Applied Probability for Computer Science

Federico Segala

Academic Year: 2025-2026

Lecture Notes

prof. Isadora Antoniano Villalobos

Table of Contents

1. Overview of Elementary Probability

This chapter is focused on providing a concise overview of some fundamental concepts in probability theory that will be essential for understanding more advanced topics that will be presented in the following chapters.

1.1. Sample Space and Events

This section is focused on providing a clear understanding of the concepts of **sample space** and **events**, which are foundational elements in probability theory. It will first introduce the definitions and then provide some examples to better understand these concepts.

Definition 1.1 (Sample Space and Events)

A collection of all elementary results, or **outcomes** of an experiment, is called a **sample space**. Any set of *outcomes* from the sample space is called an **event**. In other words, we can state that an event can be viewed as an **arbitrary subset** of the sample space. ♣

Following we introduce some common *notation* that will be used throughout the notes for referring to specific concepts:

- Ω, S are typically used to denote the **sample space**.
- $\emptyset$ is typically used to denote the empty set or event
- A, B, E and other *capital letters* are used for events
- ω, s are going to be used for **individual outcomes**
- we will use the notation $\mathbb{P}[E]$ or $P[E]$ to denote the **probability** of an event

Remark

One important aspect to consider about the empty set $\emptyset$ is that it belongs to any sample space, i.e., $\emptyset \in \Omega \ \forall \Omega$.

Remark

Regarding events and individual outcomes it is important to remember that $E \subset \Omega$ and that, to distinguish between an event with a single outcome and the outcome itself we have respectively $\omega \in \Omega$ and $\{\omega\} \subset \Omega$.

Example: Simple Die Toss

Suppose a tossed die can produce one of 6 possible outcomes: 1 dot, through 6 dots. Each outcome is an **event**. Anyway there are other possible events, such as 'observing an even number of dots', a number of dots which is less than 4, an so on.

1.1.1. Some examples of Sample Spaces

Exercise. Two identical birds are initially on two nearby trees (A and B respectively). At random intervals, a sudden noise frightens one of the birds, making it fly to the other tree. At

each event, each bird has the same probability of being the one frightened and changing tree. Find the **sample space** if *only the number of birds on tree A* is considered.

Solution. Since the problem asks for the sample space in terms of number of birds on a specific tree, i.e., A. Tree A can have either 0, 1, or 2 birds; Therefore the sample space is $\Omega = \{0, 1, 2\}$.

> **Remark**
>
> It is important to notice that we are only interested in the number of birds on a specific tree; therefore the sample space will be independent on other birds features, such as their color, size, and so on, so forth.

Exercise. Let's make things more complicated, if we were to consider both the number of birds on tree A and on tree B, what would be the sample space?

Solution. Each event in this case, can be represented as a **tuple** of the form (x, y), where x is the number of birds on tree A and y is the number of birds on tree B. Therefore the possible events are: $\Omega = \{(0, 2), (1, 1), (0, 2)\}$.

Taking into account the assignment of the previous exercises, we can notice that there are two important aspects to consider:

- the sudden noise occurs at **random intervals**
- the happening of the noise can be considered an **event**. At **each event**, each bird has the same probability of moving

Clearly, the concept we have failed to represent so far is **time**. To be more precise, we can write the sample spaces prior to any noise event as respectively: $\Omega_0 = \{0, 1, 2\}$ and $\Omega_0 = \{(0, 2), (1, 1), (2, 0)\}$.

If we wanted to describe the **full experiment** we would need to describe an **infinite sequence** of states. This is what's called a **random sequence**. We can briefly describe it as in .

$$\Omega = \{(x_0, x_1, x_2, ...) \mid x_i \in \Omega_0 \ \forall i \in \mathbb{N}\} \tag{1}$$

> **Tip**
>
> Before starting to solve any kind of probability problem, also during the exam, always start by **defining** the **sample space**. This is a fundamental step that will help at better understanding the problem and avoid mistakes throughout the solution process.

