. For example, tossing a coin, rolling dice, taking a number in a lottery, etc. In each case, there is a certain amount of randomness, or unpredictability in the experiment. To describe this mathematically, start with what we observe: the outcome. Ω is the set of all possible outcomes of the experiment: each element of Ω represents an outcome. An event is a set of outcomes belonging to the σ-field $\mathcal{F}$. The probability measure gives the probability of events. We can associate a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with any experiment. The informations allowing us to compute the actual value of $\mathbb{P}(A)$ are contained in the description of the experiment.

Example 2.2.6. Suppose the experiment is rolling a die i.e., a cube whose six faces are numbered 1 to 6. We can take $\Omega = \{1, 2, 3, 4, 5, 6\}$ as the set of outcomes and, since Ω is countable, the power set of Ω as the σ-field $\mathcal{F}$. To get the probability measure, we note that if the die is well-made, the six sides are identical except for their label. No side can be more probable than another. Therefore,

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \cdots = \mathbb{P}(\{6\}).$$

The reasoning in the preceding example was derived from symmetry considerations: the possible outcomes were indistinguishable except by their labels. In fact, this is about the only situation in which we can confidently assign probabilities by inspection. But luckily, while nature is not always obliging enough to divide itself into equally-likely pieces, one can start with the equally-likely case and then determine the probabilities in more complex situations. Which is what the subject is about.

The idea of symmetry applies to events, not just outcomes. Consider a physical experiment with finitely or countably many outcomes, labeled in some convenient fashion.

> **Symmetry principle:** If two events are indistinguishable except for the way the outcomes are labeled, they are equally likely.

For example, roll a die and consider the events "even" and "odd", i.e., $\{2, 4, 6\}$ and $\{1, 3, 5\}$. If we physically renumber the faces of the die, so that we interchange n and $7 - n$ on each face, so that $1 \leftrightarrow 6$, $2 \leftrightarrow 5$ and $3 \leftrightarrow 4$, then the events "even" and "odd" are interchanged. The symmetry principle says that the two events must have the same probability.

Example 2.2.7. Suppose the experiment is tossing a coin. We can take $\Omega = \{H, T\}$, $\mathcal{F} = 2^\Omega = \{\varnothing, H, T, \Omega\}$ and $\mathbb{P}$ defined by

$$\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(H) = p, \quad \mathbb{P}(T) = 1 - p,$$

where p is a fixed real number in $[0, 1]$. It is easily seen that all three probability axioms are satisfied. If $p = 1/2$, we say that the coin is **fair**.

Remark 2.2.8. The probability space for an experiment is not unique. This is useful in practice. It allows us to choose the probability space which works best in the particular circumstances, or to not choose one at all; we do not always have to specify the probability space. It is usually enough to know that it is there if we need it.

The simplest probability spaces are those whose sample space $\Omega = \{\omega_1, \omega_2, \dots\}$ contains countably many outcomes (we call such probability spaces, **countable probability spaces**). Recall that in such cases we may always take as σ-field $\mathcal{F}$ the power set 2^Ω. For countable probability spaces, a probability measure $\mathbb{P}$ on $\mathcal{F}$ is fully determined by the values assigned to the elementary events ω_i. Indeed, consider an event $A \subseteq \Omega$. Since A is countable (infinite or finite), then

$$A = \bigcup_{\omega \in A} \{\omega\}$$

can be expressed as a countable union of elementary events. Since these events are obviously mutually disjoint, it follows from countable additivity (or finite additivity) that

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}).$$

Suppose now that the sample space Ω is finite and that $\mathbb{P}(\{\omega\}) = p$, for each elementary event $\omega \in \Omega$. By the probability axioms, we have

$$1 = \mathbb{P}(\Omega) = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = p|\Omega|,$$

from which $p = 1/|\Omega|$. Therefore, the probability of the event A is

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = p|A| = \frac{|A|}{|\Omega|}.$$

Observation 2.2.9. *A probability space is a model for an experiment. A priori there is no reason why all outcomes should be equally probable. It is an assumption that should be made only when believed to be applicable.*

Example 2.2.10. Two fair dice are rolled. What is the probability that the sum is 7?

A convenient sample space is constructed by viewing the two dice as distinguishable, say one blue and one red, and taking $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$, where the first component accounts for the outcome of the blue die and the second component for the outcome of the red die. By symmetry, it is natural to assume that each of the $|\Omega| = 36$ outcomes is equally likely. The event "sum equals 7" is $A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ and so $\mathbb{P}(A) = |A|/|\Omega| = 1/6$.

Observation 2.2.11. *How do we know which probability space to assign to each experiment? Well, a model is a model: it may or may not relate to reality. In the previous example, we applied symmetry, as we believe that all outcomes of the experiment are equally likely.*

Example 2.2.12. Three fair coins are tossed. What is the probability of observing three heads or three tails?

One might be tempted by the following fallacious argument. There are four possible outcomes, namely 3 heads, 2 heads and 1 tail, 1 head and 2 tails, 3 tails. Since the event we are interested in consists of two of these outcomes, the desired probability is $2/4$. However, this reasoning assumes that the four outcomes are equally likely, which is not the case.

A solid argument is the following. Consider the coins to be distinguishable, say £2, £1 and 50p. We have the following possible outcomes:

£2	£1	50p
H	H	H
H	H	T
H	T	H
T	H	H
T	T	H
T	H	T
H	T	T
T	T	T

By fairness of the coins, these eight outcomes are equally likely and so the desired probability is $2/8$.

Example 2.2.13. A tea set has four cups and saucers with two cups and saucers in each of two different colors, say a and b. If the cups are placed at random on the saucers, what is the probability that no cup is on a saucer of the same color?

As a sample space, we consider the distinct ways of arranging the cups by color with the saucers fixed (suppose without loss of generality the saucers are listed as $aabb$). There are six possible ways of arranging the cups: $aabb, abba, abab, baab, baba, bbaa$. By symmetry, they are equally likely. Since only one of these arrangements has no cup on a saucer of the same color, the required probability is $1/6$.

Exercise 2.2.14. *A bag contains* 2021 *red balls and* 2021 *black balls. We remove two balls at a time repeatedly and*

- *discard them if they are of the same color;*

- *discard the black ball and return to the bag the red ball if they are of different colors.*

What is the probability that this process will terminate with one red ball in the bag? *Hint: What are the possible outcomes?*

2.3 Counting principles

We have seen in the previous examples that it is often necessary to be able to count the number of subsets of Ω with a given property. We now take a systematic look at some counting methods.

2.3.1 Multiplication rule

Take N finite sets $\Omega_1, \dots, \Omega_N$ (some of which might coincide), with cardinalities $|\Omega_k| = n_k$. We imagine to pick one element from each set: how many possible ways do we have to do so? Clearly, we have n_1 choices for the first element. Now, for each choice of the first element, we have n_2 choices for the second. Once the first two elements are picked, we have n_3 choices for the third, and so on, giving $|\Omega_1 \times \cdots \times \Omega_N| = n_1 \cdots n_N$. We refer to this as the multiplication rule.

Example 2.3.1 (Number of subsets). Suppose a set $\Omega = \{\omega_1, \dots, \omega_n\}$ has n elements. How many subsets does Ω have?

We proceed as follows. To each subset A of Ω we can associate a string of 0's and $1's$ of length n such that the i-th element of the string is 1 if $\omega_i \in A$, and 0 otherwise. For example, if $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, to the subset $A = \{\omega_1\}$ we associate the string $1, 0, 0, 0$ and to the subset $B = \{\omega_1, \omega_3, \omega_4\}$ we associate the string $1, 0, 1, 1$. This defines a bijection between the subsets of Ω and the strings of 0's and 1's of length n. Thus we have to count the number of such strings. Since for each element of the string we have 2 choices (either 0 or 1), there are 2^n strings. This shows that a set of n elements has 2^n subsets (hence the previously introduced notation 2^Ω).

2.3.2 Permutations

A **permutation** of a set A is a bijection from A to itself. In other words, it is an ordering of the elements of A. How many possible permutations of a set of n elements are there?

Label the elements of A as $\{1, 2, \dots, n\}$. We may obtain all permutations by subsequently choosing the image of element 1, then the image of element 2 and so on. We have n choices for the image of 1, then $n-1$ choices for the image of 2, $n-2$ choices for the image of 3 until we have only one choice for the image of n. Thus the total number of choices is, by the multiplication rule, $n! = n(n-1)(n-2)\cdots 1$. Thus there are $n!$ different permutations, or orderings, of n elements. Equivalently, there are $n!$ different bijections from any two sets of n elements.

Example 2.3.2. There are $52!$ possible orderings of a standard deck of cards.

2.3.3 Subsets

Let us go back to Example 2.2.13. How did we know there are exactly six possible ways of arranging the cups? Well, that corresponds to the number of ways of placing the cups of color a. Indeed, for each such a choice, the positions of the cups of color b are forced. But then this is the general problem of counting the number of ways a subset of size k can be chosen from a set of size $n \geq k$ or, equivalently, the number of subsets of size k of a set of size n.

Let us first count the number of **ordered** subsets of size k of a set of size n. We have n choices for the element in first position. For each such a choice, we have $n-1$ choices for the element in second position and so on up to $n-(k-1)$ choices for the last element in position k. Overall,

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!} \tag{2.1}$$

ordered subsets. An alternative way to obtain the above formula is the following: to pick k ordered elements out of n, first pick a permutation of the n elements ($n!$ choices), then forget all elements but the first k. Since for each choice of the first k elements there are $(n-k)!$ permutations starting with those k elements, we obtain again Equation (2.1).

On the other hand, if we are interested in **unordered** subsets, then Equation (2.1) *over-counts*: every subset is counted exactly $k!$ times (with every possible ordering of its elements). So we have to divide by $k!$.

Lemma 2.3.3. *The number of subsets of size k of a set of size n is*

$$\frac{n!}{k!(n-k)!},$$

*denoted by $\binom{n}{k}$. The numbers $\binom{n}{k}$ are called **binomial coefficients**.*

More generally, suppose we have integers $n_1, n_2, \ldots, n_k$ with $n_1 + n_2 + \cdots + n_k = n$. Then the number of ways to partition n elements into k subsets of cardinalities $n_1, \ldots, n_k$ is

$$\frac{n!}{n_1! \cdots n_k!},$$

denoted by $\binom{n}{n_1,\ldots,n_k}$. The numbers $\binom{n}{n_1,\ldots,n_k}$ are called **multinomial coefficients**.

Example 2.3.4. Suppose that in a city of n people there are and two residents A and B each having k friends in the city. Let us assume that the friends for each of A and B are selected at random. Then there are $\binom{n-1}{k}$ ways to select k friends for A (we don't count A as his/her own friend) and $\binom{n-1}{k}$ ways to select k friends for B.

Now, in how many cases A and B are not friends and do not even have a common friend? Well, we can choose k friends for A in $\binom{n-2}{k}$ ways (we are not choosing A or B as friends of A) and then k friends for B in $\binom{n-k-2}{k}$ ways.

The following result explains the name "binomial coefficient":

Theorem 2.3.5 (Binomial theorem). *The coefficient of $x^{n-k}y^k$ in the expansion of $(x+y)^n$ is $\binom{n}{k}$. In other words, the following identity holds:*

$$(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \cdots + \binom{n}{n-1}xy^{n-1} + \binom{n}{n}y^n.$$

Proof. Think of expanding

$$(x+y)^n = (x+y)(x+y)\cdots(x+y)$$

so that we get rid of all parentheses. We get each term in the expansion by selecting one of the two terms in each factor, and multiplying all the selected terms. If we choose x exactly $n-k$ times, then we must choose y exactly k times and we get a term of the form $x^{n-k}y^k$. How many times do we get this same term? Clearly, as many times as the number of ways to select the k factors that supply y (the remaining factors supply x). This can be done in $\binom{n}{k}$ ways. As a side remark, this argument also shows that $\binom{n}{k} = \binom{n}{n-k}$. □

Exercise 2.3.6. *Prove the Binomial theorem by induction.*

2.3.4 Subsets with repetitions

How many ways are there to choose k elements from a set of n elements, allowing repetitions?

Consider first the **ordered** case. We have n choices for the first element, n choices for the second element and so on. Thus there are n^k possible ways to choose k ordered elements from n, allowing repetitions.

Consider now the **unordered** case i.e., we want to choose k elements from n, allowing repetitions but discarding the order. How many ways do we have to do so? Naively dividing n^k by $k!$ doesn't give the right answer, since there may be repetitions. Instead, we count as follows. Label the n elements $\{1,\dots,n\}$ and for each element draw a $*$ each time it is picked. Note that there are k $*$'s and $n-1$ vertical lines. An example is the following

$$* * \mid * \mid \mid \cdots \mid * * *, \tag{2.2}$$

where the element 1 is picked two times, the element 2 is picked one time, the element 3 is picked zero times and so on. The above diagram uniquely identifies an unordered set of (possibly repeated) k elements. Thus we simply have to count how many such diagrams there are. The only restriction is that there must be $n-1$ vertical lines and k $*$'s. Since there are $n+k-1$ locations, we can fix such a diagram by assigning the positions of the $*$'s, which can be done in $\binom{n+k-1}{k}$ ways. This therefore counts the number of unordered subsets of k elements from n, without ordering.

Example 2.3.7 (Ordered partitions). An ordered partition of k of size n is an n-tuple $(k_1, k_2, \dots, k_n)$ of non-negative integers such that $k_1 + \cdots + k_n = k$. How many ordered partitions of k of size n are there?

We give a graphic representation of each such partition as follows: draw k_1 $*$'s followed by a vertical line, then k_2 $*$'s followed by another vertical line and so on. For example, $(1,0,3,2)$ is represented by $* \mid \mid * * * \mid * *$. Note the similarity with Equation (2.2). Now, since this determines a bijection, it again suffices to count the number of diagrams made of k $*$'s and $n-1$ vertical lines, which is $\binom{n+k-1}{k}$.

Example 2.3.8. A Bank issues bank cards with PINs consisting of 4 digits in $\{0,1,2,\dots,9\}$. How many unique PINs are there if:

1. any 4-digit code can be used;
2. the digits must be different;
3. the digits must be different and must not be in ascending or descending order?

We proceed as follows:

1. There are 10 choices for the first digit, 10 choices for the second digit and so on, giving a total of 10^4 unique PINs.

2. The number of PINs with different digits is $10 \times 9 \times 8 \times 7 = 5040$.

3. Consider a PIN on 4 different digits, say $\{1, 3, 5, 7\}$. There are $4!$ different such PINs and there is one in ascending order $1, 3, 5, 7$ and one in descending order $7, 5, 3, 1$. So there are $4! - 2 = 22$ desired PINs. This result applies to any set of 4 different digits. Since the number of such sets is $\binom{10}{4} = 210$, the number of unique PINs is $210 \times 22 = 4620$.

Example 2.3.9. Recall that bridge is played with a pack of 52 cards divided into 4 suits (clubs, diamonds, hearts, spades) where each suit has 13 denominations (Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King). In a game of bridge, what is the probability of a given player being dealt a hand (of 13 cards)

1. containing entirely one suit,

2. containing exactly five spades,

from a well-shuffled pack?

1. The number of different bridge hands is the same as the number of ways of selecting 13 cards from 52 distinct cards, where the order is not important, namely $\binom{52}{13}$. Each hand is an outcome and the $\binom{52}{13}$ outcomes are assumed equally likely. Since there are exactly 4 hands consisting entirely of one suit, the desired probability is $4/\binom{52}{13} \approx 6.3 \times 10^{-12}$.

2. Let us count the favourable outcomes. There are $\binom{13}{5}$ ways of choosing the 5 spades and $\binom{39}{8}$ ways of choosing the 8 non-spades. By the multiplication rule, we then have a total of $\binom{13}{5}\binom{39}{8}$ hands with exactly 5 spades. The desired probability is then
$$\frac{\binom{13}{5}\binom{39}{8}}{\binom{52}{13}} \approx 0.125.$$

Example 2.3.10. What is the probability that, in 6 throws of a fair die, all faces turn up?

Denoting by x_i the score on the i-th throw, an outcome is the ordered set $(x_1, x_2, \ldots, x_6)$. There are 6^6 equally likely outcomes. The favourable outcomes are the permutations of $\{1, 2, \ldots, 6\}$. Since there are $6!$ such permutations, the desired probability is $6!/6^6$.

Exercise 2.3.11. *(a) In how many ways can we arrange* 10 *people in a row?*

(b) Suppose that one of these 10 *people is John and another is Jack. In how many ways can we arrange these* 10 *people in a row so that Jack is next to John?*

(c) In how many ways can we arrange these 10 *people in a row so that Jack is to the right of John, though not necessarily next right?*

Exercise 2.3.12. *(a) In how many ways can we distribute* 10 *different marbles among* 5 *different boxes?*

(b) In how many ways can we distribute 10 *indistinguishable marbles among* 5 *different boxes?*

2.4 Properties of probability measures

After our counting detour, we now go back to a generic probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The goal is to derive several useful properties of probability measures from the probability axioms.

Lemma 2.4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ be events. The following are true:*

(a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(b) If $A \subseteq B$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$ (monotonicity);

(c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *(inclusion-exclusion).*

Proof. Notice first that, as observed in Remark 2.1.2, all the sets considered belong to $\mathcal{F}$ and so we can indeed talk about their probabilities.

(a) Since $A \cap A^c = \varnothing$ and $A \cup A^c = \Omega$, finite additivity and the 2nd axiom imply that $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$.

(b) Since $A \cap (B \setminus A) = \varnothing$ and $A \cup (B \setminus A) = B$, finite additivity and the 2nd axiom imply that $\mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$. Since $\mathbb{P}(B \setminus A) \geq 0$ (1st axiom), we then have that $\mathbb{P}(B) \geq \mathbb{P}(A)$.

(c) $A \cup B$ can be written as the disjoint union $A \cup (B \setminus A)$. We then have that,

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

where in the first equality we use additivity, in the second the fact that $B \setminus A = B \setminus (A \cap B)$ and in the third (b). □

Example 2.4.2. In a population of 1000 people, 10% are left-handed, 5% are color-blind, and of these 10 are left-handed. A person is selected at random from the population. What is the probability of a left-handed or color-blind person (or both) being selected?

Let LH and CB be the event that the selected person is left-handed and colour-blind, respectively. The desired probability is

$$\mathbb{P}(LH \cup CB) = \mathbb{P}(LH) + \mathbb{P}(CB) - \mathbb{P}(LH \cap CB) = \frac{100}{1000} + \frac{50}{1000} - \frac{10}{1000} = \frac{140}{1000} = 0.14.$$

Example 2.4.3. A fair coin is tossed 5 times. What is the probability of getting at least one head?

We can take as our sample space the set $\Omega = \{HHHHH, HHHHT, \dots, TTTTT\}$. By fairness, we may assume that all outcomes are equally likely. Let A be the event that no head is obtained. The desired probability is then

$$1 - \mathbb{P}(A) = 1 - \frac{1}{2^5} = \frac{31}{32}.$$

Example 2.4.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ be events. Although $\mathbb{P}(A) = \mathbb{P}(A \cap B)$ is obviously false in general, it is true if $\mathbb{P}(B) = 1$. Indeed, in this case

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = \mathbb{P}(A) + (1 - \mathbb{P}(A \cup B)) \geq \mathbb{P}(A).$$

The reverse inequality always holds by monotonicity.

The reader will have the chance to see the following result in analysis modules: If $f : \mathbb{R} \to \mathbb{R}$ is a continuous function at x_0 and the sequence $x_1, x_2, \dots$ converges to x_0, then the sequence $f(x_1), f(x_2), \dots$ converges to $f(x_0)$. A similar statement holds for probability measures.

Lemma 2.4.5 (Continuity of probability). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For every increasing sequence of events $A_1, A_2, \ldots$ (i.e., $A_1 \subseteq A_2 \subseteq \cdots$), we have that*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

Similarly, for every decreasing sequence of events $B_1, B_2, \ldots$ (i.e., $B_1 \supseteq B_2 \supseteq \cdots$), we have that

$$\mathbb{P}\Big(\bigcap_{i=1}^{\infty} B_i\Big) = \lim_{i\to\infty} \mathbb{P}(B_i).$$

Example 2.4.6. It is intuitively clear that the chance of obtaining no heads in an infinite sequence of tosses of a fair coin is 0. A rigorous proof goes as follows. Let A_i be the event that the first i tosses of the coin yield at least one head. Then $A_i \subseteq A_{i+1}$ for each $i \geq 1$, so that $A_1, A_2, \ldots$ is an increasing sequence. The event $A = \bigcup_{i=1}^{\infty} A_i$ is nothing but the event that heads occurs sooner or later i.e., it is the complement of the event we are interested in. By continuity of probability,

$$\mathbb{P}(A) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

However,

$$\mathbb{P}(A_i) = 1 - \Big(\frac{1}{2}\Big)^i,$$

and so $\mathbb{P}(A) = \lim_{i\to\infty} \mathbb{P}(A_i) = 1$, giving that the probability $\mathbb{P}(A^c)$ that no head ever appears is 0.

The following result, despite its simplicity, is extremely useful in probability theory. It asserts that the probability that at least one event in a sequence occurs can not exceed the sum of the probabilities of the events in the sequence.

Lemma 2.4.7 (Union bound). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \ldots$ be a sequence of events. Then*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. We only show the finite version of the statement, namely that, for each n,

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

We proceed by induction on n. The base case $n = 1$ is trivial. Therefore, suppose that $A_1, \ldots, A_{n+1} \in \mathcal{F}$. We define $A = A_1 \cup \cdots \cup A_n$ and $B = A_{n+1} \setminus A$. Then $A_1 \cup \cdots \cup A_{n+1}$ can be written as the disjoint union $A \cup B$. But then

$$\mathbb{P}(A_1 \cup \cdots \cup A_{n+1}) = \mathbb{P}(A) + \mathbb{P}(B) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(B) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(A_{n+1}) = \sum_{i=1}^{n+1} \mathbb{P}(A_i),$$

where the first equality follows from finite additivity, the first inequality follows from the induction hypothesis and the last inequality follows from monotonicity. This concludes the proof by induction. □

Exercise 2.4.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A_1, A_2, \dots$ be a countable family of events. Show that if $\mathbb{P}(A_i) = 1$, for each i, then $\mathbb{P}(\bigcap_{i=1}^{\infty} A_i) = 1$. Similarly, show that if $\mathbb{P}(A_i) = 0$, for each i, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = 0$.*

As mentioned, the union bound is a simple and yet extremely useful tool in probability. We now see it in action in the so-called **probabilistic method**.

Example 2.4.9. Our motivating question is the following: Will an arbitrary group of 6 members of a social network necessarily contain a subgroup of 3 mutual friends or a subgroup of 3 mutual strangers? Perhaps surprisingly, the answer is "Yes". A group of 5 individuals, however, does not necessarily have this property. We can model and generalize this problem via a graph, an ubiquitous object in computer science and operations research. So what is a graph? Informally speaking (which is enough for us), a **graph** is a set of points, called vertices, connected by lines, called edges. The complete graph K_n is the graph on n vertices such that any two vertices are connected by an edge. A two-coloring of the edges of K_n is an assignment of colors to its edges so that each edge is colored either red or blue.

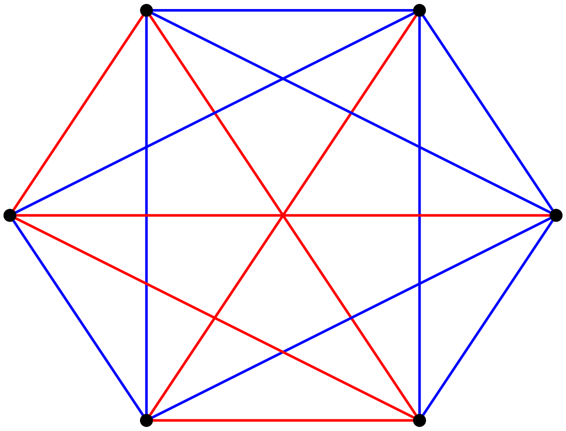


Figure 2.2: A two-coloring of K_6.

We can encode the fact of being friends by a red line and the fact of being strangers by a blue line. Therefore, generalizing our motivating question, we might ask: Does K_n always contain a monochromatic K_k i.e., a red K_k or a blue K_k, for any two-coloring? Frank Ramsey answered this question in his celebrated theorem:

Theorem 2.4.10 (Ramsey's theorem). *For any $k \geq 2$, there is a finite value of n for which any two-coloring of K_n contains a monochromatic K_k and so there is a smallest such value n, called the Ramsey number $R(k, k)$.*

We have remarked that $R(3, 3) = 6$. This is in fact not difficult to show (try!) but as soon as the value of k increases, determining $R(k, k)$ has proved to be an extremely difficult problem. At the moment, we do not even know $R(5, 5)$; we just know that it is between 43 and 48. But can we say anything about how quickly Ramsey numbers grow with k? In a seminal paper from 1947 that gave birth to what is now called the probabilistic method, Erdős showed how a lower bound for $R(k, k)$ may be obtained almost effortlessly using a probabilistic argument.

Roughly speaking, the probabilistic method works as follows: Trying to prove that a structure with a certain desired property exists, one defines an appropriate probability space of structures and then shows that the desired property holds in this space with positive probability. The method is best illustrated in action.

Consider a *random* two-coloring of K_n. The sample space is the set of all possible two-colorings of K_n. How many such colorings are there? Well, each edge can be colored either red or blue and since there are $\binom{n}{2}$ edges, we have $2^{\binom{n}{2}}$ possible colorings, where in *random* we assume that each has equal probability $2^{-\binom{n}{2}}$.

Let S be any fixed set of k vertices in K_n and let A_S be the event that S forms a monochromatic K_k. Then A_S is the union of the disjoint events $\{K_k \text{ is red}\}$ and $\{K_k \text{ is blue}\}$ and so

$$\mathbb{P}(A_S) = 2^{1-\binom{k}{2}}.$$

Let's now look at the event $\bigcup_{S\colon |S|=k} A_S$ that there is at least one monochromatic K_k. We can estimate its probability by the union bound:

$$\mathbb{P}\Big(\bigcup_{S\colon |S|=k} A_S\Big) \leq \sum_{S\colon |S|=k} \mathbb{P}(A_S) = \binom{n}{k} 2^{1-\binom{k}{2}}.$$

Therefore, if $r(k)$ denotes the largest integer n satisfying $\binom{n}{k}2^{1-\binom{k}{2}} < 1$, then

$$\mathbb{P}\Big(\bigcup_{S\colon |S|=k} A_S\Big) < 1$$

and so there must be *some* two-coloring of $K_{r(k)}$ without any monochromatic K_k i.e., $R(k,k) > r(k)$. One could in fact find an estimate for $r(k)$.

2.5 Conditional probability

Conditional probability provides us with a way to reason about the outcome of an experiment based on partial information. Suppose a certain experiment is repeated N times. On each trial we observe the occurences or non-occurences of two events A and B. Suppose we are interested only in the outcomes for which B occurs; all other trials are disregarded. The proportion of times that A occurs in this smaller collection of trials is $N(A \cap B)/N(B)$ and

$$\frac{N(A \cap B)}{N(B)} = \frac{N(A \cap B)/N}{N(B)/N}.$$

As these ratios can be thought as approximations for the probabilities, the probability that A occurs given that B occurs should intuitively be $\mathbb{P}(A \cap B)/\mathbb{P}(B)$.

Definition 2.5.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The **conditional probability** that A occurs given that B occurs is the value

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

We stress the fact that this is a definition. The next result justifies the term conditional probability:

Lemma 2.5.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $B \in \mathcal{F}$ be such that $\mathbb{P}(B) > 0$. The function $P\colon \mathcal{F} \to \mathbb{R}$ defined by $P(A) = \mathbb{P}(A|B)$ is a probability measure.*

Proof. We need to verify that the function P satisfies the three properties in Definition 2.2.1:

1. Let $A \in \mathcal{F}$. By monotonicity, $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$ and so $0 \leq \mathbb{P}(A \cap B)/\mathbb{P}(B) \leq 1$.
2. $\mathbb{P}(\varnothing|B) = \mathbb{P}(\varnothing \cap B)/\mathbb{P}(B) = 0$ and $\mathbb{P}(\Omega|B) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = 1$.
3. Let $A_1, A_2, \ldots$ be a family of pairwise disjoint events. Then

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i|B\Big) = \frac{\mathbb{P}\Big(\big(\bigcup_{i=1}^{\infty} A_i\big) \cap B\Big)}{\mathbb{P}(B)} = \frac{\mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A_i \cap B)\Big)}{\mathbb{P}(B)}$$
$$= \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}$$
$$= \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$
$$= \sum_{i=1}^{\infty} P(A_i),$$

where the first equality follows from the definition of P, the second from the definition of conditional probability, the third from the distributive property, the fourth from countable additivity for the probability measure $\mathbb{P}$, the fifth from the definition of conditional probability and the last from the definition of P. □

Lemma 2.5.2 implies that we can apply all the tools developed so far to conditional probabilities.

Example 2.5.3 (Equally likely outcomes). Let Ω be a finite set with equally likely outcomes i.e., $\mathbb{P}(A) = |A|/|\Omega|$ for each $A \subseteq \Omega$. Then, for any non-empty $B \subseteq \Omega$, we have

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

This means that, in the case of equally likely outcomes, the conditional probability of A given B counts the proportion of outcomes in B that belong to A. It also suggests that conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.

Example 2.5.4. A fair die is thrown. What is the probability of a 2 given that an even number has occurred?

Letting A to be the event that 2 is thrown and B be the event that an even number is thrown, Example 2.5.3 implies that the desired probability is $1/3$.

In many situations it is natural to assign values to some conditional probabilities and, from them, deduce the values of non-conditional probabilities.

Example 2.5.5. A student can't decide whether to study history or literature. If he takes literature, he will pass with probability $1/2$; if he takes history, he will pass with probability $1/3$. He made his decision based on a coin toss. What is the probability that he opted for history and passed the exam?

As a sample space we take {history, literature} × {pass, fail}. If A is the event that he passed, then $A =$ {history, literature} × {pass}. If B denotes the event that he opted for history, then {history} × {pass, fail}. We have

$$\mathbb{P}(B) = \mathbb{P}(B^c) = \frac{1}{2}, \quad \mathbb{P}(A|B) = \frac{1}{3}, \quad \mathbb{P}(A|B^c) = \frac{1}{2}$$

and so $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = 1/6$. Notice that making the sample space explicit was in fact not crucial in this case, as often happens with conditional probabilities.

When we are dealing with an event A which occurs if and only if each one of several events $A_1, \ldots, A_n$ has occurred i.e., $A = A_1 \cap A_2 \cap \cdots \cap A_n$, we can view the occurrence of A as the occurrence of A_1, followed by the occurrence of A_2, then of A_3 and so on. The probability of A can then be computed using the following rule:

Lemma 2.5.6 (Multiplication rule for probabilities). *Assuming that all the following conditional probabilities are positive, we have*

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i\Big) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\cdots\mathbb{P}\Big(A_n|\bigcap_{i=1}^{n-1} A_i\Big).$$

Proof.

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i\Big) &= \mathbb{P}(A_1)\cdot\frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)}\cdot\frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)}\cdot\cdots\cdot\frac{\mathbb{P}\big(\bigcap_{i=1}^{n} A_i\big)}{\mathbb{P}\big(\bigcap_{i=1}^{n-1} A_i\big)}\\ &= \mathbb{P}(A_1)\cdot\mathbb{P}(A_2|A_1)\cdot\mathbb{P}(A_3|A_1 \cap A_2)\cdot\cdots\cdot\mathbb{P}\Big(A_n|\bigcap_{i=1}^{n-1} A_i\Big).\end{aligned}$$ □

The multiplication rule is particularly useful in problems involving sequential operations.

Example 2.5.7. Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn.

One possible approach, called *parallel*, is to count the number of all card triplets that do not include a heart and divide it by the number of all possible card triplets (see Exercise 2.5.8). However, in this case, it is more convenient to adopt a *sequential* approach. For each $i \in \{1, 2, 3\}$, let A_i be the event that the i-th card is not a heart. We compute the desired probability $\mathbb{P}(A_1 \cap A_2 \cap A_3)$ that none of the three cards is a heart using the multiplication rule:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2).$$

We have $\mathbb{P}(A_1) = 39/52$, since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards, 38 of which are not hearts, and so $\mathbb{P}(A_2|A_1) = 38/51$. Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck and $\mathbb{P}(A_3|A_1 \cap A_2) = 37/50$. The desired probability is then

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{39}{52}\cdot\frac{38}{51}\cdot\frac{37}{50}.$$

Exercise 2.5.8. *Use the parallel approach to compute the probability in Example 2.5.7.*

Example 2.5.9. An urn contains 10 white balls, 9 red balls and 8 black balls. Three balls are removed. We assume that the removal of a ball (or a set of balls) is such that each ball (or each set) has the same chance of being chosen. What is the probability that all chosen balls are of different colors?

Let's adopt the parallel approach first by considering the balls to be removed simultaneously. The number of ways of choosing 3 balls of distinct colors is $\binom{10}{1}\binom{9}{1}\binom{8}{1}$, whereas the number of ways of choosing 3 balls is $\binom{27}{3}$. Therefore, the desired probability is $\binom{10}{1}\binom{9}{1}\binom{8}{1}/\binom{27}{3} = 16/65$.

We can check that the same result is obtained via the sequential approach by considering the balls to be removed one at a time, without replacement. For $i \in \{1,2,3\}$, let W_i be the event that the i-th ball chosen is white. Similarly, define R_i and B_i. The desired probability is then

$$\mathbb{P}((W_1\cap R_2\cap B_3)\cup(W_1\cap B_2\cap R_3)\cup(R_1\cap W_2\cap B_3)\cup(R_1\cap B_2\cap W_3)\cup(B_1\cap R_2\cap W_3)\cup(B_1\cap W_2\cap R_3)).$$

But

$$\mathbb{P}(W_1\cap R_2\cap B_3) = \mathbb{P}(W_1)\mathbb{P}(R_2|W_1)\mathbb{P}(B_3|W_1\cap R_2) = \frac{10}{27}\cdot\frac{9}{26}\cdot\frac{8}{25} = \frac{8}{195}.$$

Similarly, for distinct $i, j, k \in \{1,2,3\}$, we have

$$\mathbb{P}(W_i\cap R_j\cap B_k) = \frac{8}{195}.$$

Therefore, by finite additivity, the desired probability is $6\cdot\frac{8}{195} = \frac{16}{65}$.

Definition 2.5.10. Given a countable infinite collection of events $B_1, B_2, \ldots$, we say that the collection is a **partition** of Ω if $B_i \cap B_j = \varnothing$, for each $i \neq j$, and $\bigcup_{i=1}^{\infty} B_i = \Omega$. The same definition applies mutatis mutandis in the case of a finite collection.

Lemma 2.5.11 (Law of total probability). *Given a partition $B_1, B_2, \ldots$ of Ω such that $\mathbb{P}(B_i) > 0$ for each i, then*

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

A similar result holds in case the collection $B_1, B_2, \ldots, B_n$ is finite.

Proof. By definition of partition of Ω, we have that $\bigcup_{i=1}^{\infty} B_i = \Omega$. Therefore,

$$\mathbb{P}(A) = \mathbb{P}(A\cap\Omega) = \mathbb{P}\Big(A\cap\Big(\bigcup_{i=1}^{\infty}B_i\Big)\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty}(A\cap B_i)\Big) = \sum_{i=1}^{\infty}\mathbb{P}(A\cap B_i) = \sum_{i=1}^{\infty}\mathbb{P}(A|B_i)\mathbb{P}(B_i),$$

where the third equality follows from the distributive property, the fourth from countable additivity for the probability measure $\mathbb{P}$ and the last from the definition of conditional probability. □

Remark 2.5.12. The condition of positive probability may be omitted provided we agree to interpret $\mathbb{P}(A|B_i)\mathbb{P}(B_i)$ as 0 whenever $\mathbb{P}(B_i) = 0$.

The law of total probability is typically used as follows. Suppose we want to compute the probability that A occurs. Let B be another arbitrary event with $0 < \mathbb{P}(B) < 1$. There are two scenarios: either B or B^c occurs. If we know the probability of the two scenarios and the probability of A conditioned on each of them, then we can compute the probability of A.

Example 2.5.13. Tomorrow there will be either rain or snow but not both; the probability of rain is $\frac{2}{5}$ and the probability of snow is $\frac{3}{5}$. If it rains, the probability that I will be late for my lecture is $\frac{1}{5}$, while the corresponding probability in the event of snow is $\frac{3}{5}$. What is the probability that I will be late?

Let A be the event that I am late and let B be the event that it rains. The pair B, B^c is a partition of the sample space (since exactly one of them must occur). By the Law of total probability,

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) = \frac{1}{5} \cdot \frac{2}{5} + \frac{3}{5} \cdot \frac{3}{5} = \frac{11}{25}.$$

Example 2.5.14. An urn contains b black balls and r red balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is black?

Let A be the event that the second ball drawn is black and let B be the event that the first ball drawn is black. Then

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) = \frac{b-1}{b+r-1} \cdot \frac{b}{b+r} + \frac{b}{b+r-1} \cdot \frac{r}{b+r} = \frac{b}{b+r}.$$

Lemma 2.5.15 (Bayes' law). *Let A and B be two events such that $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then*

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

Proof. Exercise! □

Bayes' law tells how to update the estimate of the probability of A when new evidence restricts the sample space to B. The ratio $\mathbb{P}(B|A)/\mathbb{P}(B)$ determines "how compelling the new evidence is".

Combining Bayes' law with the law of total probability we obtain that, if $B_1, B_2, \ldots$ is a partition of Ω such that $\mathbb{P}(B_i) > 0$ for each i, then

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

Remark 2.5.16. Again, we can drop the assumption that the B_i's have positive probability by setting $\mathbb{P}(A|B_i)\mathbb{P}(B_i) = 0$ if $\mathbb{P}(B_i) = 0$.

Example 2.5.17. Going back to Example 2.5.14, suppose we are told that the second ball is black. What is the probability that the first ball was black?

Applying Bayes' theorem,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)} = \left(\frac{b}{b+r} \cdot \frac{b-1}{b+r-1}\right)\left(\frac{b}{b+r}\right)^{-1} = \frac{b-1}{b+r-1}.$$

Example 2.5.18. Consider a lab screen for a certain virus. A person that carries the virus is screened positive in only 95% of the cases (5% chance of false negative). A person who does not carry the virus is screened positive in 1% of the cases (1% chance of false positive). Given that 0.5% of the population carries the virus, what is the probability that a person who has been screened positive is actually a carrier?

We take $\Omega = \{\text{carrier, not carrier}\} \times \{+, -\}$. Let A be the event "the person is a carrier" i.e., $A = \{\text{carrier}\} \times \{+, -\}$, and let B be the event "the person was screened positive" i.e., $B = \{\text{carrier, not carrier}\} \times \{+\}$. We are given the following information

$$\mathbb{P}(A) = 0.005 \qquad \mathbb{P}(B|A) = 0.95 \qquad \mathbb{P}(B|A^c) = 0.01.$$

Therefore, taking A, A^c as our partition of Ω, we have that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} = \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \approx \frac{1}{3}.$$

Example 2.5.19. A random number N of dice is thrown. Let A_i be the event $\{N = i\}$ and suppose that $\mathbb{P}(A_i) = 1/2^i$ $(i \geq 1)$. The sum of the scores is S. Compute $\mathbb{P}(N = 2|S = 4)$.

By Bayes' law,

$$\mathbb{P}(N = 2|S = 4) = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\mathbb{P}(S = 4)}.$$

The (countable) family of events $N = 1, N = 2, N = 3, \ldots$ is a partition of Ω such that $\mathbb{P}(N = i) = 1/2^i > 0$. Therefore, by the law of total probability,

$$\mathbb{P}(S = 4) = \sum_{i=1}^{\infty} \mathbb{P}(S = 4|N = i)\mathbb{P}(N = i).$$

However, only the first four terms of the sum are non-zero. Indeed, if $i \geq 5$, then $\mathbb{P}(S = 4|N = i) = 0$. Therefore, the desired probability is

$$\mathbb{P}(N = 2|S = 4) = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\mathbb{P}(S = 4)} = \frac{\mathbb{P}(S = 4|N = 2)\mathbb{P}(N = 2)}{\sum_{i=1}^{4} \mathbb{P}(S = 4|N = i)\mathbb{P}(N = i)}.$$

We are then left to compute $\mathbb{P}(S = 4|N = i)$, for $i \in \{1, 2, 3, 4\}$. Let's consider $\mathbb{P}(S = 4|N = 3)$, the other cases being similar. This is the probability of getting a sum of 4 by throwing 3 dice. As usual, label the dice $1, 2, 3$ and let x_i be the number on die i. There are 6^3 possible outcomes and we need to count how many triples (x_1, x_2, x_3) are such that $x_1 + x_2 + x_3 = 4$. Since each x_i is at least 1, it is easy to see there are exactly 3 such triples, namely $(1, 1, 2), (1, 2, 1), (2, 1, 1)$. Therefore, $\mathbb{P}(S = 4|N = 3) = 3/6^3$.

Exercise 2.5.20. *Consider n indistinguishable balls randomly distributed in m boxes. What is the probability that exactly k boxes remain empty?*

Exercise 2.5.21. *We roll a fair four-sided die. If the result is 1 or 2, we roll once more, otherwise we stop. What is the probability that the total sum of our rolls is at least 4?*

Exercise 2.5.22. *An urn contains b blue balls and c cyan balls. A ball is drawn at random, its color noted and it is returned to the urn together with d further balls of the same color. The process is repeated indefinitely.*

- *Compute the probability that the second ball drawn is cyan.*
- *Compute the probability that the first ball drawn is cyan given that the second ball drawn is cyan.*

Exercise 2.5.23. *You are travelling on a train with your sister. Neither of you has a valid ticket, and the inspector has caught you both. He is authorized to administer a special punishment for this offence. He holds a box containing nine apparently identical chocolates, three of which are contaminated with a deadly poison. He makes each of you, in turn, choose and immediately eat a single chocolate.*

(a) If you choose before your sister, what is the probability that you will survive?

(b) *If you choose first and survive, what is the probability that your sister survives?*

(c) *If you choose first and die, what is the probability that your sister survives?*

(d) *Is it in your best interests to persuade your sister to choose first?*

(e) *If you choose first, what is the probability that you survive, given that your sister survives?*

The following example shows that an apparently irrelevant information might change probabilities in surprising ways.

Example 2.5.24 (The two children paradox). Consider the three statements:

(a) I have two children, the elder of whom is a boy;

(b) I have two children, one of whom is a boy;

(c) I have two children, one of whom is a boy born on a Thursday.

What is the probability that both children are boys in the case (a), (b) and (c)?

Since we have no further information, we will assume all outcomes are equally likely. Write BG for the event that the elder is a boy and the younger a girl (similar definitions for GB and BB). Write GT for the event that the elder is a girl and the younger a boy born on a Thursday, and write TN for the event that the elder is a boy born on a Thursday and the younger a boy born on another day (similar definitions for NT, TT and TG). Then

(a) $\mathbb{P}(BB|BG \cup BB) = 1/2$;

(b) $\mathbb{P}(BB|BB \cup BG \cup GB) = 1/3$;

(c) $\mathbb{P}(NT \cup TN \cup TT|NT \cup TN \cup TT \cup TG \cup GT) = 13/27$,

where in (c) we used the fact that

$$\mathbb{P}(NT \cup TN \cup TT) = \frac{6}{14} \cdot \frac{1}{14} + \frac{1}{14} \cdot \frac{6}{14} + \frac{1}{14} \cdot \frac{1}{14}$$

and

$$\mathbb{P}(NT \cup TN \cup TT \cup TG \cup GT) = \frac{6}{14} \cdot \frac{1}{14} + \frac{1}{14} \cdot \frac{6}{14} + \frac{1}{14} \cdot \frac{1}{14} + \frac{1}{14} \cdot \frac{7}{14} + \frac{7}{14} \cdot \frac{1}{14}.$$

Thus, learning about the gender of one child biases the probabilities for the other. Also, learning a seemingly irrelevant additional fact pulls the probabilities back towards evens.

Example 2.5.25 (Simpson's paradox). Given two events A and B with $\mathbb{P}(B) > 0$, we say that B attracts A if $\mathbb{P}(A|B) > \mathbb{P}(A)$. For a further event S with $\mathbb{P}(B \cap S) > 0$, we say that B attracts A on S if $\mathbb{P}(A|B \cap S) > \mathbb{P}(A|S)$. With this terminology, we might expect that if B attracts A both on S and on S^c, then B attracts A. The following example shows that this is false.

The interval $\Omega = (0, 1]$ can be equipped with a probability measure $\mathbb{P}$ such that $\mathbb{P}((a, b]) = b - a$ for all $0 \leq a \leq b \leq 1$. For $\varepsilon \in (0, 1/4)$, define the events $A = (\varepsilon/2, 1/2 + \varepsilon/2]$, $B = (1/2 - \varepsilon/2, 1 - \varepsilon/2]$, $S = (0, 1/2]$, $S^c = (1/2, 1]$.

Clearly, S and S^c are disjoint events such that $S \cup S^c = \Omega$. We claim that B attracts A on S and on S^c and yet B does not attract A. Indeed,

$$\mathbb{P}(A|B \cap S) = \frac{\mathbb{P}(A \cap B \cap S)}{\mathbb{P}(B \cap S)} = 1 > \mathbb{P}(A|S) = 1 - \varepsilon,$$

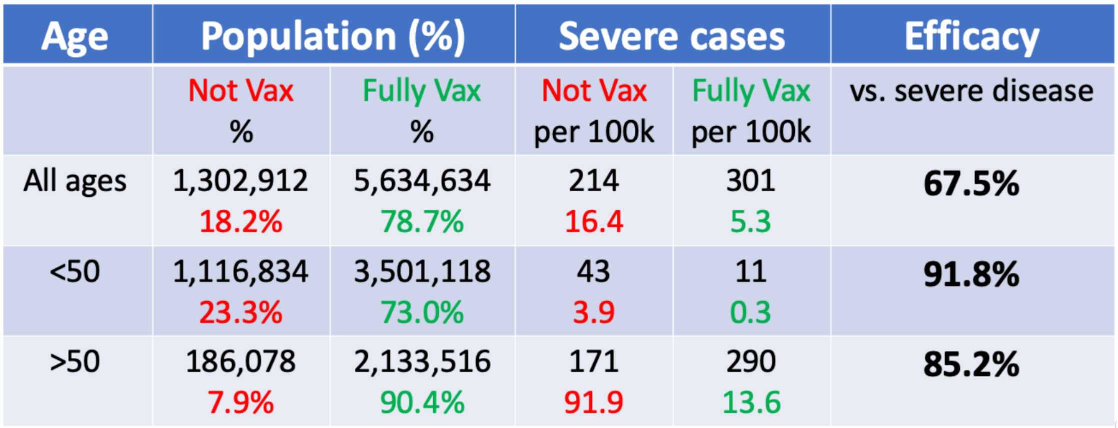
$$\mathbb{P}(A|B \cap S^c) = \frac{\mathbb{P}(A \cap B \cap S^c)}{\mathbb{P}(B \cap S^c)} = \frac{\varepsilon}{1-\varepsilon} > \mathbb{P}(A|S^c) = \varepsilon.$$

On the other hand,

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

The paradoxical outcome comes from the fact that while $\mathbb{P}(A) = 1/2$, $\mathbb{P}(A|B) = 2\varepsilon$, so conditioning on B significantly alters the probability of A.