Once we have defined the sample space, we can start defining **events**. For example, suppose the event A stands to represent the case when "initially both birds are on tree B": $A = \{(0, 2)\}$. Provided we have this information, shows the sample space for the presence on **tree A** conditioned on the fact that the starting position is the one described by A.

$$\Omega_A = \{0, 1, x_3, 1, x_4, ... \mid x_i \in \{0, 2\}\} \subset \Omega \tag{2}$$

1.1.2. Discrete vs Continuous Sample Spaces

One additional important aspect to consider is the **countability** of the sample space. Indeed we may find ourselves working with either **discrete** or **continuous** sample spaces. A discrete sample space is one that is either *finite* or *countably infinite*, meaning that its elements can be put into a one-to-one correspondence with the natural numbers. On the other hand, a continuous sample space contains an uncountably infinite number of outcomes, often represented by intervals of real numbers.

For instance, the time between consecutive noise events in the previous example could be modeled by means of a *continuous sample space*: $\Omega = \{t_0, t_1, t_2, \ldots \mid t_j \in \mathbb{R}^+\}$.

Instead of recording the birds' positions only after each noise event, we could decide to record their positions **continuously over time**. In this case the sample space would become:

$$\Omega = \left\{(X_t)_{t\geq 0} : x_t \in \{0, 1, 2\}\right\},$$

where at any given point in time t, X_t represents the number of birds on tree A.

1.2. Basic Set Operations for Events

This section is focused on introducing some basic set operations that can be performed on events. Since events are subsets of the sample space, we can apply standard set operations such as union, intersection, and complement to them.

Union of Events

The **union** of two events A and B, denoted by $A \cup B$, represents the event that either event A occurs, event B occurs, or both events occur. In other words, the union of events includes all outcomes that are in either event.

Intersection of Events

The **intersection** of two events A and B, denoted by $A \cap B$, represents the event that both event A and event B occur simultaneously. The intersection of events includes only those outcomes that are common to both events.

Complement of an Event

The **complement** of an event A, denoted by $\overline{A}$, represents the event that event A does not occur. The complement of an event includes all outcomes that are in the sample space but not in event A.

Definition 1.2 (Set Difference)

Given two sets A and B, the **set difference** of A and B, denoted by $A \setminus B$, is defined as the set of elements that are in A but not in B. Formally, it can be expressed as:

$$A \setminus B = A \cap \overline{B}$$

De Morgan's Laws

De Morgan's Laws provide a relationship between the union and intersection of sets through complementation.

Theorem 1.1 (The Morgan 1st Law)

Given sets $E_1, ..., E_n$, the complement of their union is equal to the intersection of their complements:

$$\overline{E_1 \cup E_2 \cup ... \cup E_n} = \overline{E}_1 \cap \overline{E}_2 \cap ... \cap \overline{E_n}$$

♡

Theorem 1.2 (The Morgan 2nd Law)

Given sets $E_1, ..., E_n$, the complement of their intersection is equal to the union of their complements:

$$\overline{E_1 \cap E_2 \cap ... \cap E_n} = \overline{E}_1 \cup \overline{E}_2 \cup ... \cup \overline{E_n}$$

♡

Disjoint and Exhaustive Events

There are two important concepts related to events that are worth mentioning: **disjoint** and **exhaustive** events.

Definition 1.3 (Disjoint Events)

Events A and B are **disjoint** if their intersection is empty:

$$A \cap B = \text{emptyset}$$

Events $E_1, E_2, ...$ are **mutually exclusive** or **pairwise disjoint** if any two of these events are disjoint:

$$E_i \cap E_j = \emptyset \ \forall i \neq j$$

♣

Definition 1.4 (Exhaustive Events)

A collection of events $E_1, E_2, ...$ is said to be **exhaustive** if their union covers the entire sample space:

$$E_1 \cup E_2 \cup ... = \Omega$$

♣

Partitions of the Sample Space

Defining mutual exclusivity and exhaustivity allows us to introduce the concept of **partitions** of the sample space. This is a useful concept when dealing with events that cover the entire sample space without overlapping.