More generally, Simpson's paradox refers to any instance where a positive (or negative) association between events, when conditioned by the elements of a partition, is reversed when the same events are considered without conditioning. As a concrete example of this, consider vaccine effectiveness vs severe disease for COVID-19.



Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

Figure 2.3: Vaccine effectiveness vs severe disease (Israeli data from https://www.covid-datascience.com).

Looking at the table above, one can notice that effectiveness is quite high in both < 50 and > 50 cohorts. However, these effectiveness levels are much higher than the 67.5% estimate we get if the analysis is not stratified by age. This discrepancy is an instance of the Simpson's paradox: misleading results can sometimes be obtained from observational data in the presence of *confounding* factors. In our case, age is a confounding factor. It is the fact that both vaccination status and risk of severe disease are systematically higher in the older age group that makes overall effectiveness numbers if estimated without stratifying by age misleading, producing a paradoxical result.

Exercise 2.5.26. *We have a box with 15 balls, some of which are red and the rest are blue. We pick a ball at random, note its color and put it back together with an additional ball of the same color. Then we pick a ball again. How many blue balls were in the box initially if the ball we pick the second time is blue with probability $1/3$?*

Exercise 2.5.27. *A box contains two double-headed coins, one double-tailed coin and two fair coins. You shut your eyes and pick a coin from the box.*

(a) What is the probability that you picked a double-headed coin?

(b) You toss the coin. What is the probability that it shows heads?

(c) You open your eyes and see that the coin shows heads. What is the probability that it is a double-headed coin?

(d) You shut your eyes and toss the coin again. What is the probability that it shows heads?

(e) You open your eyes and see that the coin shows heads again. What is the probability that it is a double-headed coin?

(f) You throw the coin away, shut your eyes and pick a new coin from the box. What is the probability that it is a double-headed coin?

(g) You toss the coin. What is the probability that it shows heads?

2.6 Independence

In general, the occurence of some event B changes the probability that a certain event A occurs, the original $\mathbb{P}(A)$ being replaced by $\mathbb{P}(A|B)$. If the probability remains unchanged i.e., $\mathbb{P}(A|B) = \mathbb{P}(A)$, then we call A and B independent. Since in order to talk about $\mathbb{P}(A|B)$ we need $\mathbb{P}(B) > 0$, we give the following more general definition which agrees with this special case.

Definition 2.6.1. The events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. More generally, a family of events $\{A_i : i \in I\}$ is **independent** if $\mathbb{P}(\bigcap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i)$ for each finite subset J of I. A family $\{A_i : i \in I\}$ is **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for each $i \neq j$.

If the occurrence of two events is governed by distinct and non-interacting processes, such events will turn out to be independent: This will be our *modelling assumption*.

Independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ are never independent as $\mathbb{P}(A \cap B) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$.

Example 2.6.2. Roll two dice and let A be the event "the first die is 4". Let B_1 be the event "the second die is 2". This satisfies our intuitive notion of independence since the outcome of the first dice roll has nothing to do with that of the second. To check independence, note that $\mathbb{P}(B_1) = 1/6 = \mathbb{P}(A)$ and $\mathbb{P}(A \cap B_1) = 1/36$.

Let B_2 be the event "the sum of the two dice is 3". Since $A \cap B_2 = \varnothing$, we have that $\mathbb{P}(A \cap B_2) = 0 < \mathbb{P}(A)\mathbb{P}(B_2)$ and so the events cannot be independent.

Let B_3 be the event "the sum of the two dice is 7". This time, A and B_3 are independent. Indeed, we have that $\mathbb{P}(B_3) = 6/36$ and $\mathbb{P}(A \cap B_3) = 1/36$.

Let B_4 be the event "the sum of the two dice is 9". We have that A and B_4 are not independent. Indeed, $\mathbb{P}(A \cap B_4) = 1/36$ but $\mathbb{P}(A)\mathbb{P}(B_4) = 1/6 \cdot 4/36$.

Remark 2.6.3. Independence is stronger than pairwise independence: Any independent family is clearly pairwise independent but the converse is not true. Indeed, toss two coins and consider the events "first coin gives H", "second coin gives H", "resulting number of heads is odd". They form a family which is pairwise independent but not independent.

Exercise 2.6.4. *Let A and B be events satisfying $\mathbb{P}(A), \mathbb{P}(B) > 0$ and such that $\mathbb{P}(A|B) = \mathbb{P}(A)$. Show that $\mathbb{P}(B|A) = P(B)$.*

Exercise 2.6.5. *Two fair dice are thrown. Let A be the event that the first shows an odd number, B be the event that the second shows an even number, and C be the event that either both are odd or both are even. Show that A, B, C are pairwise independent but not independent.*

Example 2.6.6 (De Méré's paradox). Let A be the event of getting at least one six with one throw of 4 fair dice and let B be the event of getting at least one double six with 24 throws of 2 fair dice? Which is more probable?

Let us first compute $\mathbb{P}(A)$. Label the dice $1, 2, 3, 4$. For each $i \in \{1, 2, 3, 4\}$, let A_i be the event of getting a six on die i when throwing the 4 dice. Clearly, $\mathbb{P}(A_i) = 1/6$ and $\mathbb{P}(A_i^c) = 5/6$. It is reasonable to *assume* that $A_1^c, A_2^c, A_3^c, A_4^c$ is an independent family (we can think of throwing 4 dice as four distinct and non-interacting processes). Therefore,

$$\mathbb{P}(A^c) = \mathbb{P}(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c) = \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\mathbb{P}(A_3^c)\mathbb{P}(A_4^c) = \left(\frac{5}{6}\right)^4,$$

from which $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) \approx 0.5179$.

Let us now compute $\mathbb{P}(B)$. For $i \in \{1, 2, \dots, 24\}$, let B_i be the event of getting a double six on the i-th throw of 2 dice. Clearly, $\mathbb{P}(B_i) = 1/36$ and $\mathbb{P}(B_i^c) = 35/36$. As above, it is reasonable to assume independence of $B_1^c, B_2^c, \dots, B_{24}^c$. Therefore,

$$\mathbb{P}(B^c) = \mathbb{P}(B_1^c \cap \dots \cap B_{24}^c) = \mathbb{P}(B_1^c) \cdot \dots \cdot \mathbb{P}(B_{24}^c) = \left(\frac{35}{36}\right)^{24},$$

from which $\mathbb{P}(B) = 1 - \mathbb{P}(B^c) \approx 0.4905$.

Example 2.6.7 (Bernoulli trials). If an experiment involves a sequence of independent but identical stages, we say that we have a sequence of independent trials. If there are only two possible results at each stage, we say that we have a sequence of independent Bernoulli trials.

Consider an experiment that consists of n independent tosses of a biased coin, in which the probability of H is p, for some $p \in [0, 1]$. What is the probability of getting exactly k heads?

Let A_i be the event "the i-th toss is H". Independence means that the events $A_1, A_2, \dots, A_n$ are independent (the occurrence of any of them is governed by distinct and non-interacting processes). Consider for example the outcome in which we have k heads followed by $n - k$ tails i.e., the elementary event $A_1 \cap A_2 \cap \dots \cap A_k \cap A_{k+1}^c \cap \dots \cap A_n^c$. Intuitively, the family $A_1, A_2, \dots, A_k, A_{k+1}^c, \dots, A_n^c$ is independent[1], as we will shortly show, and so we have that

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_k \cap A_{k+1}^c \cap \dots \cap A_n^c) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdots \mathbb{P}(A_k)\mathbb{P}(A_{k+1}^c) \cdots \mathbb{P}(A_n^c) = p^k(1-p)^{n-k}.$$

Moreover, any other elementary event consisting of k heads and $n - k$ tails will have the same probability. Therefore, by additivity, it is enough to count such elementary events. This is equivalent to counting the number of subsets of size k (the trials giving head) of a set of size n (the set of all trials). This number is $\binom{n}{k}$ and so the probability of getting exactly k heads is

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

[1] Notice how this intuitively makes sense: if A and B are independent, the occurrence of B does not provide any new information on the probability of A and so the non-occurrence of B should also provide no information on the probability of A.

As mentioned above, we now show that if $A_1, A_2 \dots, A_n$ is an independent family then, replacing A_i by A_i^c for some i, still gives an independent family. By possibly repeating the argument, it is enough to show this for one value of i, say $i = n$. Therefore, we show the following: if $A_1, A_2 \dots, A_n$ is an independent family, then $A_1, A_2 \dots, A_{n-1}, A_n^c$ is an independent family. Consider a subset J of $A_1, A_2 \dots, A_{n-1}, A_n^c$. If $A_n^c \notin J$, then $\mathbb{P}(\bigcap_{A\in J} A) = \prod_{A\in J} \mathbb{P}(A)$ by assumption. If $A_n^c \in J$ then, by possibly relabelling indices we have that J is of the form $J = \{A_1, \dots, A_\ell, A_n^c\}$ for some $\ell \in \{1, \dots, n-1\}$. Letting $B = A_1 \cap \dots \cap A_\ell$, we have that

$$\begin{aligned}
\mathbb{P}(A_1 \cap \dots \cap A_\ell \cap A_n^c) &= \mathbb{P}(B \cap A_n^c) \\
&= \mathbb{P}(B \setminus (B \cap A_n)) \\
&= \mathbb{P}(B) - \mathbb{P}(B \cap A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(A_1 \cap \dots \cap A_\ell \cap A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(A_1) \cdots \mathbb{P}(A_\ell)\mathbb{P}(A_n) \\
&= \mathbb{P}(B) - \mathbb{P}(B)\mathbb{P}(A_n) \\
&= \mathbb{P}(B)(1 - \mathbb{P}(A_n)) \\
&= \mathbb{P}(B)\mathbb{P}(A_n^c) \\
&= \mathbb{P}(A_1) \cdots \mathbb{P}(A_\ell)\mathbb{P}(A_n^c),
\end{aligned}$$

which is what we wanted to show.

Exercise 2.6.8. *Suppose A and B are events and the probability of B is either zero or one. Show that A and B are independent.*

Exercise 2.6.9. *Let $A_1, A_2, \dots, A_m$ be a family of independent events such that $\mathbb{P}(A_i) = p$ for each $i \in \{1, \dots, m\}$. Find the probability that*

(a) none of the A_i's occur;

(b) an even number of the A_i's occur.

Example 2.6.10 (Gambler's ruin). A man wants to buy a car at a cost of N units of money. He starts with k units, for some $0 < k < N$ and tries to win the remainder by the following gamble with his bank manager. He tosses a fair coin repeatedly and independently. If H comes up, then the manager pays him one unit. If T comes up, then he pays the manager one unit. He plays the game until one of two events occurs: either he runs out of money and is bankrupted or he wins enough to buy the car. What is the probability that he is ultimately bankrupted?

We want to compute the probability of the event A_k "bankrupted if starting with k units". Notice that $\mathbb{P}(A_0) = 1$ and $\mathbb{P}(A_N) = 0$. Let B be the event "first toss is H". The law of total probability tells us that

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

But if the first toss is H, he has $k+1$ units and if the first toss is T, he has $k-1$ units. Since the tosses are independent, we have that $\mathbb{P}(A_k|B) = \mathbb{P}(A_{k+1})$ and $\mathbb{P}(A_k|B^c) = \mathbb{P}(A_{k-1})$. Therefore, letting $p_k = \mathbb{P}(A_k)$, we have that

$$p_k = \frac{1}{2}(p_{k+1} + p_{k-1}), \tag{2.3}$$

with $p_0 = 1$ and $p_N = 0$. We want to compute the value of p_k by using this recurrence relation together with the two "boundary conditions". Observe first that, by Equation (2.3), the difference between consecutive p_k's is always the same: $p_k - p_{k-1} = p_{k+1} - p_k$. Letting $b_k = p_k - p_{k-1}$ this common value, we have that $b_k = b_1$ and so

$$p_k = b_1 + p_{k-1} = b_1 + (b_1 + p_{k-2}) = \cdots = kb_1 + p_0.$$

Substituting N for k, we get $0 = p_N = Nb_1 + p_0 = Nb_1 + 1$, from which $b_1 = -1/N$ and so $p_k = 1 - k/N$.

Notice that, for each fixed k, the probability p_k he is bankrupted starting with k units tends to 1 as $N \to \infty$.

Exercise 2.6.11. *In this exercise we consider gambler's ruin in the case the coin has probability p of getting heads and probability q of getting tails, where $p + q = 1$ and $p \neq 1/2$. Using the previous notation, proceed as follows (each step is deduced from the previous):*

- *Show that* $p_k = p \cdot p_{k+1} + q \cdot p_{k-1}$;
- *Deduce that*
$$p_{k+1} - p_k = \frac{q}{p} \cdot (p_k - p_{k-1});$$
- *Deduce that*
$$b_k = \left(\frac{q}{p}\right)^{k-1} \cdot b_1;$$
- *Conclude that*
$$p_k = 1 - \frac{1 - \left(\frac{q}{p}\right)^k}{1 - \left(\frac{q}{p}\right)^N}.$$

Remark 2.6.12. Let's make a comment about the previous exercise in the realistic situation that our gambler plays against a gambling machine. Gambling machines in most countries permit by law a certain degree of "unfairness" by taking $p < 1/2$. This allows the house to make an income. Suppose that $p = 0.47$ and that the gambler starts with 10 units and aims at reaching 20 units. The probability he is bankrupted turns out to be roughly 77% (check it yourself!). Therefore, a "slightly unfair" game at each round can become devastatingly unfair in the long run.

Chapter 3

Random variables

Most of the times we are not interested in an experiment itself but rather in some consequence of its random outcome. A random variable can be thought of as a numerical "summary" of a certain aspect of the experiment. It is nothing but a function from the sample space Ω to the real line $\mathbb{R}$, where the "random" in the name comes from the experiment:

1. Chance determines the random outcome $\omega \in \Omega$;

2. The outcome ω determines a certain quantity of interest.

In other words, a random variable X represents an unknown quantity that varies with the outcome of a random event. Before the random event, we know which values X could possibly assume, but we do not know which one it will take until the random event happens. The terminology may appear confusing: a variable is a function? This is because the words "random variable" were in use long before the connection between probability and analysis was discovered.

Example 3.0.1. Consider the experiment of tossing a coin twice. We can take as sample space $\Omega = \{HH, HT, TH, TT\}$. For any outcome $\omega \in \Omega$, we let $X(\omega)$ be the number of heads in the outcome. Therefore, $X(HH) = 2$, $X(HT) = X(TH) = 1$ and $X(TT) = 0$.

Example 3.0.2. Consider the experiment of throwing a fair die once. We can take $\Omega = \{1, 2, 3, 4, 5, 6\}$. Suppose that we gamble on the outcome of the experiment in such a way that the profit is

$$\begin{array}{ll} -1 & \text{if the outcome is in } \{1,2,3\}; \\ 0 & \text{if the outcome is } 4; \\ 2 & \text{if the outcome is in } \{5,6\}; \end{array}$$

where negative profits are positive losses. If the outcome is ω, then our profit is $X(\omega)$, where $X\colon \Omega \to \mathbb{R}$ is defined by $X(1) = X(2) = X(3) = -1$, $X(4) = 0$, $X(5) = X(6) = 2$.

Crucially, the function $X\colon \Omega \to \mathbb{R}$ has to be sufficiently well-behaved so that we can talk about probabilities with which X assumes certain values:

Definition 3.0.3. A **random variable** is a function $X\colon \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. We denote by $\mathrm{Im}(X)$ the image of X i.e., the values taken by X.

Remark 3.0.4. Notice that whenever we talk about a random variable we implicitly assume an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

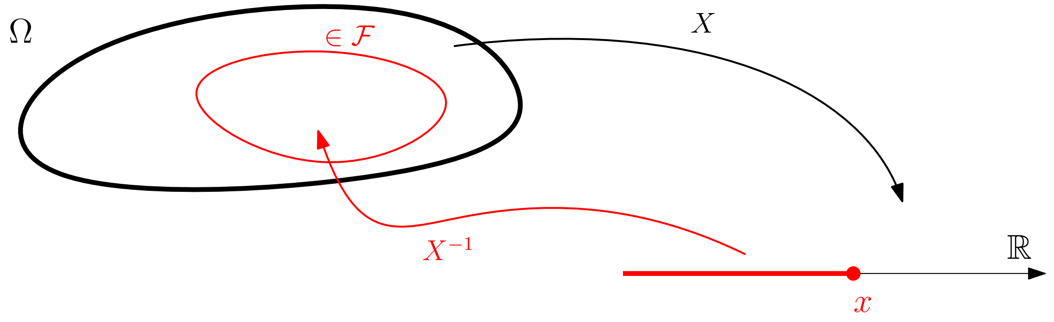


Figure 3.1: Visualization of a random variable X and the property $\{\omega \in \Omega : X(\omega) \le x\} \in \mathcal{F}$.

Example 3.0.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where $\mathcal{F}$ is the power set of Ω. Then obviously any function $X : \Omega \to \mathbb{R}$ is a random variable. Recall that if Ω is countable, we can always take its power set as our σ-field $\mathcal{F}$.

In general, the numerical value of a random variable is more likely to lie in certain subsets of $\mathbb{R}$, depending on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the function X itself. The following notion describes the distribution of the likelihoods of possible values of X. As mentioned above, it is the reason behind the technical requirement $\{\omega \in \Omega : X(\omega) \le x\} \in \mathcal{F}$ in the definition of random variable.

Definition 3.0.6. The **distribution function** of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by $F_X(x) = \mathbb{P}(X \le x)$. Here and in the following we use the shorthands $X \le x$ or $\{X \le x\}$ for the event $\{\omega \in \Omega : X(\omega) \le x\}$.

We are interested in two types of random variables:

Definition 3.0.7. The random variable X is **discrete** if it takes values in some countable subset of $\mathbb{R}$. The **probability mass function (pmf)** of a discrete random variable X is the function $f_X : \mathbb{R} \to [0, 1]$ given by $f_X(x) = \mathbb{P}(X = x)$.

The random variable X is **continuous** if its distribution function can be expressed as

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

for some integrable function $f_X : \mathbb{R} \to [0, \infty)$ called the **probability density function (pdf)** of X.

Remark 3.0.8. We will sometimes drop the subscript X in F_X or f_X when it is clear to which random variable we are referring.

Remark 3.0.9. The definition of random variable requires that $\{X \le x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. But what about $\{X = x\}$? It turns out that, since $\{x\}$ is a Borel set (Example 2.1.15), Theorem 3.0.33 will indeed imply that $\{X = x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$ and so it makes sense to write $\mathbb{P}(X = x)$.

The name continuous comes from the fact (a generalization of the Fundamental theorem of calculus) that the function F_X defined by $F_X(x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$ is continuous. This is in sharp contrast to discrete random variables, whose distribution functions are never continuous (only right-continuous as we will see shortly).

Remark 3.0.10. Observe that, knowing the probability mass function of a discrete random variable, we can immediately compute its distribution function using countable additivity. Indeed,

$$\{X \leq x\} = \bigcup_{k:\ k\leq x \text{ and } k\in \operatorname{Im}(X)} \{X = k\},$$

where the union is countable as $\operatorname{Im}(X)$ is countable, and so

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{k:\ k\leq x \text{ and } k\in \operatorname{Im}(X)} \mathbb{P}(X = k) = \sum_{k:\ k\leq x \text{ and } k\in \operatorname{Im}(X)} f_X(k).$$

Example 3.0.11. Let us compute the probability mass function and the distribution function of the discrete random variable in Example 3.0.1. In view of Remark 3.0.10, we start with the pmf. The values taken by X are $0, 1, 2$ and the pmf f_X is given by $f_X(0) = \mathbb{P}(TT) = 1/4$, $f_X(1) = \mathbb{P}(HT \cup TH) = 1/2$ and $f_X(2) = \mathbb{P}(HH) = 1/4$. For all $x \notin \{0, 1, 2\}$, we have $f_X(x) = 0$. Obtained the pmf, it is now easy to compute the distribution function:

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0; \\ 1/4 & \text{if } 0 \leq x < 1; \\ 3/4 & \text{if } 1 \leq x < 2; \\ 1 & \text{if } x \geq 2. \end{cases}$$

In order to compute the pmf of a discrete random variable X, do the following. For each possible value x of X:

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$;
2. Add their probabilities to obtain $f_X(x)$.

Example 3.0.12 (Uniform random variable). The random variable X is uniform on $[a, b]$ if it has distribution function

$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < a; \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b; \\ 1 & \text{if } x > b. \end{cases}$$

Such a random variable X is continuous, as it admits probability density function given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b; \\ 0 & \text{otherwise.} \end{cases}$$

The idea here is that we are picking a value "at random" from $[a, b]$ (values outside the interval are impossible, and all those inside have the same probability density). Therefore, the probability that $X \leq c$ for some $c \in [a, b]$ should intuitively be $\frac{c-a}{b-a}$.

If the distribution function F_X of a continuous random variable X is differentiable at some $x \in \mathbb{R}$, the value $f_X(x)$ of the probability density function at x can be found by taking the derivative of F_X at x.

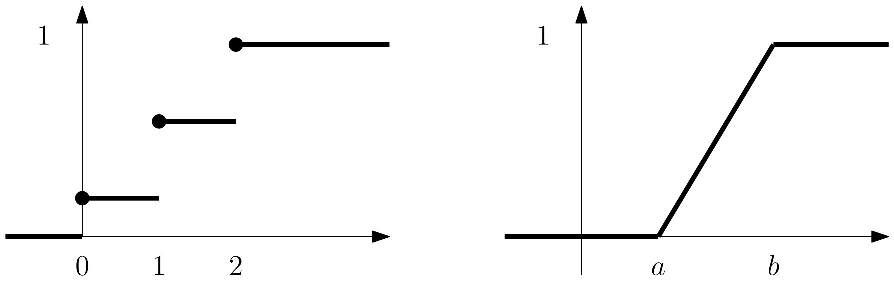


Figure 3.2: Distribution function of the discrete random variable in Example 3.0.1 (left) and of the uniform random variable on $[a, b]$ (right).