We say that a collection of **mutually exclusive** and **exhaustive** events $E_1, E_2, ...$ forms a **partition** of the sample space Ω. This concept is clearly shown in

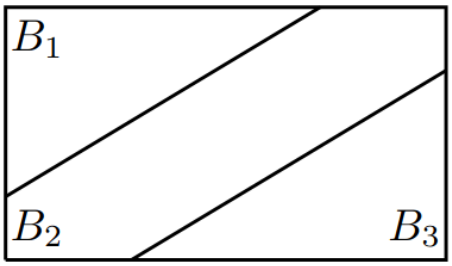


Figure 1.1: Example of a partition of the sample space

> **Remark**
>
> Any event $A \subset \Omega$ can be written in terms of the union of its intersections with the elements of the partition. This is illustrated in :

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \tag{3}$$

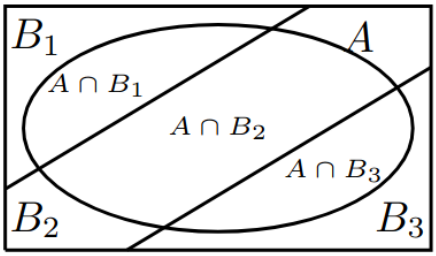


Figure 1.2: Event represented in terms of a partition of the sample space

1.3. Definition of Probability

In this section we are going to try to provide a formal definition for the notion of **probability**. First of all the reason why we introduced sample spaces and events in the first place, is that events are the entities for which we can compute *probabilities*.

In a very coarse way, we can think of probability as a **measure** that assigns to each event a number between 0 and 1, representing the **likelihood** of that event occurring. In a more technical language, a probability is a **function** which maps each event from the sample space to a real number in the interval $[0, 1]$.

A little Digression

In mathematics, a function che be described as the following:

$$f(x) = \sin(x)$$
$$f : \mathbb{R} \to [-1, 1]$$

We can see that we need to define both the **domain** and the **codomain** of the function along with the **rule** that describes how to map elements from the domain to elements in the codomain. Since probability is a function we'll need to do the same for it.

Sigma Algebra

Now that we have seen how a probability can be seen as a function, to formally define it, we need to specify its domain, codomain, and the mapping rule. This will give us a solid foundation to work with probabilities in a rigorous manner.

As per the **codomain**, we have already mentioned that probabilities are real numbers in the interval $[0, 1] \in \mathbb{R}$. The matter becomes a little bit more complicated when it comes to defining the **domain**. We can imagine the domain of the probability function as a *collection of events* with some specific properties. This collection is called a **sigma-algebra** (or **σ-algebra**). Typically, a sigma-algebra is denoted by the symbol $\mathfrak{M}$.

Definition 1.5 (Sigma Algebra)

A collection $\mathfrak{M}$ of events is a **σ - algebra** on a sample space Ω if:

- it includes the sample space:

$$\Omega \in \mathfrak{M}$$

- every event in $\mathfrak{M}$ is contained along with its complement:

$$E \in \mathfrak{M} \Rightarrow \overline{E} \in \mathfrak{M}$$

- every finite or countable collection of events in $\mathfrak{M}$ is contained along with its union:

$$E_1, E_2, \ldots \in \mathfrak{M} \Rightarrow E_1 \cup E_2 \cup \ldots \in \mathfrak{M}$$

♣

We can notice the following important aspects about sigma-algebras:

- $\mathfrak{M} = \{\emptyset, \Omega\}$ is the smallest possible sigma-algebra on Ω, called **degenerate**
- $\mathfrak{M} = 2^{\Omega} = \{E : E \subset \Omega\}$ is the largest possible sigma-algebra on Ω, called **power set**

Remark

When $\Omega \subseteq \mathbb{N}$, is **countable** the most common choice for the associated sigma algebra is the **power set** $\mathfrak{M} = 2^{\Omega}$.

On the other hand, when dealing with **uncountable** $\Omega \subseteq \mathbb{R}$, the power set *too large* to be useful. In this case a common choice for the sigma-algebra is the **Borel Sigma Algebra**, denoted by $\mathcal{B}$, which contains all possible sets that one could practically think about except for everything that could get created by some strange recursive process that resembles the construction of *fractals*.

Axiomatic Definition of Probability

Now that we have defined the sigma-algebra, we can finally provide a formal definition of probability.

Definition 1.6 (Probability)

Assume a sample space Ω and a sigma-algebra of events $\mathfrak{M}$ defined on it. **Probability**

$$\mathbb{P} \rightarrow [0, 1] \tag{1.4}$$

is a function of events with the domain $\mathfrak{M}$ and the range $[0, 1]$ that satisfies the following conditions (which are called the **axioms of probability**):

- **Unit Measure**: the sample space has unit probability: $\mathbb{P}[\Omega] = 1$
- **Sigma-Additivity**: for any finite or countable collection of mutually exclusive events $E_1, E_2, \ldots \in \mathfrak{M}$, the probability of their union is equal to the sum of their individual probabilities:

$$\mathbb{P}[E_1 \cup E_2 \cup \ldots] = \mathbb{P}[E_1] + \mathbb{P}[E_2] + \ldots$$

♣

It is good to notice that, from the first properties, we can derive that the computing the probability of the sample space amounts to say: '*something happened*'. The second property becomes fundamental when dealing with events that can be broken down into simpler, mutually exclusive events, allowing us to compute their probabilities more easily.

All rules of probability are a direct consequence of . This will allow us to compute probabilities for all events in our interest. Following we outline some of the most important probability rules that will be useful in the next chapters.

- $\mathbb{P}[\emptyset] = 0$: this is easy to verify; indeed the we know from the axioms that $\mathbb{P}[\Omega] = 1$. From the second axiom we know that the union of any disjoint event has probability equal to their sum, that is $\mathbb{P}(\Omega) + \mathbb{P}[\emptyset] = 1 \Rightarrow \mathbb{P}[\emptyset] = 0$
- $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$; we can actually notice that the following relation holds: $\mathbb{P}[A \cup B] = \mathbb{P}\left[A \cap \overline{B}\right] + \mathbb{P}\left[B \cap \overline{A}\right] + \mathbb{P}[A \cap B]$.

We can see how the second formulation is supported by the second axiom in that the three members of the summation form a **partition** of the event $A \cup B$. The reason why the first formulation is commonly preferred, is that that, if A and B are **independent** the probability of their intersection is given by $\mathbb{P}[A]\mathbb{P}[B]$.

Intuitively, saying that two experiments are independent means that the outcome of one does not affect the outcome of the other one. To be more formal we should also consider that the occurrence of an event doesn't even influence the events' probability in the other experiment.

We would even like to be more formal about the definition of independency but it's not possible without first introducing the concept of **conditional probability**.

Inclusion - Exclusion Principle

There are many cases in which we may be interested in computing the probability (or the amount of elements) of the union of multiple events. We already know how to do that in case of two events. In case we have three or more events we can generalize through the **Inclusion-Exclusion Principle**.We know that in the two event case we have:

$$\text{count}(A_1 \cup A_2) = \text{count}(A_1) + \text{count}(A_2) - \text{count}(A_1 \cap A_2)$$

Suppose now that we have three non-disjoint events A_1, A_2, A_3 as represented in .

The number elements in the union in such case would be given by:

$$\text{count}\left(\bigcup_{i=1}^{3} A_i\right) = \sum_{i=1}^{3} \text{count}(A_i) - \sum_{i<j} \text{count}\big(A_i \cap A_j\big) + \text{count}\left(\bigcap_{i=1}^{3} A_i\right)$$