Before providing several examples of important discrete random variables, we make the following observation. If X is a discrete random variable with pmf f_X then, by definition, $\mathrm{Im}(X)$ is a countable subset of $\mathbb{R}$. Moreover, the following two properties hold:

$$f_X(x) \geq 0 \text{ for each } x \in \mathbb{R} \text{ and } f_X(x) = 0 \text{ if } x \notin \mathrm{Im}(X), \tag{3.1}$$

$$\sum_{x\in\mathrm{Im}(X)} f_X(x) = \mathbb{P}\Big(\bigcup_{x\in\mathrm{Im}(X)} \{\omega \in \Omega : X(\omega) = x\}\Big) = \mathbb{P}(\Omega) = 1. \tag{3.2}$$

Equation (3.2) is sometimes written as $\sum_{x\in\mathbb{R}} f_X(x) = 1$ in light of the fact that only countably many values of x make non-zero contributions to this sum. The two properties above essentially characterize mass functions of discrete random variables in the sense of the following theorem.

Theorem 3.0.13. *Let $S = \{s_i : i \in I\}$ be a countable set of distinct real numbers and let $\{\pi_i : i \in I\}$ be a collection of real numbers satisfying*

$$\pi_i \geq 0 \textit{ for each } i \in I, \textit{ and } \sum_{i\in I} \pi_i = 1. \tag{3.3}$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a discrete random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ such that the pmf of X is given by $f_X(s_i) = \pi_i$ for $i \in I$, and $f_X(s) = 0$ for $s \notin S$.

Proof. Since S is countable, we can build our probability space as follows. We let $\Omega = S$, $\mathcal{F} = 2^\Omega$ and, for each $A \in \mathcal{F}$, we define

$$\mathbb{P}(A) = \sum_{i:\ s_i\in A} \pi_i.$$

It is easy to see that such a $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Finally, we define our discrete random variable X as the function $X : \Omega \to \mathbb{R}$ given by $X(\omega) = \omega$ for each $\omega \in \Omega$. For $i \in I$, we have that $f_X(s_i) = \mathbb{P}(X = s_i) = \pi_i$, whereas for $s \notin S$, we have that $f_X(s) = \mathbb{P}(X = s) = \mathbb{P}(\varnothing) = 0$, as desired. □

Theorem 3.0.13 is very useful, since for many purposes it allows us to forget about probability spaces: it is enough to say "let X be a random variable taking the value s_i with probability π_i and satisfying (3.3)" and we can be sure that such a random variable exists without having to construct it explicitly. This reasoning can be applied in order to check that the examples below provide indeed (discrete) random variables.

Example 3.0.14 (Constant random variable). Let $c \in \mathbb{R}$ and let $X \colon \Omega \to \mathbb{R}$ be the function given by $X(\omega) = c$ for each $\omega \in \Omega$. This is a random variable with pmf

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 0 & \text{if } x \neq c; \\ 1 & \text{if } x = c. \end{cases}$$

and distribution function

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < c; \\ 1 & \text{if } x \geq c. \end{cases}$$

Example 3.0.15 (Bernoulli random variable). A coin is tossed once and let p be the probability of H. Let X be 1 if the toss gives H and 0 otherwise. This is a random variable with pmf

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} 1-p & \text{if } x = 0; \\ p & \text{if } x = 1; \\ 0 & \text{otherwise.} \end{cases}$$

and distribution function

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0; \\ 1-p & \text{if } 0 \leq x < 1; \\ 1 & \text{if } x \geq 1. \end{cases}$$

We refer to X as a Bernoulli random variable with parameter p, denoted $X \sim \mathit{Bernoulli}(p)$. In practice, the Bernoulli random variable is used to model probabilistic situations with just two outcomes, such as:

(a) The state of a telephone at a given time that can be either free or busy;

(b) A person who can be either healthy or sick with a certain disease;

(c) The preference of a person who can be either for or against a certain political candidate.

Moreover, we will see how more complicated random variables can be obtained by combining multiple Bernoulli random variables.

Example 3.0.16 (Binomial random variable). A coin is tossed n times. At each toss, the coin gives H with probability p, independently of prior tosses. Let X be the number of heads in the n-toss sequence. We refer to X as a binomial random variable with parameters (n, p), denoted $X \sim \mathit{Binomial}(n, p)$. We essentially already computed its pmf in Example 2.6.7. It is given by

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

for $k \in \{0, 1, \ldots, n\}$. Instead of tossings of a coin, we can more generally take n independent Bernoulli trials with probability p of success.

The Binomial theorem implies that f_X is a legitimate pmf. Indeed,

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

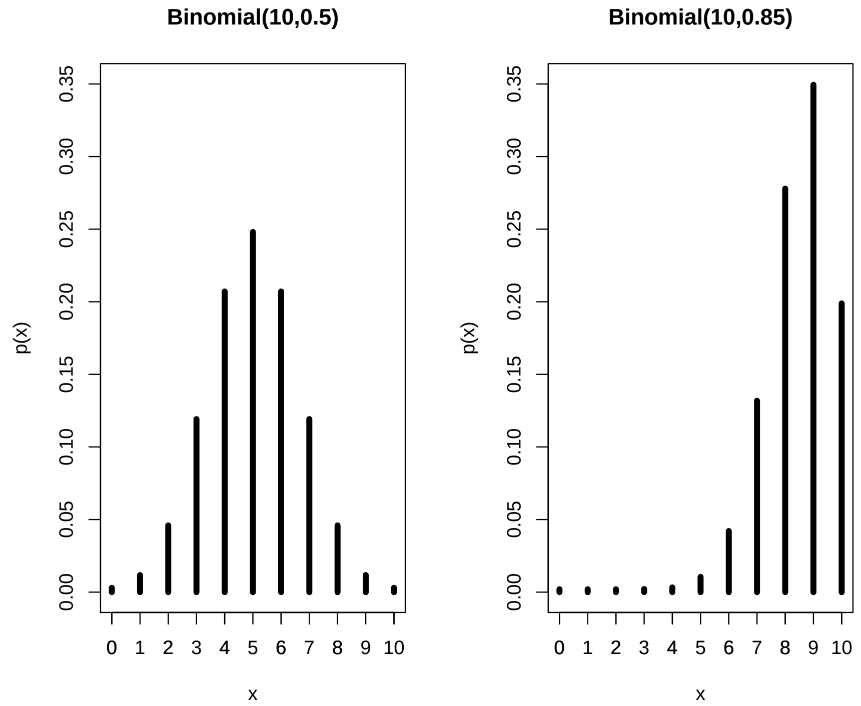


Figure 3.3: The pmf of a binomial random variable. If $p = 1/2$, the pmf is symmetric around $n/2$. Otherwise, it is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

Example 3.0.17 (Geometric random variable). Suppose that we repeatedly and independently toss a coin with probability p of getting H. The geometric random variable is the number X of tosses needed for a head to come up for the first time. Its pmf is given by

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

for $k = 1, 2, \ldots$. We refer to such an X as a geometric random variable with parameter p, denoted $X \sim \mathit{Geometric}(p)$. Notice that $f_X(k)$ gives indeed the pmf of a discrete random variable as $f_X(k) \geq 0$ for $k = 1, 2, \ldots$ and

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

where in the last equality we used the sum of a geometric series.

More generally, we can interpret the geometric random variable in terms of repeated independent trials until the first "success". Each trial has probability p of success and the number of trials until (and including) the first success is modeled by the geometric random variable.

Example 3.0.18 (Poisson random variable). A random variable X is said to be Poisson with parameter $\lambda > 0$, denoted $X \sim \mathit{Poisson}(\lambda)$, if it has pmf given by

$$f_X(k) = \mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

for $k = 0, 1, 2, \ldots$. Notice that this gives indeed the pmf of a discrete random variable as $f_X(k) \geq 0$ for $k = 0, 1, \ldots$ and

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

where in the last equality we used the definition of e^λ.

How does a Poisson random variable with parameter λ arise? It turns out that it is a limit of a binomial random variable with parameters $(n, \lambda/n)$. Indeed,

$$\begin{aligned}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} &= \frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k}\end{aligned}$$

and for any fixed k, taking the limit $n \to \infty$, we have that the quantity above tends to $e^{-\lambda}\frac{\lambda^k}{k!}$. In other words, a Poisson random variable with parameter λ approximates a binomial random variable with parameters (n, p) provided $\lambda = np$, n is large and p is small. It appears abundantly in life, for example, when counting the number of radio-active decays in a unit of time or the number of cars involved in accidents in a city on a given day.

Example 3.0.19 (Negative binomial random variable). Consider a sequence of independent Bernoulli trials with probability p of success. The negative binomial random variable is the number X of trials required to achieve a number r of successes. Therefore, the values taken by X are $r, r+1, r+2, \ldots$. Since $\{X = x\}$ if and only if we have $r-1$ successes in the first $x-1$ trials and a success at the x-th trial, then independence implies that the pmf is given by

$$\mathbb{P}(X = x) = \binom{x-1}{r-1}p^{r-1}(1-p)^{(x-1)-(r-1)} \cdot p = \binom{x-1}{r-1}p^r(1-p)^{x-r},$$

for $x = r, r+1, \ldots$.

Let X be a random variable and let $A \subseteq \mathbb{R}$. How can we compute probabilities of the form $\mathbb{P}(X \in A)$ (assuming for the moment $X \in A$ is indeed an event)? If X is described both in terms of an underlying experiment and via its pmf, then we can typically proceed in two different ways, as shown in the following example.

Example 3.0.20. Let $X \sim \textit{Geometric}(p)$. Compute $\mathbb{P}(X > k)$.

$$\begin{aligned}\mathbb{P}(X > k) = 1 - \mathbb{P}(X \le k) = 1 - \sum_{i=1}^{k}\mathbb{P}(X = i) &= 1 - \sum_{i=1}^{k}(1-p)^{i-1}p \\ &= 1 - p\sum_{i=1}^{k}(1-p)^{i-1} \\ &= 1 - p \cdot \frac{1-(1-p)^k}{1-(1-p)} \\ &= (1-p)^k,\end{aligned}$$

where in the second equality we used finite additivity and in the fifth the formula for a geometric progression.

Alternatively, we can use the experimental description of X. The event $\{X > k\}$ coincides with the event that the first k tosses are all tails and so, by independence, $\mathbb{P}(X > k) = (1-p)^k$.

Exercise 3.0.21. *Determine whether the following functions $f\colon \mathbb{N} \to [0,1]$ are probability mass functions of a discrete random variable:*

- $f(x) = \frac{1}{x(x+1)}$;
- $f(x) = \frac{4}{x(x+1)(x+2)}$.

Exercise 3.0.22. *An airplane engine breaks down during a flight with probability $1-p$. An airplane lands safely only if at least half of its engines are functioning upon landing. What is preferable: a two-engine airplane or a four-engine airplane?*

Exercise 3.0.23. *There are n white balls and m black balls in an urn. Each time, we take out one ball (with replacement) until we have a black ball. What is the probability that we need at least k trials?*

One might worry that the condition in the definition of a random variable is too stringent. Luckily, this is not the case: almost all reasonable functions turn out to be random variables. We now provide several ways of construting new random variables.

Since real numbers have addition and multiplication, we can perform such operations on real-valued functions pointwise. If $f_1, f_2 \colon \Omega \to \mathbb{R}$ are two functions, then the pointwise sum $f_1 + f_2 \colon \Omega \to \mathbb{R}$ is defined by $(f_1 + f_2)(\omega) = f_1(\omega) + f_2(\omega)$ for each $\omega \in \Omega$. We define the pointwise product $f_1 f_2$ and the pointwise scalar λf_1 similarly.

Proposition 3.0.24. *Let X and Y be random variables and let $\lambda \in \mathbb{R}$. The following are random variables:*

(a) λX;

(b) $X + Y$;

(c) XY;

(d) $Z(\omega) = \begin{cases} Y(\omega)/X(\omega) & \text{if } X(\omega) \neq 0; \\ 0 & \text{if } X(\omega) = 0. \end{cases}$

(e) $\max\{X, Y\}$;

(f) $\min\{X, Y\}$.

Proof. We prove only (b) as the other proofs are similar. We need to show that $\{\omega : X(\omega) + Y(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Since σ-fields are closed under complementation, it is then enough to show that $\{\omega : X(\omega) + Y(\omega) > x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Observe that, since there exists a rational number between any two real numbers, we have

$$\{\omega : X(\omega) + Y(\omega) > x\} = \bigcup_{r \in \mathbb{Q}} \{\omega : X(\omega) > r, Y(\omega) > x - r\}.$$

But for fixed r and x, $\{\omega : X(\omega) > r\} \in \mathcal{F}$ and $\{\omega : Y(\omega) > x - r\} \in \mathcal{F}$, as X and Y are random variables. Therefore their intersection $\{\omega : X(\omega) > r, Y(\omega) > x - r\}$ belongs to $\mathcal{F}$ and hence the countable union $\bigcup_{r \in \mathbb{Q}} \{\omega : X(\omega) > r, Y(\omega) > x - r\}$ belongs to $\mathcal{F}$ as well. □

Example 3.0.25. A binomial random variable with parameters (n, p) is a sum of n Bernoulli random variables each with parameter p.

Consider a probability model of today's weather and let the random variable X be the temperature in degrees Celsius. The transformation $Y = 1.8X + 32$ gives the temperature in degrees Fahrenheit. In this case, Y is a linear function of X but we may also consider nonlinear functions. For example, if we wish to display temperatures on a logarithmic scale, we would want to use the function $Y = \log(X)$.

More generally, given a random variable $X\colon \Omega \to \mathbb{R}$, and a function $g\colon \mathbb{R} \to \mathbb{R}$, we can consider the composition function $Y = g(X)$ of g after X i.e., the function $Y\colon \Omega \to \mathbb{R}$ mapping $\omega \in \Omega$ to $g(X(\omega)) \in \mathbb{R}$. It turns out that if g is continuous, we obtain another random variable:

Theorem 3.0.26. *Let X be a random variable and $g\colon \mathbb{R} \to \mathbb{R}$ a continuous function. Then $Y = g(X)$ is a random variable.*

Example 3.0.27. Let X be a random variable. Then $\sin(X), e^X, \log(X), X^n$ are all random variables.

Example 3.0.28. Consider the experiment of repeatedly and independently tossing a coin with probability p of getting H. We know that the number of tosses until the first H appears is a geometric random variable with pmf given by

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

for $k = 1, 2, \ldots$. Let now Y be the number of T before the first H. Then $Y = X - 1$ is another random variable whose pmf can be easily computed from that of X as follows:

$$f_Y(k) = \mathbb{P}(Y = k) = \mathbb{P}(X = k+1) = (1-p)^k p,$$

for $k = 0, 1, 2, \ldots$.

Example 3.0.29. Consider again the discrete random variable X in Example 3.0.1. It takes values $0, 1, 2$ and its pmf is given by $f_X(0) = 1/4$, $f_X(1) = 1/2$ and $f_X(2) = 1/4$. Then $Y = g(X) = (X-1)^2$ is another discrete random variable taking values in $0, 1$. The pmf of Y is given by

$$f_Y(0) = \mathbb{P}(Y = 0) = \mathbb{P}(X = 1) = 1/2,$$
$$f_Y(1) = \mathbb{P}(Y = 1) = \mathbb{P}(\{X = 0\} \cup \{X = 2\}) = \mathbb{P}(X = 0) + \mathbb{P}(X = 2) = 1/4 + 1/4 = 1/2,$$

where we have used finite additivity.

We now observe some properties that the distribution function of a generic random variable satisfies. Hence the following result holds for both discrete and continuous random variables.

Lemma 3.0.30. *The distribution function F of a random variable X satisfies the following properties:*

(a) It is monotonically increasing i.e., if $x \le y$, then $F(x) \le F(y)$.

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

(c) It is right-continuous i.e., $F(x+h) \to F(x)$ as h tends to 0 from the positive side.

Proof. We show only (a). If $x \le y$, then $\{\omega : X(\omega) \le x\} \subseteq \{\omega : X(\omega) \le y)\}$ and so

$$F(x) = \mathbb{P}(\{\omega : X(\omega) \le x\}) \le \mathbb{P}(\{\omega : X(\omega) \le y\}) = F(y)$$

by monotonicity of probability. Notice that this also implies that the limits in (b) exist, as $0 \le F(x) \le 1$ is bounded. □

A remarkable and reassuring fact is that the three properties in Lemma 3.0.30 in fact characterize distribution functions of random variables, as shown by the following result which can be seen as a generalization of Theorem 3.0.13 to generic random variables.

Theorem 3.0.31. *Let $F\colon \mathbb{R} \to \mathbb{R}$ be a function satisfying (a), (b) and (c) in Lemma 3.0.30. Then there exists a random variable X with distribution function F.*

The result above tells us that instead of directly providing a random variable, we can simply provide a function $F\colon \mathbb{R} \to \mathbb{R}$ satisfying (a), (b) and (c) in Lemma 3.0.30. Notice that it justifies the existence of the uniform random variable on $[a, b]$ (which was defined by providing its distribution function).

Exercise 3.0.32. *Determine whether the following functions $F\colon \mathbb{R} \to \mathbb{R}$ are distribution functions of a random variable:*

- $F(x) = \frac{x^2}{1+x^2}$;
- $F(x) = \frac{1}{\pi}(\arctan(x) + \frac{\pi}{2})$.

It turns out that, for a random variable X, not only it makes sense to compute $\mathbb{P}(X \leq x)$, which is the same as $\mathbb{P}(X \in (-\infty, x])$, but also $\mathbb{P}(X \in A)$ for all Borel sets $A \subseteq \mathbb{R}$. We first need to check that indeed $\{X \in A\}$ is an event whenever A is a Borel set. Recall that, in particular, every open set in $\mathbb{R}$ is a Borel set and the family $\mathcal{B}$ of Borel sets is extremely rich.

Theorem 3.0.33. *Let X be a random variable and let $A \in \mathcal{B}$ be a Borel set in $\mathbb{R}$. Then $\{X \in A\} \in \mathcal{F}$ i.e., $\{X \in A\}$ is an event.*

Proof. We proceed as follows. We let $\mathcal{G}$ be the family of all $A \subseteq \mathbb{R}$ such that $\{X \in A\} \in \mathcal{F}$ and show that $\mathcal{G}$ is a σ-field on $\mathbb{R}$ containing all open intervals in $\mathbb{R}$. Since we know $\mathcal{B}$ is generated by the open intervals (Proposition 2.1.13) and hence is the smallest σ-field on $\mathbb{R}$ containing the open intervals, we obtain that $\mathcal{B} \subseteq \mathcal{G}$, as desired.

Since $\{a < X < b\} \in \mathcal{F}$ for each real numbers $a < b$ (why?), $\mathcal{G}$ contains all open intervals. It remains to check that $\mathcal{G}$ is indeed a σ-field on $\mathbb{R}$. Clearly, $\mathbb{R} \in \mathcal{G}$, as $\{\omega : X(\omega) \in \mathbb{R}\} = \Omega \in \mathcal{F}$. Let now $A \in \mathcal{G}$. Then $\{X \in A\} \in \mathcal{F}$ and so $\{X \in A^c\} = \{X \in A\}^c \in \mathcal{F}$ as $\mathcal{F}$ is closed under complementation. Suppose finally that $A_1, A_2, \ldots \in \mathcal{G}$. Then $\{X \in \bigcup_n A_n\} = \bigcup_n \{X \in A_n\} \in \mathcal{F}$ as $\mathcal{F}$ is closed under countable unions. □

Corollary 3.0.34. *Let X be a discrete random variable and let $A \subseteq \mathbb{R}$ be any set. Then $\{X \in A\} \in \mathcal{F}$.*

Proof. Since X is discrete, $\mathrm{Im}(X)$ is a countable set. We then write $\{X \in A\}$ as a countable union of members of $\mathcal{F}$ as follows:

$$\{\omega : X(\omega) \in A\} = \bigcup_{x \in A \cap \mathrm{Im}(X)} \{\omega : X(\omega) = x\}.$$

By Theorem 3.0.33, since X is a random variable and $\{x\}$ is a Borel set, each set in the countable union belongs to $\mathcal{F}$. But since $\mathcal{F}$ is a σ-field, the countable union must belong to $\mathcal{F}$ as well, as claimed. □

We have seen that if X is a random variable and $g\colon \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X)$ is a random variable. It turns out that if X is a *discrete* random variable, we can in fact compose it with *any* function g to obtain a new (discrete) random variable, as the following result shows.

Corollary 3.0.35. *Let X be a discrete random variable and let $g\colon \mathbb{R} \to \mathbb{R}$ be an arbitrary function. Then $g(X)$ is a discrete random variable.*

Proof. Clearly, $g(X)$ takes countably many values, as X does. We then need to show that $g(X)$ is indeed a random variable, namely that for each $x \in \mathbb{R}$, $\{\omega \in \Omega : g(X(\omega)) \leq x\} \in \mathcal{F}$. In order to do that, we simply write our set as a countable union of members of $\mathcal{F}$:

$$\{\omega : g(X(\omega)) \leq x\} = \bigcup_{c \in \mathrm{Im}(X)\colon\, g(c) \leq x} \{\omega : X(\omega) = c\}.$$

By Theorem 3.0.33, since X is a random variable and $\{c\}$ is a Borel set, each set in the countable union belongs to $\mathcal{F}$. But since $\mathcal{F}$ is a σ-field, the countable union must belong to $\mathcal{F}$ as well. □

Knowing the distribution function of a random variable X, it is easy to compute the probabilities of the events $\{X > x\}$ and $\{x \leq X \leq y\}$:

Lemma 3.0.36. *Let F be the distribution function of the random variable X. Then*

(a) $\mathbb{P}(X > x) = 1 - F(x)$*;*

(b) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$*.*

Proof. (a) Since $\{X > x\} = \{X \leq x\}^c$, we have that

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

(b) Ω can be written as the disjoint union $\{X \leq x\} \cup \{x < X \leq y\} \cup \{X > y\}$. Therefore, by finite additivity and (a),

$$1 = F(x) + \mathbb{P}(x < X \leq y) + (1 - F(y)),$$

as claimed. □

We somehow convinced ourselves that in the case of a discrete random variable X, the probability mass function is more informative than the distribution function (see Remark 3.0.10). As the following result shows, the probability mass function indeed captures all the information in the probability space that is relevant to X: we can compute the probability of every event defined just in terms of X by simply knowing the pmf of X.

Lemma 3.0.37. *Let X be a discrete random variable with pmf $f(x)$ and let $A \subseteq \mathbb{R}$ be any set. Then*

(a) The set $\{x \in \mathbb{R} : f(x) \neq 0\}$ is countable.

(b) $\mathbb{P}(X \in A) = \sum_{x \in A} f(x)$ *where, in view of (a), the sum is understood to be a countable sum (it contains only countably many non-zero terms).*

Proof. (a) It follows from the fact that X takes countably many values.

(b) We have seen in the proof of Corollary 3.0.34 that we can write the event $\{X \in A\}$ as the countable union

$$\{X \in A\} = \bigcup_{x \in A \cap \mathrm{Im}(X)} \{X = x\}.$$

Countable additivity then implies that

$$\mathbb{P}(X \in A) = \sum_{x \in A \cap \mathrm{Im}(X)} \mathbb{P}(X = x) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} f(x),$$

as claimed. □

The analogue of Lemma 3.0.37 and its consequences in the case of continuous random variables are given by the following.

Lemma 3.0.38. *If X is a continuous random variable with pdf $f(x)$, then*

(1) $\mathbb{P}(x < X \leq y) = \int_x^y f(u)$.

(2) $\mathbb{P}(X = x) = 0$*, for each* $x \in \mathbb{R}$.

(3) $\mathbb{P}(x < X \leq y) = \mathbb{P}(x \leq X \leq y) = \mathbb{P}(x < X < y) = \mathbb{P}(x \leq X < y)$.

(4) $\int_{-\infty}^{\infty} f(u) = 1$.

Loosely speaking, the reason behind (2) is that there are uncountably many possible values for X and this number is so large that the probability of X taking any particular value is 0.

Proof. (1) By Lemma 3.0.36, definition of density function and additivity of integration, we have

$$\mathbb{P}(x < X \leq y) = F(y) - F(x) = \int_{-\infty}^{y} f(u) - \int_{-\infty}^{x} f(u) = \int_{x}^{y} f(u).$$

(2) For each $n \in \mathbb{N}$, we have that $\{X = x\} \subseteq \{x - 1/n < X \leq x\}$. Therefore, monotonicity and (1) imply that

$$\mathbb{P}(X = x) \leq \mathbb{P}\Big(x - \frac{1}{n} < X \leq x\Big) = \int_{x-\frac{1}{n}}^{x} f(u).$$

But $\int_{x-\frac{1}{n}}^{x} f(u)$ tends to 0 as $n \to \infty$.

(3) By finite additivity and (2),

$$\mathbb{P}(x \leq X \leq y) = \mathbb{P}(x < X \leq y) + \mathbb{P}(X = x) = \mathbb{P}(x < X \leq y).$$

The other equalities are proved similarly.

(4) By Lemma 3.0.30, $\lim_{x\to\infty} F(x) = 1$ and so $\int_{-\infty}^{\infty} f(u) = \lim_{x\to\infty} F(x) = 1$. □

Exercise 3.0.39. *Let X be a random variable with distribution function*

$$F(x) = \begin{cases} 0 & \text{if } x < 0; \\ x^2 & \text{if } 0 \leq x \leq 1; \\ 1 & \text{if } x > 1. \end{cases}$$

Is X discrete or continuous? Compute $\mathbb{P}(1/4 < X < 5)$, $\mathbb{P}(0.2 < X < 0.8)$ *and* $\mathbb{P}(X = 1/2)$.

Chapter 4

Discrete random variables

4.1 Expectation of discrete random variables

Suppose we have an experiment and a discrete random variable X arising from the experiment. We repeat the experiment a large number N of times and record the N values taken by X. Intuitively, we would expect that $\{X = x\}$ occurs approximately $\mathbb{P}(X = x)N$ many times. So the average of the values taken by X would approximately be

$$\frac{\sum_x x\mathbb{P}(X = x)N}{N} = \frac{\sum_x x f_X(x)N}{N} = \sum_x x f_X(x).$$

Definition 4.1.1. Let X be a discrete random variable with pmf $f_X(x)$. The **expected value** (or **expectation**, or **mean**) of X, denoted by $\mathbb{E}(X)$, is

$$\mathbb{E}(X) = \sum_x x f_X(x),$$

provided that $\sum_x |x f_X(x)|$ converges. We use again the convention that $\sum_x x f_X(x)$ denotes the sum over the values of x for which $f_X(x) \neq 0$. Hence we are dealing with either a finite sum or the sum of a series, as X takes countably many values (see Lemma 3.0.37(a)).

Remark 4.1.2. Since the expectation is defined only if $\sum_x |x f_X(x)|$ converges, the expectation is a real number. Indeed, it can be shown that if $\sum_x |x f_X(x)|$ converges, then $\sum_x x f_X(x)$ converges as well. However, asking only for convergence of $\sum_x x f_X(x)$ would not be enough for our purposes. Indeed, the following undesirable behavior might occur: Given a series and $x \in \mathbb{R}$, there might exist a rearrangement[1] of the series which converges to x. A famous example is the so-called alternating harmonic series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$: it converges to $\ln 2$ but we can rearrange its terms to make it convergent to any $x \in \mathbb{R}$! Luckily, if the requirement in Definition 4.1.1 holds, then all rearrangements converge to the same real number and we do not have to worry about the order of summation.

We will soon give a precise mathematical meaning to the notion of expectation (thanks to the Weak law of large numbers). For the time being, the intuitive idea of $\mathbb{E}(X)$ as the average value of X in a long run of independent trials is enough.

[1] A rearrangement of a certain series is a series obtained by changing the order of summation of its terms.

Example 4.1.3. Let X be the score obtained when throwing a fair die. Then

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

Simulating 10, 100 and 1000 throws, we obtain the following averages of the scores:

3	1	4	6	6	6	1	3	6	1

10 Throws: average = 3.7

5	6	4	4	4	2	3	3	5	6
5	1	5	3	4	1	6	5	5	2
6	4	5	2	1	6	6	3	4	2
4	3	2	2	2	2	6	4	4	3
3	5	2	2	5	6	3	1	1	5
1	5	4	4	3	5	3	3	2	4
6	4	2	3	3	4	2	5	4	6
5	4	1	2	1	2	3	1	4	4
5	6	3	4	4	5	3	4	4	4
6	1	4	6	5	2	1	3	3	1

100 Throws: average = 3.54

3	3	6	5	6	1	6	1	2	5	3	1	2	3	3	2	3	2	5	2
6	4	3	6	3	3	2	4	2	3	6	6	3	6	2	4	4	5	3	1
5	1	2	5	6	4	1	3	6	5	2	6	5	4	6	4	3	5	6	2
5	5	6	1	5	1	1	2	5	1	6	6	5	1	4	2	5	3	4	4
6	3	5	6	5	4	3	5	3	1	5	4	3	2	5	3	6	4	3	3
6	5	4	3	6	6	3	5	4	4	4	3	1	4	5	3	5	4	1	6
4	6	5	4	1	2	2	1	3	5	1	3	5	6	6	2	3	2	4	4
4	3	1	6	2	1	1	4	5	6	6	4	3	3	4	5	4	4	1	4
5	1	4	3	1	2	3	1	1	3	2	1	5	4	3	2	4	6	2	4
5	4	4	3	4	2	3	4	4	5	3	6	5	2	2	6	2	5	6	6
6	3	1	6	4	1	6	5	6	6	6	2	2	5	5	1	3	2	2	4
3	2	4	2	6	2	2	3	2	6	5	2	1	4	6	1	1	1	3	4
1	6	4	1	4	1	5	6	4	5	4	6	4	2	2	6	6	2	3	1
3	3	3	6	4	6	5	1	3	5	5	5	3	3	5	6	4	4	5	4
4	4	1	3	4	3	3	2	5	5	4	2	5	5	3	5	4	3	1	2
6	6	4	5	1	2	5	6	2	4	3	5	1	6	6	2	1	3	2	6
5	2	2	3	1	3	2	1	6	5	4	5	6	4	3	4	3	2	6	1
5	1	5	5	5	1	2	6	3	5	2	3	6	1	4	3	1	4	5	6
1	6	2	1	4	4	1	2	4	4	5	4	1	2	6	1	2	2	5	3
2	2	3	6	1	5	3	2	5	4	1	1	2	4	2	5	3	3	4	2
2	6	3	2	2	6	3	5	2	5	3	6	1	5	2	3	1	2	1	5
5	2	3	5	3	1	6	2	5	6	5	3	1	5	6	3	4	1	2	4
3	4	2	6	6	2	5	1	4	1	4	1	1	1	4	1	1	1	3	5
1	2	1	5	2	6	6	6	2	1	6	3	5	3	5	2	6	2	1	2
5	4	5	6	6	4	4	4	3	1	3	1	4	2	5	5	4	6	1	5
2	1	3	5	5	3	4	2	3	2	4	3	4	4	4	3	5	6	4	4
4	1	5	4	1	5	5	3	1	6	3	6	1	1	6	6	4	1	5	1
4	2	6	3	4	2	1	4	1	3	1	2	2	4	4	3	6	4	1	4
4	3	3	5	5	6	4	4	5	2	4	2	3	2	6	6	2	5	6	6
4	2	5	6	4	4	1	1	4	2	2	4	2	4	4	4	5	5	2	1
3	3	4	1	5	5	2	1	3	3	4	4	4	4	6	6	1	6	4	2
5	2	5	3	1	3	5	1	3	6	5	5	4	1	4	4	4	6	1	6
1	2	2	4	6	2	4	3	1	6	2	5	6	2	4	1	1	1	2	5
6	4	4	1	6	3	5	3	3	6	3	5	3	2	3	3	6	1	6	6
2	6	1	4	6	3	5	4	3	4	4	4	6	6	1	5	4	1	4	4
3	5	2	1	3	2	3	3	4	6	2	3	6	6	6	5	2	1	2	2
2	3	2	5	3	4	2	5	3	6	5	3	3	4	6	3	3	1	2	1
2	4	2	1	4	5	3	4	4	1	2	5	4	5	3	1	3	6	5	4
5	2	6	3	4	1	2	5	4	3	1	2	6	3	6	3	6	4	5	4
5	1	1	2	6	6	3	2	6	2	3	1	4	1	5	5	3	1	1	4
1	5	2	6	2	1	4	4	4	1	2	1	3	4	1	6	2	4	5	4
6	5	5	6	6	1	4	1	3	2	4	4	1	6	4	1	1	5	6	4
2	3	5	1	3	4	6	3	6	2	3	6	3	4	4	1	1	2	1	4
3	2	3	4	6	6	2	6	6	1	3	6	2	2	6	3	4	6	3	6
2	2	3	5	3	4	1	4	1	2	4	6	5	2	1	1	6	2	2	6
1	1	3	4	4	3	2	6	6	3	5	4	4	6	2	2	2	6	1	5
1	4	3	2	1	1	1	5	3	3	2	4	1	6	1	4	6	6	1	1
6	2	2	6	1	2	4	3	6	2	6	2	3	1	2	6	1	5	2	1
5	1	1	5	1	4	3	2	6	1	3	1	6	3	5	2	3	4	4	5
4	3	1	1	4	1	5	3	3	4	5	1	3	4	1	6	5	1	1	1

1,000 Throws: average = 3.47

It appears that the larger the number of throws, the closer the average is to the actual value of the expectation of X.

Example 4.1.4. Let $X \sim Bernoulli(p)$. Then

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

Example 4.1.5. Let $X \sim Binomial(n, p)$. Then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \\ &= np,\end{aligned}$$

where in the fourth equality we used the change of variable $j = k - 1$ and in the last the Binomial theorem.

Example 4.1.6. Let $X \sim Poisson(\lambda)$. Then

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

where in the third equality we used the change of variable $j = k - 1$ and in the last the definition of e^{λ}.

Example 4.1.7. Let $X \sim Geometric(p)$. Then

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=1}^{\infty} k(1-p)^{k-1} p &= p + \sum_{k=2}^{\infty} k(1-p)^{k-1} p \\ &= p + \sum_{s=1}^{\infty} (s+1)(1-p)^s p \\ &= p + \sum_{s=1}^{\infty} s(1-p)^s p + \sum_{s=1}^{\infty} (1-p)^s p \\ &= p + (1-p) \sum_{s=1}^{\infty} s(1-p)^{s-1} p + p(1-p) \sum_{s=1}^{\infty} (1-p)^{s-1} \\ &= p + (1-p)\mathbb{E}(X) + p(1-p) \cdot \frac{1}{1-(1-p)},\end{aligned}$$

where in the third equality we used the change of variable $s = k-1$ and in the last equality we used the formula for the sum of a geometric series. Solving for $\mathbb{E}(X)$, we obtain $\mathbb{E}(X) = 1/p$.

Exercise 4.1.8. *Let X and Y be discrete random variables with pmf's*

$$f_X(x) = \frac{4}{x(x+1)(x+2)} \quad \textit{and} \quad f_Y(x) = \frac{1}{x(x+1)},$$

respectively, where $x = 1, 2, \ldots$. Check whether X and Y admit an expectation and, if so, compute the value.

Given a discrete random variable X, how do we compute the expectation of the discrete random variable $Y = g(X)$? Well, according to the definition, we first have to compute the pmf of the newly defined $Y = g(X)$. We have seen a first example of this procedure in Example 3.0.29. We now work out the general case:

Lemma 4.1.9. *Let X be a discrete random variable and let $g\colon \mathbb{R} \to \mathbb{R}$. The pmf of $Y = g(X)$ is*

$$f_Y(y) = \sum_{x\colon\, g(x)=y} f_X(x).$$

Proof. We have that the composition function $Y = g \circ X$ acts as follows: $\omega \in \Omega \mapsto X(\omega) \in \mathbb{R} \mapsto g(X(\omega)) \in \mathbb{R}$. We rewrite the event $\{\omega \in \Omega : Y(\omega) = y\}$, in whose probability $f_Y(y)$ we are interested in, as a countable union of pairwise disjoint events:

$$\{\omega \in \Omega : Y(\omega) = y\} = \{\omega \in \Omega : g(X(\omega)) = y\} = \bigcup_{x\colon\, g(x)=y} \{\omega \in \Omega : X(\omega) = x\}.$$

By countable additivity,

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x\colon\, g(x)=y} \mathbb{P}(X = x) = \sum_{x\colon\, g(x)=y} f_X(x),$$

as claimed. □

Since doing the above procedure for each specific Y and then applying the definition of expectation becomes pretty tedious, the following important result settle once and for all the computation we need.

Theorem 4.1.10 (Law of the unconscious statistician, LOTUS). *Let X be a discrete random variable and let $g\colon \mathbb{R} \to \mathbb{R}$. Then*

$$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x).$$

Proof. Let $Y = g(X)$. We have seen in Lemma 4.1.9 that the pmf of Y is

$$f_Y(y) = \sum_{x\colon\, g(x)=y} f_X(x).$$

Therefore, by definition of expectation,

$$\begin{aligned}\mathbb{E}(Y) &= \sum_y y f_Y(y) \\ &= \sum_y y \sum_{x\colon g(x)=y} f_X(x) \\ &= \sum_y \sum_{x\colon g(x)=y} y f_X(x) \\ &= \sum_y \sum_{x\colon g(x)=y} g(x) f_X(x) \\ &= \sum_x g(x) f_X(x).\end{aligned}$$

□

Example 4.1.11. Let X be the score obtained when throwing a fair die and let $Y = (X-3)^2$. Compute $\mathbb{E}(Y)$.

We can proceed in two different ways. Either we first compute the pmf of Y as in Lemma 4.1.9 and use the definition of expectation, or we simply use LOTUS. It will become clear that the latter procedure should be preferred. Let us first find the pmf of Y. The values taken by are $0, 1, 4, 9$ and $f_Y(0) = f_X(3) = 1/6$, $f_Y(1) = f_X(2) + f_X(4) = 2/6$, $f_Y(4) = f_X(1) + f_X(5) = 2/6$, $f_Y(9) = f_X(6) = 1/6$. Using the definition of expectation, we then obtain

$$\mathbb{E}(Y) = \sum_y y f_Y(y) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} + 4 \cdot \frac{2}{6} + 9 \cdot \frac{1}{6} = \frac{19}{6}.$$

We now use LOTUS. Since the values taken by X are $1, 2, 3, 4, 5, 6$, each with probability $1/6$, we obtain:

$$\mathbb{E}(Y) = \sum_x (x-3)^2 f_X(x) = 4 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} = \frac{19}{6}.$$

Example 4.1.12. Let $X \sim Poisson(\lambda)$. Compute the expectation of $Y = e^X$.

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = \sum_{k=0}^{\infty} e^k \cdot f_X(k) = \sum_{k=0}^{\infty} e^k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e)^k}{k!} = e^{-\lambda} \cdot e^{\lambda e},$$

where in the second equality we used LOTUS.

Expected values often provide a convenient vehicle for choosing optimally between several candidate decisions that result in different expected rewards. If we view the expected reward of a decision as its “average payoff over a large number of trials,” it is reasonable to choose a decision with maximum expected reward.

Example 4.1.13 (The quiz problem). Consider a quiz game where a person is given two questions and must decide which question to answer first. Question 1 will be answered correctly with probability 0.8, and the person will then receive as prize \$100, while question 2 will be answered correctly with probability 0.5, and the person will then receive as prize \$200. If the first question attempted is answered incorrectly, the quiz terminates i.e., the person is not allowed to attempt the second question. If the first question is answered correctly, the person is allowed to attempt the second question. Which question should be answered first to maximize the expected value of the total prize money received?

The answer is not obvious because there is a tradeoff: attempting first the more valuable but also more difficult question 2 carries the risk of never getting a chance to attempt the easier question 1. Let us view the total prize money received as a random variable X and compute $\mathbb{E}(X)$ under the two possible question orders.

(a) *Answer question 1 first:* Then the pmf of X is $f_X(0) = 0.2$, $f_X(100) = 0.8 \cdot 0.5$, $f_X(300) = 0.8 \cdot 0.5$ and so

$$\mathbb{E}(X) = 0 \cdot 0.2 + 100 \cdot 0.8 \cdot 0.5 + 300 \cdot 0.8 \cdot 0.5 = 160.$$

(b) *Answer question 2 first:* Then the pmf of X is $f_X(0) = 0.5$, $f_X(200) = 0.5 \cdot 0.2$, $f_X(300) = 0.5 \cdot 0.8$ and so

$$\mathbb{E}(X) = 0 \cdot 0.5 + 200 \cdot 0.5 \cdot 0.2 + 300 \cdot 0.5 \cdot 0.8 = 140.$$

Therefore, it is preferable to attempt the easier question 1 first.

Example 4.1.14. Consider the following two discrete random variables X_1 and X_2, each taking three values and with pmf given by

$$\mathbb{P}(X_1 = 49) = \mathbb{P}(X_1 = 51) = \frac{1}{4} \quad \text{and} \quad \mathbb{P}(X_1 = 50) = \frac{1}{2};$$

$$\mathbb{P}(X_2 = 0) = \mathbb{P}(X_2 = 50) = \mathbb{P}(X_2 = 100) = \frac{1}{3}.$$

We have that

$$\mathbb{E}(X_1) = 49 \cdot \frac{1}{4} + 51 \cdot \frac{1}{4} + 50 \cdot \frac{1}{2} = 50 \quad \text{and} \quad \mathbb{E}(X_2) = 0 \cdot \frac{1}{3} + 50 \cdot \frac{1}{3} + 100 \cdot \frac{1}{3} = 50.$$

They have the same expected value but X_1 is much less "dispersed" than X_2.

In view of the previous example, we would like to introduce a measure of "dispersion". One way could be to measure how far things are from the expected value, on average. This leads to the notion of variance.

Definition 4.1.15. The **variance** of a discrete random variable X is the quantity

$$\operatorname{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

The **standard deviation** of X is the quantity $\sqrt{\operatorname{var}(X)}$. The **$k$-th moment** of X is $\mathbb{E}(X^k)$.

Notice that in the previous definition we ask that the expectations involved exist.

Remark 4.1.16. How do we compute the variance of a discrete random variable X? We can use the definition of expectation and first compute the pmf of the random variable $(X - \mathbb{E}(X))^2$. Unsurprisingly, a faster way consists in relying on LOTUS, as shown in the following. Let $g(X)$ be the random variable $(X - \mathbb{E}(X))^2$ i.e., the function mapping $\omega \in \Omega$ to $(X(\omega) - \mathbb{E}(X))^2$. LOTUS allows us to write

$$\operatorname{var}(X) = \sum_x (x - \mathbb{E}(X))^2 f_X(x).$$

Clearly, $\operatorname{var}(X) \geq 0$, as the factors of each summand are non-negative. But when is that $\operatorname{var}(X) = 0$? Well, $\operatorname{var}(X) = 0$ if and only if $(x - \mathbb{E}(X))^2 f_X(x) = 0$ for each x. This means that, for each x such that $f_X(x) > 0$, we have $x - \mathbb{E}(X) = 0$. But then the random variable X is not really "random": its value is equal to $\mathbb{E}(X)$ with probability 1.

Example 4.1.17. Let X be the score obtained when throwing a fair die. Compute $\mathrm{var}(X)$.

We know that $\mathbb{E}(X) = 7/2$ and so

$$\begin{aligned}\mathrm{var}(X) &= \sum_x (x - \mathbb{E}(X))^2 f_X(x)\\ &= \frac{1}{6}\left(\left(1-\frac{7}{2}\right)^2 + \left(2-\frac{7}{2}\right)^2 + \left(3-\frac{7}{2}\right)^2 + \left(4-\frac{7}{2}\right)^2 + \left(5-\frac{7}{2}\right)^2 + \left(6-\frac{7}{2}\right)^2\right)\\ &= \frac{35}{12}.\end{aligned}$$

Exercise 4.1.18. *Show that $\mathrm{var}(X_1) \neq \mathrm{var}(X_2)$, where X_1 and X_2 are the random variables in Example 4.1.14.*

Proposition 4.1.19. *Let X be a discrete random variable and let $a, b \in \mathbb{R}$. Then*

(a) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

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

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

Proof. We repeatedly use LOTUS and Remark 4.1.16.

(a)

$$\mathbb{E}(aX+b) = \sum_x (ax+b) f_X(x) = a\sum_x x f_X(x) + b\sum_x f_X(x) = a\mathbb{E}(X) + b.$$

(b)

$$\begin{aligned}\mathrm{var}(aX+b) &= \sum_x (ax + b - \mathbb{E}(aX+b))^2 f_X(x)\\ &= \sum_x (ax - a\mathbb{E}(X))^2 f_X(x)\\ &= a^2 \sum_x (x - \mathbb{E}(X))^2 f_X(x)\\ &= a^2 \mathrm{var}(X).\end{aligned}$$

(c)

$$\begin{aligned}\mathrm{var}(X) &= \sum_x (x - \mathbb{E}(X))^2 f_X(x)\\ &= \sum_x (x^2 - 2x\mathbb{E}(X) + \mathbb{E}(X)^2) f_X(x)\\ &= \sum_x x^2 f_X(x) - 2\mathbb{E}(X)\sum_x x f_X(x) + \mathbb{E}(X)^2 \sum_x f_X(x)\\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2\\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2.\end{aligned}$$

□

(a) and (b) show the behavior of expectation and variance of $g(X)$, when g is a linear function. (c) provides an alternative way of computing the variance.

Example 4.1.20. Let $X \sim \mathit{Bernoulli}(p)$. Recall that $\mathbb{E}(X) = p$. LOTUS and (c) above then imply that $\mathrm{var}(X) = (1^2 \cdot p + 0^2 \cdot (1-p)) - p^2 = p(1-p)$.

Exercise 4.1.21. *Show that the variance of the Poisson random variable with parameter $\lambda > 0$ is λ.*

4.2 Multiple discrete random variables

It is often the case that each outcome of an experiment generates several real numbers of interest. We have seen how to treat these as individual random variables but it is often important to consider their "joint behavior". For example, complicated systems are monitored by several computers that work together to run the system. If one fails or makes an error, the others can override it and the system fails only when a majority of computers fail. If X_i denotes the time until the i-th processor fails, then the time until the system fails depends jointly on the collection of random variables $X_1, \dots, X_n$. As a concrete easy example, consider the following.

Example 4.2.1. We flip a fair coin twice and let X_1 be the number of heads on the first flip, X_2 be the number of heads on the second flip and $Y = 1 - X_1$. Clearly, all these random variables take values in $\{0, 1\}$ and have the same pmf (the constant function $1/2$). So we might think that the pair (X_1, X_2) "behaves" like the pair (X_1, Y). But they are in fact "different". For example, in (X_1, Y), the value of Y is completely determined by that of X_1, whereas the values of X_1 and X_2 are independent. This is not reflected by the pmf of the single random variables and so we need to find a way to encode the information about their "collective behavior". We will focus on the case of two random variables.