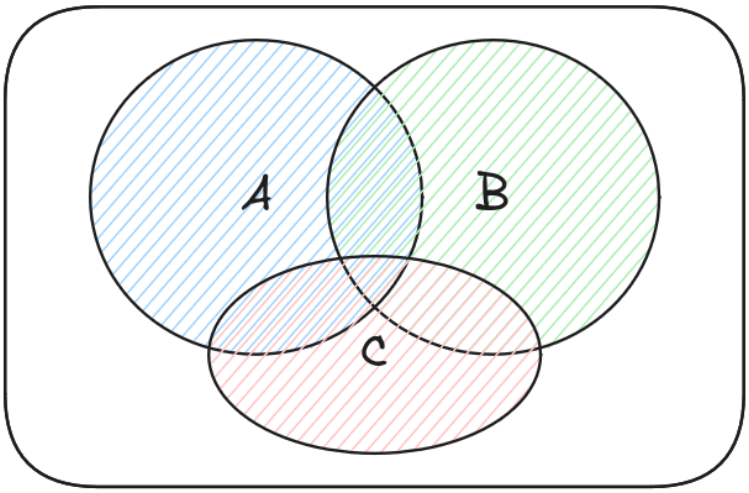


Figure 1.3: Three non-disjoint events

We are not going to see the full generalized version of the formula, mainly because it is so ugly. Anyway as a general principle we can say that to compute the count of the union multiple events:

- we first sum the counts of each individual event
- we subtract the counts ef each pairwise union (even number of events)
- we add back the counts of each triple-wise union (odd number of events)
- we keep on alternating between subtracting the count of "even-unions" and adding the count of "odd-unions" until we reach the union of all events

1.4. Conditional Probability

After defining the basic notion of probability, it's time to move on to a slightly more powerful concept, which is the one of **conditional probability**. The following definition gives us a way to compute conditional probabilities of two events.

Definition 1.7 (Conditional Probability)

Given two events A and B we can define the conditional probability of one event *given* that the other one has occurred as follows:

$$\mathbb{P}[\![A|B]\!] = \frac{\mathbb{P}[\![A \cap B]\!]}{\mathbb{P}[\![B]\!]} \tag{1.5}$$

♣

To understand the meaning of we can consider the following example.

Example: Computation of conditional probability

Suppose we throw a die 2 times and observe that the sum is 7. We can assert the following:

- the probability that the event $(6, 6)$ happened is 0, since it is impossible to obtain a 7 with $(6, 6)$
- the probability the rolling $(3, 4)$ can be computed even without looking at . Indeed if we know that the sum of the tosses is 7, we can manually compute the sample space: $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$. By simply noticing that the event $(3, 4)$ is present only once in the sample space we conclude that its probability is given by $\frac{1}{6}$.

Formula Derivation

The reason behind the formulation of is the following: before an experiment the sample space Ω is the set of all possible experiments outcomes. Due to the fact that we are considering a probability we are dealing with a sigma algebra $\mathfrak{M}$, we can make safely state the following:

$$A, B \in \mathfrak{M} \Rightarrow A \cap B \in \mathfrak{M} \Rightarrow \mathbb{P}[\![A]\!], \mathbb{P}[\![B]\!], \mathbb{P}[\![A \cap B]\!] \text{ are known}$$

or at least can be computed in some way. We can say that $\mathbb{P} : \mathfrak{M} \to [0, 1]$ is a **prior probability**. If now we perform the experiment and know that B happens we can update the probability to incorporate the new knowledge by computing a new $\mathbb{P} \mid B : \mathfrak{M} \to [0, 1]$, a **posterior probability**. In practical terms, if B happened, $\overline{B}$ becomes impossible and we can **restrict our sample space** to only those outcomes in which B happens. Therefore the new sample space becomes $\Omega_B = \{\omega \in \Omega \mid \omega \in B\}$.