Definition 4.2.2. Let X_1 and X_2 be two discrete random variables. Their **joint pmf** is the function defined by

$$f_{X_1,X_2}(x_1, x_2) = \mathbb{P}(\{X_1 = x_1\} \cap \{X_2 = x_2\}).$$

We usually denote $\mathbb{P}(\{X_1 = x_1\} \cap \{X_2 = x_2\})$ by $\mathbb{P}(X_1 = x_1, X_2 = x_2)$.

Notice that $f_{X_1,X_2}(x_1, x_2)$ is a non-negative function from $\mathbb{R}^2$ to $\mathbb{R}$ which is non-zero only on a countable set of points of $\mathbb{R}^2$, namely the vectors whose i-th component is one of the countably many values X_i can take. Moreover, since Ω can be written as the union of pairwise disjoint events of the form $\{X_1 = x_1\} \cap \{X_2 = x_2\}$, we have that

$$\sum_{x_1,x_2} f_{X_1,X_2}(x_1, x_2) = \sum_{x_1,x_2} \mathbb{P}(X_1 = x_1, X_2 = x_2) = \mathbb{P}(\Omega) = 1.$$

Back to our example, we have that f_{X_1,X_2} is the constant function $1/4$, whereas $f_{X_1,Y}$ is $1/2$ at the points $(0, 1)$ and $(1, 0)$ and 0 at the points $(0, 0)$ and $(1, 1)$.

As with the case of one random variable, the purpose of introducing the joint pmf is to extract all the information in the probability measure $\mathbb{P}$ that is relevant to the random variables we are considering. So we should be able to compute the probability of any event defined just in terms of the random variables by simply using their joint pmf. The following analogue of Lemma 3.0.37 in the case of multiple random variables shows that we can indeed do that.

Proposition 4.2.3. *Let X_1 and X_2 be two discrete random variables and let $A \subseteq \mathbb{R}^2$ be any set. Then*

$$\mathbb{P}((X_1, X_2) \in A) = \sum_{(x_1,x_2)\in A} f_{X_1,X_2}(x_1, x_2).$$

The following important result shows that if we know the joint mass function of two random variables, we can find all their separate mass functions. In this context, $f_X(x)$ and $f_Y(y)$ are called **marginal pmf**.

Corollary 4.2.4. *Let X and Y be discrete random variables. Then*

$$f_X(x) = \sum_y f_{X,Y}(x,y) \quad \text{and} \quad f_Y(y) = \sum_x f_{X,Y}(x,y).$$

Proof. By countable additivity, we have that

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x,y).$$

The expression for $f_Y(y)$ is obtained similarly. □

You can think of Corollary 4.2.4 as follows. The joint pmf is represented by a (countable) table, where the number in each square (x,y) is the value $f_{X,Y}(x,y)$. To compute the marginal $f_X(x)$ for a given value of x, we simply add the numbers in the column corresponding to x. Similarly, to compute the marginal pmf $f_Y(y)$ for a given value of y, we add the numbers in the row corresponding to y.

Given a pair (X,Y) of discrete random variables and a function $g\colon \mathbb{R}^2 \to \mathbb{R}$, we can build a new discrete random variable $Z = g(X,Y)$ defined by $Z(\omega) = g(X,Y)(\omega) = g(X(\omega), Y(\omega))$. Similarly to Lemma 4.1.9, the new random variable Z has pmf

$$f_Z(z) = \sum_{(x,y)\colon\, g(x,y)=z} f_{X,Y}(x,y).$$

We then have the following generalized version of the LOTUS, whose proof we omit.

Theorem 4.2.5. $\mathbb{E}(g(X,Y)) = \sum_{x,y} g(x,y) f_{X,Y}(x,y)$.

Corollary 4.2.6 (Linearity of expectation). *Let X and Y be discrete random variables and $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Proof. We have:

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_{x,y} (ax + by) f_{X,Y}(x,y) \\
&= a \sum_x \sum_y x f_{X,Y}(x,y) + b \sum_x \sum_y y f_{X,Y}(x,y) \\
&= a \sum_x \sum_y x f_{X,Y}(x,y) + b \sum_y \sum_x y f_{X,Y}(x,y) \\
&= a \sum_x x \sum_y f_{X,Y}(x,y) + b \sum_y y \sum_x f_{X,Y}(x,y) \\
&= a \sum_x x f_X(x) + b \sum_y y f_Y(y) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y).
\end{aligned}$$

The first equality follows from Theorem 4.2.5. The exchange in the order of summation in the third equality is possible thanks to the absolute convergence of the series $\sum_{x,y}(ax + by) f_{X,Y}(x,y)$. The fifth equality follows from Corollary 4.2.4. □

Corollary 4.2.6 is extremely useful and generalizes to n random variables:

$$\boxed{\mathbb{E}(a_1X_1 + \cdots + a_nX_n) = a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n).}$$

Example 4.2.7. By using the definition of expectation, we have seen that the expectation of a Binomial random variable X with parameters n and p is np. A faster way to obtain this result is the following.

Let X_j be the random variable taking value 1 if the j-th flip results in heads and 0 otherwise (hence $X_j \sim Bernoulli(p)$). As X counts the number of heads in the n flips, we have that $X = X_1 + \cdots + X_n$ and so

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np.$$

Example 4.2.8. Similarly to the previous example, we can use linearity of expectation to compute the expectation of the negative binomial random variable. Let X_1 be the number of trials required for the 1-st success, X_2 the additional number of trials for the 2-nd success and so on until X_r which is the additional number of trials for the r-th success. Then $X = X_1 + \cdots + X_r$. But the trials are independent and so $X_1, \ldots, X_r$ is a family of geometric random variables, each with parameter p. Therefore,

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_r) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_r) = r \cdot \frac{1}{p}.$$

Example 4.2.9 (Coupon collector). Each packet of a product is equally likely to contain any one of n different types of coupon, independently of every other packet. What is the expected number of packets you must buy to obtain at least one of each type of coupon?

Let R be the number of packets required to complete a set of n distinct coupons. We need to compute $\mathbb{E}(R)$. Let T_1 be the number of packets required to obtain the first coupon, T_2 the further number of packets required to obtain a second type of coupon, T_3 the further number required for a third type and so on. Then, $R = \sum_{i=1}^n T_i$. It is easy to see that

$$\mathbb{P}(T_k = r) = \left(\frac{k-1}{n}\right)^{r-1}\left(\frac{n-(k-1)}{n}\right).$$

Hence T_k is a geometric random variable with parameter $\frac{n-k+1}{n}$ and so with mean $\frac{n}{n-k+1}$. Since $R = \sum_{k=1}^n T_k$, we can then conclude by linearity of expectation that

$$\mathbb{E}(R) = n\sum_{k=1}^{n}\frac{1}{k},$$

which is roughly $n \log n$.

Exercise 4.2.10. *A box contains the numbers $1, 2, \ldots, 10$. We pick three numbers at random from the box and compute their sum X. Find $\mathbb{E}(X)$.*

Exercise 4.2.11. *We toss a fair coin 20 times. What is the expected number of runs of 3 heads?*

Exercise 4.2.12. *Let X be the number of fixed points in a random permutation of n items, say for example the number of students in a class of size n who receive their own homework after shuffling. Show that $\mathbb{E}(X) = \mathrm{var}(X) = 1$.*

Exercise 4.2.13. *Given a permutation of the numbers $1, 2, \ldots, n$, a number is called a record if it is bigger than all the preceding numbers. The first number is always a record. For example, in the permutation $3\ 2\ 1\ 5\ 7\ 4\ 6$, the numbers $3, 5, 7$ are records. Let X be the number of records in a random permutation. Compute $\mathbb{E}(X)$.*

4.3 Conditioning discrete random variables

Consider our usual setting $(\Omega, \mathcal{F}, \mathbb{P})$ of a probability space and let X be a discrete random variable. Suppose that we know that some event B occurs with $\mathbb{P}(B) > 0$. We have seen that this gives rise to a (conditional) probability measure, namely the function $P\colon \mathcal{F} \to \mathbb{R}$ defined by $P(A) = \mathbb{P}(A|B)$ (see Lemma 2.5.2). It makes therefore sense to consider the pmf of X with respect to the (conditional) measure P.

Definition 4.3.1. Let X be a discrete random variable and let B be an event with $\mathbb{P}(B) > 0$. The **conditional probability mass function of X given B** is the function $f_{X|B}(x) = \mathbb{P}(X = x|B)$.

Note that, by definition of conditional probability,

$$f_{X|B}(x) = \mathbb{P}(X = x|B) = \frac{\mathbb{P}(X = x, B)}{\mathbb{P}(B)}.$$

This function is clearly non-negative and $\sum_x f_{X|B}(x) = 1$ (hence it is a legitimate pmf). Indeed, the event B can be written as the countable union of pairwise disjoint events of the form $\{X = x\} \cap B$ (where x ranges through the countably many values taken by X) and so

$$\mathbb{P}(B) = \sum_x \mathbb{P}(X = x, B) = \sum_x f_{X|B}(x) \cdot \mathbb{P}(B) = \mathbb{P}(B) \sum_x f_{X|B}(x).$$

Let now X and Y be two discrete random variables associated with the same probability space. If we know that the value of Y is y (with $f_Y(y) > 0$), we can consider the conditional pmf of X given the event $\{Y = y\}$. Definition 4.3.1 adapts as follows: the **conditional pmf of X given $Y = y$** is the function

$$f_{X|Y}(x|y) \overset{\text{def}}{=} \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

The conditional pmf is particularly useful if we want to compute the joint pmf. Indeed, we have $f_{X,Y}(x, y) = f_{X|Y}(x|y) \cdot f_Y(y)$.

Example 4.3.2. Consider four independent rolls of a 6-sided die. Let X be the number of 1's and Y be the number of 2's obtained. What is the joint pmf of the discrete random variables X and Y?

Intuitively, X and Y are "related" and $f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y)$ should be easier to compute than $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. We then try to compute $f_Y(y)$ and $f_{X|Y}(x|y)$ and multiply them to get $f_{X,Y}(x, y)$. Notice first that X and Y are nothing but Binomial random variables with parameters $n = 4$ and $p = 1/6$. Indeed, nothing prevents you to think of the die as a biased coin in which the face 2 represents the outcome heads and all the other faces the outcome tails. Therefore,

$$f_Y(y) = \binom{4}{y} \left(\frac{1}{6}\right)^y \left(\frac{5}{6}\right)^{4-y},$$

for $y \in \{0, 1, 2, 3, 4\}$. Suppose now we have observed that $Y = y$. Then X is the number of 1's in the remaining $4 - y$ rolls, each of which can take one of the remaining values $\{1, 3, 4, 5, 6\}$ with probability $1/5$. This is again a Binomial random variable with parameters $n = 4 - y$ and $p = 1/5$ and so

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \binom{4-y}{x} \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{4-y-x},$$

for $x, y \in \{0, 1, 2, 3, 4\}$ with $0 \leq x + y \leq 4$.

Example 4.3.3. Professor May B. Right often has her facts wrong and answers each of her students' questions incorrectly with probability $1/4$, independently of other questions. In each lecture, May is asked 0, 1, or 2 questions with equal probability $1/3$. What is the probability that, in a lecture, she gives at least one wrong answer?

Let X and Y be the number of questions May is asked and the number of questions she answers wrong in the given lecture, respectively. Recalling that $f_{X,Y}(x,y) = f_{Y|X}(y|x) \cdot f_X(x)$, the desired probability is

$$f_{X,Y}(1,1) + f_{X,Y}(2,1) + f_{X,Y}(2,2) = \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{3}{4} \cdot 2 \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{3}.$$

The conditional pmf can also be used to compute one marginal pmf given the other. Indeed, Corollary 4.2.4 implies that

$$f_X(x) = \sum_y f_{X,Y}(x,y) = \sum_y f_{X|Y}(x|y) \cdot f_Y(y),$$

which is morally the same as the law of total probability.

4.4 Conditional expectation of discrete random variables

A conditional pmf can be thought of as an ordinary pmf over a new universe determined by the conditioning event. In the same spirit, a conditional expectation is the same as an ordinary expectation, except that it refers to the new universe.

Definition 4.4.1. Let X and Y be discrete random variables. The **conditional expectation of X given the event B** is

$$\mathbb{E}(X|B) = \sum_x x f_{X|B}(x),$$

provided that the series is absolutely convergent.

Adapting the above to events of the form $\{Y = y\}$, we obtain the **conditional expectation of X given $Y = y$**:

$$\mathbb{E}(X|Y = y) = \sum_x x f_{X|Y}(x|y).$$

Expectation and conditional expectation are related via the following important result. In words, it basically says that "the unconditional average can be obtained by averaging the conditional averages".

Theorem 4.4.2 (Total expectation theorem). *Let X and Y be discrete random variables. Then*

$$\mathbb{E}(X) = \sum_y \mathbb{E}(X|Y = y) \cdot f_Y(y),$$

provided that the expectations exist.

Proof. Recall that $f_X(x) = \sum_y f_{X|Y}(x|y)f_Y(y)$. Therefore,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_x x f_X(x) \\
&= \sum_x x \sum_y f_{X|Y}(x|y)f_Y(y) \\
&= \sum_x \sum_y x f_{X|Y}(x|y)f_Y(y) \\
&= \sum_y \sum_x x f_{X|Y}(x|y)f_Y(y) \\
&= \sum_y f_Y(y) \sum_x x f_{X|Y}(x|y) \\
&= \sum_y f_Y(y) \cdot \mathbb{E}(X|Y=y),
\end{aligned}$$

where the exchange of summation is possible thanks to absolute convergence. □

Corollary 4.4.3. *Let $B_1, B_2, \dots$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for each i. Then*

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|B_i) \cdot \mathbb{P}(B_i).$$

Proof. Let Y be the discrete random variable that takes the value i if and only if B_i occurs. Clearly,

$$f_Y(i) = \mathbb{P}(Y = i) = \begin{cases} \mathbb{P}(B_i) & \text{for } i = 1, 2, \dots; \\ 0 & \text{otherwise.} \end{cases}$$

By the Total expectation theorem,

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|Y=i) \cdot f_Y(i) = \sum_i \mathbb{E}(X|B_i) \cdot \mathbb{P}(B_i),$$

as $\omega \in B_i$ if and only if $\omega \in \{Y = i\}$. □

Theorem 4.4.2 and Corollary 4.4.3 are the “expectation versions” of the law of total probability.

Example 4.4.4. We have already computed the expectation of the geometric random variable. To show the versatility of the Total expectation theorem, we provide yet another computation. Recall that the pmf of a geometric random variable X with parameter p is given by $f_X(x) = (1-p)^{x-1}p$. We use Corollary 4.4.3 by conditioning on the outcome of the first toss (as it is good practice when we have repeated independent trials). Therefore, consider the events $\{X = 1\}$ (i.e., the first toss gives heads) and its complement $\{X > 1\}$. Clearly, $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X > 1) = 1 - p$.

Given that the first toss is heads, the expected number of tosses before getting heads should be 1 i.e., $\mathbb{E}(X|X = 1) = 1$. Indeed,

$$\mathbb{E}(X|X=1) = \sum_{x=1}^{\infty} x\mathbb{P}(X = x|X = 1) = 1,$$

as only the first term of the series is non-zero.

Intuitively, given that the first toss is tails, the expected number of tosses before getting heads should be $\mathbb{E}(X|X>1)=\mathbb{E}(X)+1$. Let's check it. We have that

$$\mathbb{E}(X|X>1)=\sum_{x=1}^{\infty}x\mathbb{P}(X=x|X>1)=\sum_{x=2}^{\infty}x\mathbb{P}(X=x|X>1).$$

But now observe that, for each $x\geq 2$,

$$\mathbb{P}(X=x|X>1)=\frac{\mathbb{P}(X=x,X>1)}{\mathbb{P}(X>1)}=\frac{\mathbb{P}(X=x)}{\mathbb{P}(X>1)}=(1-p)^{x-2}p=\mathbb{P}(X=x-1).$$

Therefore,

$$\begin{aligned}\mathbb{E}(X|X>1)&=\sum_{x=2}^{\infty}x\mathbb{P}(X=x-1)\\&=\sum_{x=2}^{\infty}(x-1+1)\mathbb{P}(X=x-1)\\&=\sum_{x=2}^{\infty}(x-1)\mathbb{P}(X=x-1)+\sum_{x=2}^{\infty}\mathbb{P}(X=x-1)\\&=\mathbb{E}(X)+1.\end{aligned}$$

Corollary 4.4.3 then implies that

$$\mathbb{E}(X)=\mathbb{E}(X|X=1)\mathbb{P}(X=1)+\mathbb{E}(X|X>1)\mathbb{P}(X>1)=1\cdot p+(1+\mathbb{E}(X))(1-p),$$

from which we obtain $\mathbb{E}(X)=1/p$.

Example 4.4.5. A coin is tossed repeatedly and heads appears at each toss with probability p, where $0<p<1$. Find the expected length of the initial run (this is a run of heads if the first toss gives heads, and of tails otherwise).

As in the previous example, we condition on the result of the first toss. Therefore, let H be the event that the first toss gives heads and let H^c be the event that the first toss gives tails. The pair H, H^c forms a partition of the sample space. Let X be the length of the initial run. We have that

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

for $k=1,2,\ldots$, since if H occurs, then $\{X=k\}$ occurs if and only if the first toss is followed by exactly $k-1$ heads and then a tail. Similarly,

$$\mathbb{P}(X=k|H^c)=(1-p)^{k-1}p,$$

for $k=1,2,\ldots$. Therefore,

$$\mathbb{E}(X|H)=\sum_{k=1}^{\infty}kp^{k-1}(1-p)=\frac{1}{1-p}$$

and

$$\mathbb{E}(X|H^c)=\sum_{k=1}^{\infty}k(1-p)^{k-1}p=\frac{1}{p},$$

where we used the fact that the two sums are nothing but the expectations of geometric random variables with parameters $1-p$ and p, respectively (see Example 4.1.7). But then Corollary 4.4.3 implies that

$$\mathbb{E}(X) = \mathbb{E}(X|H)\mathbb{P}(H) + \mathbb{E}(X|H^c)\mathbb{P}(H^c) = \frac{1}{1-p} \cdot p + \frac{1}{p} \cdot (1-p).$$

Exercise 4.4.6. *Show that the geometric random variable X has the **lack of memory** property. Namely, $\mathbb{P}(X > m+n|X > m) = \mathbb{P}(X > n)$, for each m and n in $\mathbb{N}$.*

Exercise 4.4.7. *Let N be the number of tosses of a fair coin up to and including the appearance of the first head. By conditioning on the result of the first toss, show that $\mathbb{E}(N) = 2$.*

Exercise 4.4.8. *I try to open a door with one of the 5 similar keys in my pocket; one of them is correct, the other four will not turn.*

(a) Let X be the number of attempts necessary if I choose a key at random from my pocket and drop those that fail on the floor. Compute $\mathbb{E}(X)$.

(b) Let Y be the number of attempts necessary if I choose a key at random from my pocket and put those that fail back in my pocket. Compute $\mathbb{E}(Y)$.

Exercise 4.4.9. *The probability of obtaining a head when a certain coin is tossed is p. The coin is tossed repeatedly until n heads occur in a row. Let X be the total number of tosses required for this to happen. Compute $\mathbb{E}(X)$.*

4.5 Independence of discrete random variables

Definition 4.5.1. Two discrete random variables X and Y are **independent** if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for each pair $(x,y) \in \mathbb{R}^2$. More generally, a family of n discrete random variables $X_1, \dots, X_n$ is independent if

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

for each $(x_1, \dots, x_n) \in \mathbb{R}^n$. Finally, an arbitrary family of random variables is independent if each finite subfamily is.

Notice that X and Y are independent if and only if the events $\{X = x\}$ and $\{Y = y\}$ are independent for each $(x,y) \in \mathbb{R}^2$. Recall that $f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y)$. Therefore, X and Y are independent if and only if $f_{X|Y}(x|y) = f_X(x)$ for each y with $f_Y(y) > 0$ and for each x i.e., the experimental value of Y tells us nothing about the value of X.

Example 4.5.2. Consider again the random variables X_1, X_2 and $Y = 1 - X_1$ in Example 4.2.1. We have that X_1 and X_2 are independent but X_1 and Y are not.

Example 4.5.3. Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2, respectively. What is the pmf of $X_1 + X_2$? We need to determine $\mathbb{P}(X_1 + X_2 = n)$.

We use the Law of total probability and independence:

$$\begin{aligned}
\mathbb{P}(X_1+X_2=n) = \sum_{k=0}^{n}\mathbb{P}(X_1=k, X_2=n-k) &= \sum_{k=0}^{n}\mathbb{P}(X_1=k)\mathbb{P}(X_2=n-k) \\
&= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!}\cdot e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!} \\
&= e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^k\lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!},
\end{aligned}$$

where in the last equality we used the Binomial theorem. The computation shows that X_1+X_2 is a Poisson random variable with parameter $\lambda_1+\lambda_2$.

Exercise 4.5.4. *Let X_1 and X_2 be independent Poisson random variables with parameters λ_1 and λ_2, respectively.*

(i) Compute the conditional pmf of X_1 given that $X_1+X_2=n$.

(ii) Using (i), compute the conditional expectation of X_1 given that $X_1+X_2=n$.

(iii) Suppose that $\lambda_1=\lambda_2$. Explain how in this case we can instead use symmetry and linearity of expectation in order to deduce that $\mathbb{E}(X_1|X_1+X_2=n)=\frac{1}{2}\mathbb{E}(X_1+X_2|X_1+X_2=n)$ and compute the actual value of $\mathbb{E}(X_1|X_1+X_2=n)$.

Theorem 4.5.5. *Let X and Y be independent discrete random variables. Then*

(a) For any sets $A, B\subseteq\mathbb{R}$, we have that $\mathbb{P}(X\in A, Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B)$.

(b) For any functions $g,h\colon\mathbb{R}\to\mathbb{R}$, we have that $g(X)$ and $h(Y)$ are independent.

Notice that (a) extends to a family $X_1,\dots,X_n$ of n independent discrete random variables: for any sets $S_1,\dots,S_n\subseteq\mathbb{R}$, we have

$$\mathbb{P}(X_1\in S_1,\dots,X_n\in S_n)=\prod_{i=1}^{n}\mathbb{P}(X_i\in S_i).$$

Proof. (a) We have that $\{X\in A, Y\in B\}=\bigcup_{x\in A,\ y\in B}\{X=x, Y=y\}$, where the union is countable since both X and Y are discrete random variables. But then countable additivity and independence imply that

$$\begin{aligned}
\mathbb{P}(X\in A, Y\in B) &= \sum_{x\in A}\sum_{y\in B}\mathbb{P}(X=x, Y=y) \\
&= \sum_{x\in A}\sum_{y\in B}\mathbb{P}(X=x)\mathbb{P}(Y=y) \\
&= \Big(\sum_{x\in A}\mathbb{P}(X=x)\Big)\Big(\sum_{y\in B}\mathbb{P}(Y=y)\Big) \\
&= \mathbb{P}(X\in A)\mathbb{P}(Y\in B).
\end{aligned}$$

(b) It will be part of a homework assignment. □

Exercise 4.5.6. *Let X and Y be independent geometric random variables with pmf's $f_X(x) = (1-\lambda)\lambda^{x-1}$ and $f_Y(y) = (1-\mu)\mu^{y-1}$, respectively. Find the pmf of $Z = \min\{X, Y\}$.*

It is in general not true that, given two discrete random variables X and Y, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ holds. Consider for example the random variable X taking values 1 and -1, each with probability $1/2$. Then $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = 1$. Taking $Y = X$ we see that indeed $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ does not hold. However, the situation changes if X and Y are independent:

Theorem 4.5.7. *Let X and Y be independent discrete random variables. Then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof. We use Theorem 4.2.5 with the function $g\colon \mathbb{R}^2 \to \mathbb{R}$ given by $g(x,y) = xy$:

$$\mathbb{E}(XY) = \sum_x \sum_y xy f_{X,Y}(x,y) = \sum_x \sum_y xy f_X(x) f_Y(y) = \sum_x x f_X(x) \sum_y y f_Y(y) = \mathbb{E}(X)\mathbb{E}(Y),$$

where the second equality follows by independence. □

Remark 4.5.8. If X and Y are independent, we have seen that $g(X)$ and $h(Y)$ are independent as well and so, by Theorem 4.5.7, $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$.

Recall that the expectation is linear. In particular, $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ for any two random variables X and Y. Although not true in general[2], variance is linear for families of independent random variables.

Theorem 4.5.9. *If X and Y are independent discrete random variables, then $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.*

Proof. We have

$$\begin{aligned}
\mathrm{var}(X+Y) &= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) - (\mathbb{E}(X))^2 - (\mathbb{E}(Y))^2 - 2\mathbb{E}(X)\mathbb{E}(Y) \\
&= \mathrm{var}(X) + \mathrm{var}(Y),
\end{aligned}$$

where in the first equality we used Proposition 4.1.19(c), in the second we used LOTUS to expand the first term and linearity of expectation to expand the second term, and in the last equality we used Theorem 4.5.7. □

Theorem 4.5.9 generalizes as follows:

If $X_1, \dots, X_n$ are *independent* discrete random variables, then

$$\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n).$$

Example 4.5.10. Let us compute the variance of a Binomial random variable X with parameters n and p. Recall from Example 4.2.7 that X can be written as the sum of n Bernoulli random variables X_i with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$. By independence of coin tosses, $X_1, \dots, X_n$ are independent and so $\mathrm{var}(X) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n) = np(1-p)$.

[2] For a counterexample consider again $X = Y$, where X takes values 1 and -1, each with probability $1/2$.

We now introduce an indicator of "dependence" between two random variables:

Definition 4.5.11. The **covariance** of the discrete random variables X and Y is

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]$$

and X and Y are **uncorrelated** if $\operatorname{cov}(X,Y)=0$.

The covariance of two random variables is a measure of their tendency to be larger than their expected value together. A negative covariance means that when one of the random variables is larger than its mean, the other is more likely to be less than its mean. By linearity of expectation, we have that

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

and so $\operatorname{cov}(X,Y)=0$ if and only if $\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$. Even though independence implies uncorrelation, the converse is not true, as shown in the following:

Example 4.5.12. Let X be a discrete random variable such that $f_X(x) = f_X(-x)$ for each $x \in \operatorname{Im}(X)$. Suppose that $\mathbb{E}(X^3)$ exists and let $Y = X^2$. Clearly, X and Y are not independent. However, by LOTUS, we have

$$\mathbb{E}(XY) = \mathbb{E}(X^3) = \sum_{x>0} x^3(f_X(x) - f_X(-x)) = 0.$$

Similarly,

$$\mathbb{E}(X) = \sum_{x>0} x(f_X(x) - f_X(-x)) = 0$$

and so $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

4.6 Weak law of large numbers

Throughout this section, we assume that the random variables we are working with are discrete. Notice however that all stated results hold for any type of random variables.

Very often, in probability, we want to assert that some sequence of random variables tends to a limit in a suitable probabilistic sense. Limit theorems are useful for several reasons:

- They provide an interpretation of expectations in terms of a long sequence of identical independent experiments.
- They allow for an approximate analysis of the properties of random variables such as $\frac{X_1+\cdots+X_n}{n}$, where in contrast an exact analysis might reveal to be a complicated task.
- They describe the long term behavior of a *stochastic process*, where a stochastic process is nothing but a sequence $\{X_n\}$ of random variables indexed by time n.

We begin by considering the following classical situation. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (i.i.d. for short) random variables with expectation μ and variance σ^2. We look at the random variable

$$M_n = \frac{X_1+\cdots+X_n}{n}.$$

By linearity of expectation,

$$\mathbb{E}(M_n) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \mu.$$

Since $X_1, \ldots, X_n$ are independent, Proposition 4.1.19 and Theorem 4.5.9 imply that

$$\operatorname{var}(M_n) = \frac{1}{n^2}\operatorname{var}(X_1 + \cdots + X_n) = \frac{1}{n^2}(\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)) = \frac{\sigma^2}{n}.$$

In particular, the variance of M_n decreases to 0 as n increases. This phenomenon is the subject of the so-called Laws of large numbers (Weak and Strong), asserting that the random variables M_n converge to μ in a precise sense. We will see how this provides mathematical justification for the loose interpretation of the expectation of a random variable X as the average of a large number of independent samples drawn from the distribution of X.

In order to make the discussion above more precise, we need to introduce some probability inequalities. We remark that they hold for continuous random variables as well (with almost identical proofs, provided we define the expectation and variance of a continuous random variable) but we should content ourselves with discrete ones.

Theorem 4.6.1 (Markov's inequality). *Let X be a non-negative random variable. Then, for each $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. Fix $a > 0$ and consider the discrete random variable Y_a defined by

$$Y_a = \begin{cases} 0 & \text{if } X < a; \\ a & \text{if } X \geq a; \end{cases}$$

By construction, $X \geq Y_a$ (this should be understood as $X(\omega) \geq Y_a(\omega)$ for each $\omega \in \Omega$). But then, using monotonicity of expectation (show that it indeed holds!), we have

$$\mathbb{E}(X) \geq \mathbb{E}(Y_a) = a\mathbb{P}(Y_a = a) = a\mathbb{P}(X \geq a),$$

as claimed. □

In words, if a non-negative random variable has small expectation, then the probability that it takes a large value is small.

Theorem 4.6.2 (Chebyshev's inequality). *Let X be a random variable. Then, for each $c > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{\operatorname{var}(X)}{c^2}.$$

Proof. We apply Markov's inequality to the non-negative random variable $(X - \mathbb{E}(X))^2$ and take $a = c^2$:

$$\mathbb{P}((X - \mathbb{E}(X))^2 \geq c^2) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{c^2} = \frac{\operatorname{var}(X)}{c^2}.$$

But the event $\{(X - \mathbb{E}(X))^2 \geq c^2\}$ is the same as the event $\{|X - \mathbb{E}(X)| \geq c\}$ and the conclusion follows. □

In words, if a random variable has small variance, then the probability that it takes a value far from its expectation is small.

Example 4.6.3. Let X be the random variable counting the number of students in a class of size n who receive their own homework after shuffling. Recall from Exercise 4.2.12 that $\mathbb{E}(X) = \mathrm{var}(X) = 1$. We now want to estimate $\mathbb{P}(X \geq 20)$. By monotonicity and Chebyshev's inequality,

$$\mathbb{P}(X \geq 20) \leq \mathbb{P}(|X-1| \geq 19) \leq \frac{1}{19^2}.$$

Notice this is independent of the class size n.

Example 4.6.4. Let X be a discrete random variable with $\mathbb{E}(X) = \mathbb{E}(X^2) = 0$. Then $X = 0$ almost surely i.e., $\mathbb{P}(X = 0) = 1$. We observed this in Remark 4.1.16. We now deduce it using Chebyshev's inequality. Since $\mathrm{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = 0$, Chebyshev's inequality implies that $\mathbb{P}(|X| \geq c) = 0$ for each $c > 0$. But $\{|X| > 0\} = \bigcup_{k\in\mathbb{N}}\{|X| > 1/k\}$ and so the union bound implies that

$$\mathbb{P}(|X| > 0) \leq \sum_{k=1}^{\infty} \mathbb{P}(|X| > 1/k) = 0,$$

from which we obtain that $\mathbb{P}(X = 0) = 1$.

Exercise 4.6.5. *Let X be a discrete random variable with $\mathbb{E}(X) = 10$ and $\mathrm{var}(X) = 5$. Estimate the probability $\mathbb{P}(3 < X < 15)$.*

Let's now go back to our sequence $X_1, X_2, \ldots$ of i.i.d. random variables with expectation μ and variance σ^2. We have seen that the random variable

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

has expectation $\mathbb{E}(M_n) = \mu$ and variance $\mathrm{var}(M_n) = \sigma^2/n$. Applying Chebyshev's inequality to M_n and taking $c = \varepsilon$, we have that for each $\varepsilon > 0$,

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

But for fixed $\varepsilon > 0$, the RHS goes to 0 as $n \to \infty$. We have therefore proved the following:

Theorem 4.6.6 (Weak law of large numbers, WLLN). *Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ. Then, for each $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0$$

as $n \to \infty$.

Remark 4.6.7. Notice that, if we drop the independence assumption, the theorem will in general be false.

A consequence of the WLLN is the following interpretation of the expectation of a random variable: The arithmetic average of a sequence of independent observations of a random variable X converges with high probability to $\mathbb{E}(X)$. More precisely,

> we can estimate the expectation of a random variable with any amount of precision with arbitrary probability if we use a sufficiently large number of samples of its value.

The WLLN is not completely satisfactory as it just states that the probability $\mathbb{P}(|M_n - \mu| \geq \varepsilon)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$. Still, for any $n \in \mathbb{N}$, this probability might be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ. The problem of the WLLN is that it deals with a somewhat weak notion of convergence, called convergence in probability. What would back our intuitive notion of expectation though, is another notion of convergence, called convergence almost surely, according to which M_n converges to μ with probability 1. This is the content of the so-called Strong law of large numbers, which implies that, for any given $\varepsilon > 0$, the difference $|M_n - \mu|$ exceeds ε only finitely many times. We will not go into details.

Chapter 5

Continuous random variables

In this section we are essentially going to revisit the notions introduced in Chapter 3, this time in the context of continuous random variables. Recall that a random variable X is continuous if its distribution function F_X can be expressed as

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u) \, \mathrm{d}u,$$

for some integrable function $f_X \colon \mathbb{R} \to [0, \infty)$ called the pdf of X. Moreover, if the distribution function is differentiable at $x \in \mathbb{R}$, then

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}.$$

The following properties hold for a continuous random variable X and its pdf f_X:

1. $\mathbb{P}(x < X \leq y) = \int_x^y f_X(u) \, \mathrm{d}u$.
2. $\mathbb{P}(X = x) = 0$.
3. $\int_{-\infty}^{\infty} f_X(u) \, \mathrm{d}u = 1$.

Definition 5.0.1. Let X be a continuous random variable with pdf $f_X(x)$. The **expectation** of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x,$$

provided that the following improper integral converges:

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

Remark 5.0.2. The convergence of the improper integral guarantees that the expectation is well-defined and finite.

Example 5.0.3. Recall that the uniform random variable X on the interval $[a, b]$ has pdf given by

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b; \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

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

We have seen that if X is an arbitrary random variable and $g\colon \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X)$ is a random variable. In fact, if X is discrete, $g(X)$ is discrete for any function g. On the other hand, if X is continuous, $g(X)$ can be either continuous or discrete. The former occurs for example if g is the identity, the latter by taking g as follows (see hw9):

$$g(x) = \begin{cases} 1 & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Given the pdf of X, how do we compute the pdf of $g(X)$? The standard technique is illustrated in the following example.

Example 5.0.4. Let X be a continuous random variable with pdf

$$f_X(x) = \begin{cases} 2x & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Find the pdf of $Y = 2X - 1$.

We first obtain the distribution function of Y and then differentiate. We have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(2X - 1 \le y) = \mathbb{P}\Big(X \le \frac{y+1}{2}\Big) = F_X\Big(\frac{y+1}{2}\Big).$$

Therefore, for each $-1 < y < 1$, we have

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{1}{2} \cdot f_X\Big(\frac{y+1}{2}\Big) = \frac{y+1}{2},$$

where the second equality follows from the Chain rule.

Example 5.0.5. Given a continuous random variable X, we compute the pdf of $Y = X^2$. As in the previous example, we first obtain the distribution function of Y and then differentiate. For each $y > 0$, we have

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

where in the last equality we used Lemma 3.0.36. Therefore,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{f_X(\sqrt{y})}{2\sqrt{y}} + \frac{f_X(-\sqrt{y})}{2\sqrt{y}},$$

where the last equality follows from the Chain rule.

If we are just interested in the expectation of $g(X)$, we can however skip the computation of the pdf of $g(X)$, thanks to the following result. It is the continuous analogue of the Law of the unconscious statistician.

Theorem 5.0.6 (LOTUS). *If X and $g(X)$ are continuous random variables, then*

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

Example 5.0.7. Let X be a continuous random variable with pdf

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Compute the expectation of $Y = X^2$.

We can proceed in two ways. Either we compute the pdf of Y and use the definition of expectation or just apply LOTUS. As for the first way, we have seen in Example 5.0.5 that

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} = \begin{cases} \dfrac{3y}{2\sqrt{y}} & \text{if } 0 < y < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = \int_0^1 \frac{3}{2} y^{3/2}\,\mathrm{d}y = \frac{3}{5}.$$

Using LOTUS, we immediately get

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_0^1 3x^4\,\mathrm{d}x = \frac{3}{5}.$$

Definition 5.0.8. Let $\lambda > 0$. A continuous random variable X with pdf given by

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

is called **exponential** with parameter λ, denoted by $X \sim Exp(\lambda)$.

What is the distribution function of X? Since X is continuous, we have that

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u = \int_0^x \lambda e^{-\lambda u}\,\mathrm{d}u = 1 - e^{-\lambda x}.$$

Therefore, for any $a \geq 0$,

$$\mathbb{P}(X > a) = 1 - \mathbb{P}(X \leq a) = 1 - (1 - e^{-\lambda a}) = e^{-\lambda a} \tag{5.1}$$

and so the probability that X exceeds a falls exponentially. The exponential random variable is a good model for the amount of time until a certain event occurs. For example, the amount of time until a piece of equipment breaks down, until a car accident occurs or until the next earthquake. Using integration by parts, it is easy to see that the expectation of an exponential random variable X with parameter λ is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_0^{\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}.$$

Example 5.0.9. Suppose that the duration of a phone call in minutes is an exponential random variable X with parameter $\lambda = 1/10$. What is the probability that the phone call lasts more than 10 minutes? This is just $\mathbb{P}(X > 10) = e^{-1}$. Suppose now we know that the phone call has already lasted 10 minutes. What is the probability that it will last at least 10 more minutes? The probability we are interested in is

$$\mathbb{P}(X > 20 | X > 10) = \frac{\mathbb{P}(X > 20, X > 10)}{\mathbb{P}(X > 10)} = \frac{\mathbb{P}(X > 20)}{\mathbb{P}(X > 10)} = \frac{e^{-2}}{e^{-1}} = e^{-1}.$$

The same argument used in the previous example shows that if Y is exponential, then

$$\mathbb{P}(Y > t + s | Y > s) = \mathbb{P}(Y > t),$$

for each $t, s > 0$. But this is the lack of memory property and we have seen that, in the discrete world, the geometric random variable has this property. It turns out that the exponential random variable can be viewed as the continuous analogue of the geometric random variable in the following sense. Suppose that $X \sim \mathit{Geometric}(p)$, for p small, and recall that $\mathbb{P}(X > n) = (1-p)^n$. Consider now the rescaled random variable $X/\mathbb{E}(X) = pX$. We have that

$$\mathbb{P}(pX > t) = \mathbb{P}(X > t/p) \approx (1-p)^{\frac{t}{p}} \approx e^{-t},$$

where in the last approximation we used the fact that $e^{-1} = \lim_{n\to\infty}(1 - 1/n)^n$. In other words, the rescaled geometric pX behaves like the exponential with parameter $\lambda = 1$ when p is small.

Definition 5.0.10. The **variance** of a continuous random variable X is defined exactly as in the discrete case: $\text{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$.

By LOTUS, the variance can be computed as follows:

$$\begin{aligned}\text{var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x - 2\mathbb{E}(X)\int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x + \mathbb{E}(X)^2 \int_{-\infty}^{\infty} f_X(x)\, \mathrm{d}x \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2,\end{aligned}$$

exactly as in the discrete case.

Example 5.0.11. Let $X \sim \mathit{Exp}(\lambda)$. Then $\text{var}(X) = 1/\lambda^2$. Indeed, we know that $\mathbb{E}(X) = 1/\lambda$. Moreover, by LOTUS and integration by parts, we have that

$$\mathbb{E}(X^2) = \int_0^{\infty} x^2 f_X(x)\, \mathrm{d}x = \frac{2}{\lambda^2}.$$

Exercise 5.0.12. *Let X be the uniform random variable on the interval $[a, b]$.*

1. *Compute* $\text{var}(X)$*;*
2. *Compute the pdf of $Y = X^2$.*

Exercise 5.0.13. *During any 8-hour shift, the proportion of time X that a machine is down for maintenance or repairs has pdf given by:*

$$f_X(x) = \begin{cases} 2(1-x) & \textit{if } 0 \le x \le 1; \\ 0 & \textit{otherwise.} \end{cases}$$

The total cost in \$ of this downtime, due to lost production, maintenance and repair, is given by $C = 200(5 + 10X + 2X^2)$. Compute $\mathbb{E}(C)$.

Exercise 5.0.14. *The radius of a circle is a uniform random variable on $[1, 2]$. What is the pdf of the area of the circle?*

Exercise 5.0.15. *The amount of time needed to wash a car at a car washing station is an exponential random variable with expected value of 15 minutes. You arrive while the washing station is occupied and one other car is waiting for a washing. The owner of this car informs you that the car in the washing station has already been there for 10 minutes. What is the probability that the car in the washing station will need no more than 5 other minutes?*

5.1 Multiple continuous random variables

Definition 5.1.1. Two continuous random variables X and Y admit **joint pdf** if there exists a non-negative integrable function $f_{X,Y}\colon \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y,$$

for every $B \subseteq \mathbb{R}^2$ for which the double integral exists.

We will not need the proper definition of double integral; it will be enough to know that it computes the volume of the 3-dimensional region between the portion of the plane B and the surface $z = f_{X,Y}(x,y)$.

In many cases, the double integral can be computed as an *iterated integral*. For example, if B is the rectangle $[a,b] \times [c,d] \subseteq \mathbb{R}^2$, we have that

$$\mathbb{P}((X,Y) \in B) = \mathbb{P}(a \le X \le b, c \le Y \le d) = \int_c^d \left(\int_a^b f_{X,Y}(x,y)\,\mathrm{d}x \right) \mathrm{d}y,$$

where in the inner integral we integrate with respect to x (and hence y is treated as a constant) and the result is then integrated with respect to y.

Remark 5.1.2. Contrary to the discrete case, a joint pdf might not exist. Here is the intuition. Take X as the uniform random variable on $[0,1]$ and $Y = X$. Suppose that X and Y admit a joint pdf $f_{X,Y}$ and let $B = \{(x,y) \in [0,1] \times [0,1] : x = y\}$ be the main diagonal of the unit square. We have that $\mathbb{P}((X,Y) \in B) = 1$ and so

$$1 = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

This double integral gives the volume of the region under the surface $z = f_{X,Y}(x,y)$ and above B. But B has area 0, and so we expect the volume to be 0, not 1!

As in the discrete case, we can recover the densities of X and Y from the joint pdf:

Lemma 5.1.3. *For continuous random variables X and Y with joint pdf $f_{X,Y}$, we have*

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

Example 5.1.4 (Two-dimensional uniform pdf). Suppose $S \subseteq \mathbb{R}^2$ has finite area. We say that the pair (X,Y) of continuous random variables is **uniformly distributed** over S if the joint pdf of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\operatorname{area}(S)} & \text{if } (x,y) \in S; \\ 0 & \text{otherwise.} \end{cases}$$

The idea is that we pick a point (X,Y) inside S at random and all points are equally likely. This is similar to the 1-dimensional case of the uniform random variable on the interval $[a,b]$ (see Example 3.0.12).

By definition of joint pdf, we then have that, for $B \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{(x,y)\in B\cap S} \frac{1}{\mathrm{area}(S)}\,\mathrm{d}x\mathrm{d}y = \frac{\mathrm{area}(B\cap S)}{\mathrm{area}(S)},$$

where the second equality follows from the fact that $f_{X,Y}(x,y) = 0$ for each $(x,y) \notin S$, and the last equality follows from the fact that the volume of the solid with base $B \cap S$ and constant height $\frac{1}{\mathrm{area}(S)}$ is $\frac{\mathrm{area}(B\cap S)}{\mathrm{area}(S)}$.

As a concrete example, consider the following situation. Suppose that a point is chosen at random from an open unit disk. What is the probability that the sum of its coordinates is larger than 1?

Let $B_1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ be the open unit disk centered at the origin and let (X,Y) be the coordinates of our random point. It is reasonable to assume that (X,Y) is uniformly distributed over B_1:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } (x,y) \in B_1; \\ 0 & \text{otherwise.} \end{cases}$$

We need to compute $\mathbb{P}(X+Y>1)$. It is then enough to compute the area of the intersection between B_1 and the half-plane $\{(x,y) \in \mathbb{R}^2 : x+y > 1\}$. Drawing a picture, it is easy to see that this area is $\frac{\pi}{4} - \frac{1}{2}$ and so $\mathbb{P}(X+Y>1) = \frac{1}{4} - \frac{1}{2\pi}$.

Example 5.1.5 (Buffon's needle). We throw a needle of length ℓ at random on a surface marked with horizontal lines at distance d (see Figure 5.1). Assume $\ell < d$, so that the needle can intersect at most one horizontal line. What is the probability that the needle will intersect one of these lines?

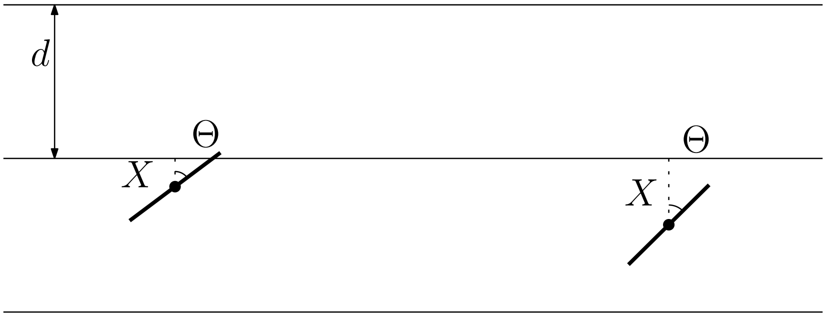


Figure 5.1

Consider the midpoint of the needle and the vertical segment between the midpoint and the closest horizontal line (the dotted lines in Figure 5.1). Let X be the length of this segment and let Θ be the acute angle between the needle and the segment. The pair of random variables (X,Θ) uniquely determines the position of the needle and we may assume it is uniformly distributed over $R = [0,\frac{d}{2}] \times [0,\frac{\pi}{2}]$. We then have that

$$f_{X,\Theta}(x,\theta) = \begin{cases} \dfrac{4}{\pi d} & \text{if } (x,\theta) \in R; \\ 0 & \text{otherwise.} \end{cases}$$

The needle will intersect one of the lines if and only if

$$\frac{X}{\cos\Theta} < \frac{\ell}{2}.$$

Therefore, the desired probability is

$$\iint_{(x,\theta)\in A} f_{X,\Theta}(x,\theta)\,\mathrm{d}x\mathrm{d}\theta,$$

where $A = \{(x,\theta) \in \mathbb{R}^2 : 0 \le x \le d/2, 0 \le \theta \le \pi/2, x < \ell\cos\theta/2\}$. The double integral can be computed as follows:

$$\iint_{(x,\theta)\in A} f_{X,\Theta}(x,\theta)\,\mathrm{d}x\mathrm{d}\theta = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\ell\cos\theta}{2}} \frac{4}{\pi d}\,\mathrm{d}x\mathrm{d}\theta = \frac{2\ell}{\pi d}.$$

This suggests a way to calculate π: Throw the needle a large number of times, count the number of intersections in the first n tosses and divide by n. This will give an estimate of the true probability $2\ell/\pi d$ and so

$$\pi \approx \frac{2n\ell}{(\#\text{intersections in first } n \text{ tosses}) \cdot d}.$$

A generalized version of LOTUS still holds:

Lemma 5.1.6. *Let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y)$ and let $g\colon \mathbb{R}^2 \to \mathbb{R}$. Then*

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

In particular, $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

Exercise 5.1.7. *We throw a dart at a circular target centered at the origin and of radius r. We assume that we always hit the target and that all points of impact (X,Y) are equally likely. Compute $\mathbb{P}(Y \le r/2)$.*

Exercise 5.1.8. *A husband and wife agree to meet at a street corner between 15:00 and 16:00 to go shopping together. The one who arrives first will await the other for 15 minutes, and then leave. What is the probability that the two meet within the given time interval, assuming that they can arrive at any time with the same probability?*

5.2 Conditioning continuous random variables

Definition 5.2.1. Let X be a continuous random variable and let A be an event with $\mathbb{P}(A) > 0$. The **conditional pdf** of X is the nonnegative integrable function $f_{X|A}$ satisfying

$$\mathbb{P}(X \in B|A) = \int_B f_{X|A}(x)\,\mathrm{d}x,$$

for every $B \subseteq \mathbb{R}$ for which the integral exists.

Definition 5.2.2. The **conditional expectation** of a continuous random variable X is defined as

$$\mathbb{E}(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x)\,\mathrm{d}x.$$

We then have the following continuous analogue of Corollary 4.4.3.

Theorem 5.2.3. *Let $A_1, \dots, A_n$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i and let X be a continuous random variable. Then*

(a) $f_X(x) = \sum_{i=1}^n f_{X|A_i}(x) \cdot \mathbb{P}(A_i)$.

(b) $\mathbb{E}(X) = \sum_{i=1}^n \mathbb{E}(X|A_i) \cdot \mathbb{P}(A_i)$.

Example 5.2.4. The metro train arrives every 15 minutes starting at 6am. You walk into the station every morning between 7:10am and 7:30am with the time of arrival in this interval being a uniform random variable. What is the pdf of the time you have to wait for the first train?

Let X be the time of arrival. We know it is a uniform random variable on the interval between 7:10 and 7:30. Let Y be the waiting time. We want $f_Y(y)$. As the waiting time depends on whether you manage to take the 7:15 train or not, we consider the following partition:

$$A_1 = \{7:10 \le X \le 7:15\} \quad A_2 = \{7:15 < X \le 7:30\}.$$

Conditioned on A_1, the arrival time is uniform on the interval 7:10-7:15 and so the waiting time is uniform on $[0, 5]$. In other words,

$$f_{Y|A_1}(y) = \begin{cases} \dfrac{1}{5} & \text{if } 0 \le y \le 5; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, conditioned on A_2, the arrival time is uniform on the interval 7:15-7:30 and so the waiting time is uniform on $[0, 15]$. In other words,

$$f_{Y|A_2}(y) = \begin{cases} \dfrac{1}{15} & \text{if } 0 \le y \le 15; \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 5.2.3, we have that

$$f_Y(y) = \mathbb{P}(A_1) f_{Y|A_1}(y) + \mathbb{P}(A_2) f_{Y|A_2}(y).$$

Since X is uniform on the interval 7:10-7:30 of length 20, we know that $\mathbb{P}(A_1) = 5/20$ and $\mathbb{P}(A_2) = 15/20$. Combining, we obtain

$$f_Y(y) = \begin{cases} 1/10 & \text{if } 0 \le y \le 5; \\ 1/20 & \text{if } 5 < y \le 15. \end{cases}$$

Continuing the analogy with discrete random variables, we would now like to condition on events of the form $Y = y$. But we know that if Y is continuous, $\mathbb{P}(Y = y) = 0$. How do we interpret then probabilities of the form $\mathbb{P}(X \in A | Y = y)$? We will make use of the following notion.

Definition 5.2.5. Let X and Y be continuous random variables with joint pdf $f_{X,Y}$. For any fixed y with $f_Y(y) > 0$, the **conditional pdf of X given that $Y = y$** is defined by

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

Notice that it agrees with the definition of conditional pmf in the discrete case. Viewing $f_{X|Y}(x|y)$ as a function of x, it has the same shape as $f_{X,Y}$. The normalization by $f_Y(y)$ implies that $f_{X|Y}(x|y)$ is a legitimate pdf:

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

But how do we interpret $f_{X|Y}(x|y)$? Fix small δ_1 and δ_2 and consider the following conditional probability:

$$\begin{aligned}
\mathbb{P}(x \le X \le x+\delta_1 | y \le Y \le y+\delta_2) &= \frac{\mathbb{P}(x \le X \le x+\delta_1, y \le Y \le y+\delta_2)}{\mathbb{P}(y \le Y \le y+\delta_2)} \\
&= \frac{\int_x^{x+\delta_1}\int_y^{y+\delta_2} f_{X,Y}(x,y)\,\mathrm{d}y\mathrm{d}x}{\int_y^{y+\delta_2} f_Y(y)\,\mathrm{d}y} \\
&\approx \frac{f_{X,Y}(x,y)\delta_1\delta_2}{f_Y(y)\delta_2} \\
&= f_{X|Y}(x|y)\delta_1.
\end{aligned}$$

Letting $\delta_2 \to 0$, we have that $\mathbb{P}(x \le X \le x+\delta_1 | Y = y)$ should approximately be $f_{X|Y}(x|y)\delta_1$ for small δ_1. We then make the following definition in the continuous case:

$$\mathbb{P}(X \in A | Y = y) \overset{\text{def}}{=} \int_A f_{X|Y}(x|y)\,\mathrm{d}x.$$

Example 5.2.6. We throw a dart at a circular target of radius r. We assume that we always hit the target and that all points of impact (X, Y) are equally likely. In other words, we assume that the joint pdf of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{\pi r^2} & \text{if } x^2+y^2 \le r^2; \\ 0 & \text{otherwise.} \end{cases}$$

What is the conditional pdf $f_{X|Y}(x|y)$? We first compute the marginal $f_Y(y)$. By Lemma 5.1.3,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \begin{cases} 0 & \text{if } |y| > r; \\ \displaystyle\int_{x:x^2+y^2\le r^2} \frac{1}{\pi r^2}\,\mathrm{d}x = \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} \frac{1}{\pi r^2}\,\mathrm{d}x = \frac{2}{\pi r^2}\sqrt{r^2-y^2} & \text{if } |y| \le r. \end{cases}$$

Therefore, $f_{X|Y}(x|y) = \frac{1}{2\sqrt{r^2-y^2}}$.

Definition 5.2.7. The conditional expectation $\mathbb{E}(X|Y=y)$ is defined as $\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x$.

We have that $\mathbb{E}(g(X)|Y=y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\,\mathrm{d}x$. Moreover, the following version of the total expectation theorem holds:

Theorem 5.2.8. *Let X and Y be continuous random variables. Then*

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

Independence is defined exactly as in the discrete case.

Definition 5.2.9. Two continuous random variables X and Y are **independent** if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for each x, y.

Exactly as in the discrete case, if X and Y are independent, then the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. Indeed,

$$\begin{aligned}
\mathbb{P}(X \in A, Y \in B) &= \int_{X\in A}\int_{Y\in B} f_{X,Y}(x,y)\ \mathrm{d}y\mathrm{d}x \\
&= \int_{X\in A}\int_{Y\in B} f_X(x)f_Y(y)\ \mathrm{d}y\mathrm{d}x \\
&= \int_{X\in A} f_X(x)\int_{Y\in B} f_Y(y)\ \mathrm{d}y\mathrm{d}x \\
&= \Big(\int_{Y\in B} f_Y(y)\ \mathrm{d}y\Big)\Big(\int_{X\in A} f_X(x)\ \mathrm{d}x\Big) \\
&= \mathbb{P}(Y \in B)\mathbb{P}(X \in A).
\end{aligned}$$

The following properties, which were shown for discrete random variables, remain true in the continuous case.

Lemma 5.2.10. *Let X and Y be independent continuous random variables and let g and h be two functions such that $g(X)$ and $h(Y)$ are continuous. The following hold:*

- *$g(X)$ and $h(Y)$ are independent;*
- *$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$;*
- *$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$;*
- *$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$.*

Exercise 5.2.11. *A service station has a slow server (server 1) and a fast server (server 2). Upon arrival at the station, you are routed to server i with probability p_i, for $i \in \{1,2\}$, where $p_1 + p_2 = 1$. The service time at server i is an exponential random variable with parameter λ_i. What is the pdf of your service time at the station?*

5.3 Normal random variables

Definition 5.3.1. A continuous random variable X is **normal** if it has pdf given by

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

for some parameters μ, σ with $\sigma > 0$. We write $X \sim N(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma = 1$, X is called **standard normal**.

It can be shown that

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

and so $f_X(x)$ is a legitimate pdf. The parameter μ is the "center" of the density. Indeed, $f_X(x)$ is symmetric around μ i.e., $f_X(\mu + x) = f_X(\mu - x)$. The parameter σ is the "spread" of the density. The graph of $f_X(x)$ has a characteristic bell shape symmetric around the line $x = \mu$.

The importance of the normal random variable is mainly due to the Central limit theorem. Loosely speaking, it asserts that the distribution of the sum of a large number of i.i.d. random variables is approximated by the normal distribution.

It turns out that the parameters μ and σ are nothing but the expectation and the standard deviation, respectively:

Lemma 5.3.2. *Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{E}(X) = \mu$ and $\operatorname{var}(X) = \sigma^2$.*

Theorem 5.3.3. *Normality is preserved by linear transformations. Namely, if $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$.*

Proof. We look for the pdf of Y and obtain it by differentiating the distribution function. Suppose that $a > 0$ (the case $a \le 0$ is similar).

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}\Big(X \le \frac{y-b}{a}\Big) = F_X\Big(\frac{y-b}{a}\Big).$$

By the Chain rule and recalling the pdf of X,

$$f_Y(y) = \frac{\mathrm{d}F_Y}{\mathrm{d}y}(y) = \frac{1}{a} \cdot f_X\Big(\frac{y-b}{a}\Big) = \frac{1}{a} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{((y-b)/a-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}},$$

as claimed. □

If X is normal, we have that

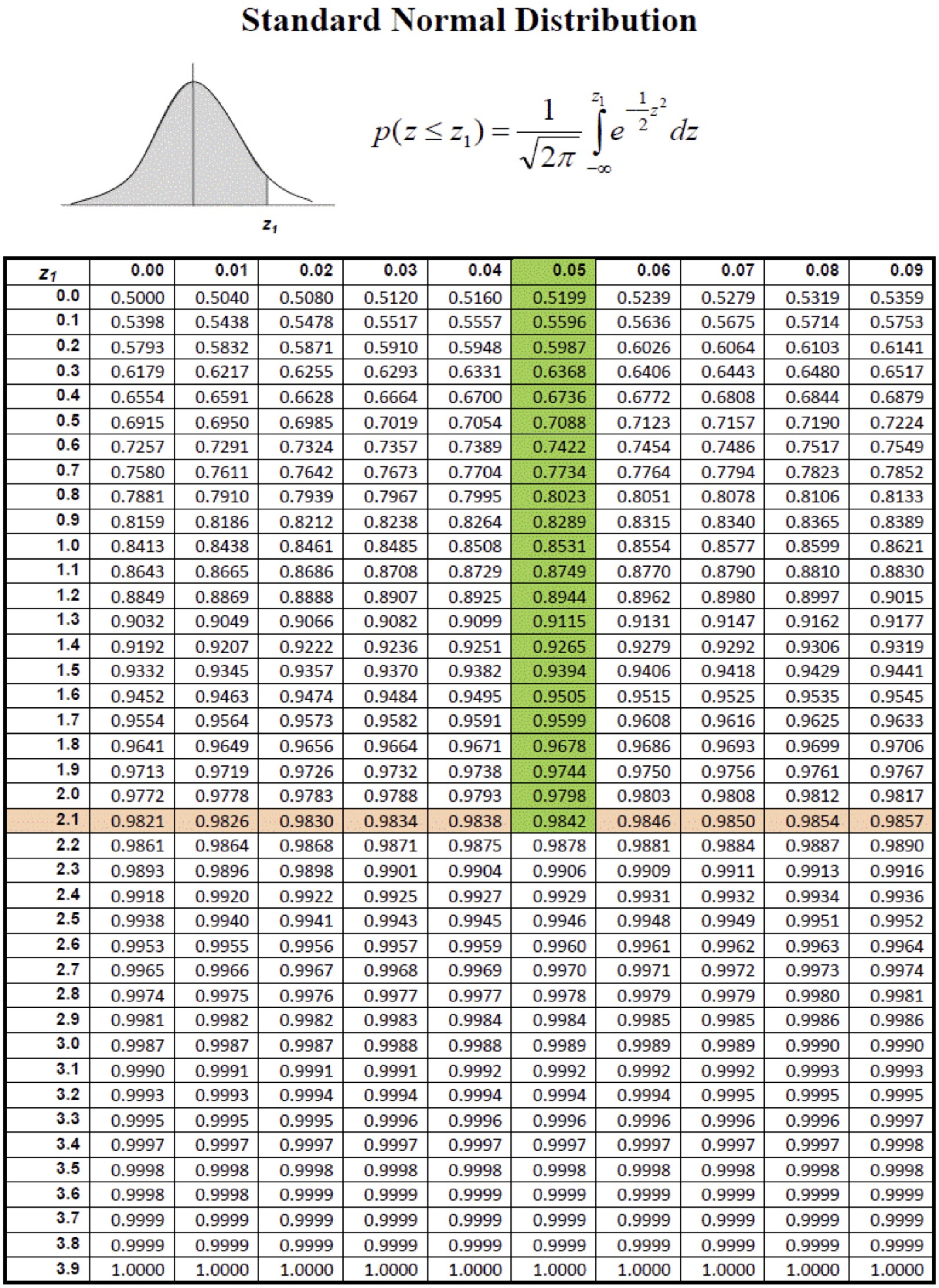
$$\mathbb{P}(X \le x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x.$$

Unfortunately, the function e^{-x^2} has no elementary antiderivative i.e., its antiderivative cannot be expressed as a sum, product, composition of finitely many polynomials, rational functions, trigonometric and exponential functions, and their inverse functions. On the other hand, in order to compute probabilities involving the normal random variable, we need to somehow compute the integral above. The lack of an elementary antiderivative is bypassed by computing approximations of the integral above, in the case $\mu = 0$ and $\sigma = 1$, via numerical integration. These approximated values are then stored in tables (see Figure 5.2) and allow to determine an approximate value of $\mathbb{P}(X \le x)$, for each x. Notice that the distribution function $F_X(x) = \mathbb{P}(X \le x)$ of a standard normal is usually denoted by $\Phi(x)$.

But in order to use these tables, how do we pass from a normal random variable X with parameters μ and σ to a standard normal? The answer is already in Theorem 5.3.3:

$$Y = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

We can then use this linear transformation and its inverse to jump from generic normal to standard normal and vice versa.



Standard Normal Distribution

$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2}\, dz$$

z_1	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Figure 5.2: Table storing values of $\Phi(z)$.

Example 5.3.4. The annual snowfall at a particular location is modelled as a normal random variable with mean $\mu = 60$ (in inches) and $\sigma = 20$. What is the probability that this year's snowfall will be at least 80 inches?

Let X be the snow accumulation. We need to compute $\mathbb{P}(X \geq 80) = 1 - \mathbb{P}(X \leq 80)$. To compute the latter, since we need to resort to the tables, we first pass to the standard normal random variable $Y = \frac{X-60}{20}$. We have that $\mathbb{P}(X \leq 80) = \mathbb{P}(20Y + 60 \leq 80) = \mathbb{P}(Y \leq 1)$. We then check the approximate value of $\mathbb{P}(Y \leq 1)$ in the tables: it is 0.8413 (see Figure 5.2).

Example 5.3.5. A binary message is transmitted as a signal which is either -1 or 1. The communication channel corrupts the transmission with additive normal noise with mean $\mu = 0$ and variance σ^2. The receiver concludes that the signal -1 (or 1) was transmitted if the value received is smaller than 0 (or at least 0). What is the probability of an error?

Let N be the noise and S be the signal. We have an error if -1 is transmitted and $N \geq 1$ (as this gives $N + S \geq 0$) or if 1 is transmitted and $N < -1$ (as this gives $N + S < 0$). We want to compute $\mathbb{P}(N \geq 1)$ and $\mathbb{P}(N < -1)$. As N is normal with $\mu = 0$, we know that these two values are the same. We then pass to the standard normal $N' = \frac{N-\mu}{\sigma} = \frac{N}{\sigma}$ and compute

$$\mathbb{P}(N \geq 1) = 1 - \mathbb{P}(N < 1) = 1 - \mathbb{P}(\sigma N' < 1) = 1 - \mathbb{P}(N' < 1/\sigma).$$

Exercise 5.3.6. *Let $X \sim N(5, 16)$.*

1. *Compute $\mathbb{P}(X > 3)$;*
2. *Find the value of c such that $\mathbb{P}(|X - 5| < c) = 0.9$.*

Exercise 5.3.7. *Prove Lemma 5.3.2.*

5.4 Central limit theorem

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2 (not necessarily normal random variables) and let $S_n = X_1 + \cdots + X_n$. The Weak law of large numbers tells us that the distribution of S_n/n concentrates around its mean μ as n becomes large. The Central limit theorem goes further and quantifies the behavior of the "fluctuations" of S_n around its mean $n\mu$.

Consider the following rescaling of $S_n - \mathbb{E}(S_n)$:

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{var}(S_n)}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Theorem 5.4.1 (Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2 and let*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then $\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z)$ for any $z \in \mathbb{R}$, where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x.$$

The Central limit theorem has several important consequences:

> It first tells us that the "fluctuations" of S_n around its mean $n\mu$ are of order $\sqrt{n}$. Moreover, the behavior of these fluctuations is universal: no matter what the distribution of the X_i's is, the asymptotic distribution of the "fluctuations" is standard normal.

It also answers the question: How does S_n behave for large n?

For n large, probabilities of the form $\mathbb{P}(S_n \leq c)$ can be approximated as follows:

1. Compute mean $n\mu$ and variance $n\sigma^2$ of S_n.
2. Compute $z = \frac{c-n\mu}{\sigma\sqrt{n}}$.
3. Use the approximation $\mathbb{P}(S_n \leq c) \approx \Phi(z)$, where the value of $\Phi(z)$ can be found from tables.

Let us justify these steps. For large n, the Central limit theorem implies that

$$\Phi(z) \approx \mathbb{P}(Z_n \leq z) = \mathbb{P}\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\Big) = \mathbb{P}(S_n \leq z\sigma\sqrt{n} + n\mu).$$

Therefore, letting z as in 2., we obtain the approximation in 3.

Example 5.4.2. The number of students X who are going to fail in the exam is a Poisson random variable with mean 100. What is the probability that at least 120 students will fail?

Since X is Poisson with mean 100 (and so $\lambda = 100$), we know that the desired probability is

$$\mathbb{P}(X \geq 120) = 1 - \mathbb{P}(X \leq 119) = 1 - \sum_{k=0}^{119} \frac{e^{-100}100^k}{k!}.$$

If we are just interested in an approximate value of this complicated sum, we can use the procedure described above. We can express X as a sum of 100 independent Poisson random variables $X_1, \ldots, X_{100}$, each with mean 1 and variance 1 (see Example 4.5.3). Checking Figure 5.2, we then have that

$$\mathbb{P}(X \leq 119) \approx \Phi\Big(\frac{119 - 100 \cdot 1}{1 \cdot \sqrt{100}}\Big) = \Phi(1.9) = 0.9713.$$

Example 5.4.3. We load on a plane 100 packages whose weights are independent random variables uniformly distributed between 5kg and 50kg. What is the probability that the total weight will exceed 3000kg?

Let S_{100} be the sum of weights of the 100 packages. We compute an approximate value for the desired probability $\mathbb{P}(S_{100} > 3000)$ by following the procedure above. We first need mean and variance of a uniform random variable on $[5, 50]$: $\mu = 27.5$ and $\sigma^2 = 168.75$. Letting $z = \frac{3000-100\cdot 27.5}{\sqrt{168.75\cdot 100}} = 1.92$, we get $\mathbb{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726$.

Example 5.4.4. A machine processes parts one at a time. The processing times of different parts are independent random variables uniformly distributed on $[1, 5]$. Find an approximate value for the probability that the number of parts processed within 320 time units is at least 100.

Let N_{320} be the random number of interest. We need to approximate $\mathbb{P}(N_{320} \geq 100)$. It is not clear how to express N_{320} as a sum of i.i.d. random variables. On the other hand, let X_i be the processing time of the i-th part and let $S_{100} = X_1 + \cdots + X_{100}$. The events $\{N_{320} \geq 100\}$ and $\{S_{100} \leq 320\}$ are the same and so we can apply our approximation procedure to S_{100}. We have that $\mathbb{E}(X_i) = 3$ and $\mathrm{var}(X_i) = 4/3$. Therefore, $z = \frac{320-100\cdot 3}{\sqrt{100\cdot 4/3}} = 1.73$ and so $\mathbb{P}(N_{320} \geq 100) = \mathbb{P}(S_{100} \leq 320) \approx \Phi(1.73) = 0.9582$.

Exercise 5.4.5. *A local telephone exchange with* 2000 *subscribers is to be connected to a central exchange by trunk lines. Suppose that during the busy period each subscriber requires a trunk line for an average of* 2 *minutes per hour. How many lines should be installed to ensure that at least* 99% *of calls find an idle trunk line?*

Bibliography

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2nd edition, 2008.

[2] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 3rd edition, 2001.

[3] G. Grimmett and D. Welsh. *Probability - An Introduction*. Oxford University Press, 2nd edition, 2014.

[4] M. Mitzenmacher and E. Upfal. *Probability and Computing - Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[5] J.R. Norris. Probability. http://www.statslab.cam.ac.uk/~james/Lectures/p.pdf, 2017.

[6] D. Stirzaker. *Elementary Probability*. Cambridge University Press, 2nd edition, 2003.

[7] S.S. Venkatesh. *The Theory of Probability*. Cambridge University Press, 2013.

[8] J.B. Walsh. *Knowing the Odds - An Introduction to Probability*. AMS, 2012.