Let's understand the reason why we need to divide by $\mathbb{P}[\![B]\!]$ in :

- suppose we are interested in the whole sample space Ω
- prior to knowing that B happened, the probability of Ω is obviously 1: $\mathbb{P}[\![\Omega]\!] = 1$
- after knowing that B happened, the new sample space becomes $\Omega_B = \Omega \cap B = B$

- since the new sample space is $\Omega \cap B$, its probability must still be 1: $\mathbb{P}[\![\Omega \cap B]\!] = 1$, but since $\Omega \cap B = B$ we have that $\mathbb{P}[\![B \mid B]\!] = 1$
- since we want to make sure that $\mathbb{P}[\![B \mid B]\!] = 1$ we can compute $\frac{\mathbb{P}[\![\Omega \cap B]\!]}{\mathbb{P}[\![B]\!]} = 1$, in order to normalize the probability to 1

We can now focus on the single posterior probability of the event A given that B happened following a similar reasoning:

- for any event A prior to knowing that B happened, its probability is given by $\mathbb{P}[\![A]\!]$
- after knowing that B happened, the new sample space becomes $\Omega_B = \Omega \cap B = B$
- since we need to compute the probability of A in the new sample space, we need to restrict A to only those outcomes in which B happens, therefore the new event becomes $A \cap B$. Thus the probability of A in the new sample space becomes $\mathbb{P}[\![A \cap B]\!]$
- to scale everything back to the previous sample space we need to divide by $\mathbb{P}[\![B]\!]$.

The previous reasoning is clearly represented in the Venn diagram shown in .

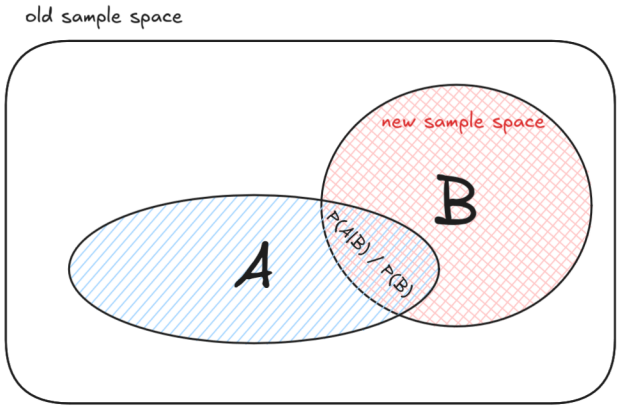


Figure 1.4: Venn diagram representation of conditional probability

We are now ready to provide a formal definition of independency between two events.

Definition 1.8 (Independent Events)

Given two events A and B, we say that they are independent if by knowing that one event has changed the probability of the other event remains the same. Formally, this can be expressed as:

$$\mathbb{P}[\![A \mid B]\!] = \frac{\mathbb{P}[\![A \cap B]\!]}{\mathbb{P}[\![B]\!]} = \mathbb{P}[\![A]\!] \tag{1.6}$$

♣

⚠ Warning

Although the notions of independency and disjointedness may seem similar at a first glance, they are actually the **opposite**. Indeed if two events are disjoint, the occurrence of one

event implies that the other event cannot occur, which means that knowing one event has occurred changes the probability of the other event to zero. Therefore, disjoint events are not independent.

From the previous definition we can derive the following important relation for independent events.

Theorem 1.3

Given two independent events A, B we can look at their definition in order to derive a different formulation for their intersection:

$$\mathbb{P}[\![A \cap B]\!] = \mathbb{P}[\![A]\!]\mathbb{P}[\![B]\!] \tag{1.7}$$

1.5. Conditional Probability and Bayes' theorem

This section will introduce one of the most important results in all probability theory. Before doing so, we show a couple of examples.

1.5.1. Exercise - Crashes in a Computer Program (Baron 2.35)

Problem Statement

A new computer program consists of two modules. The first module contains an error with probability 0.2. The second module is more complex and has a probability of 0.4 of containing an error, *independently of the first one.*

An error in the first module alone causes the program to crash with probability 0.5. For the second module, an error causes a crash with probability 0.8. If there are errors in both modules the probability of a crash rises to 0.9. Suppose that the program has crashed. What is the probability of error in both modules? ∎

Definition of the Sample Space and Sigma Algebra

Let's first try to map all the information provided into terms we are already familiar with. The **experiment** basically consists in running the program. The **sample space Ω** should contain all results of this experiment, that is, the program itself and the result of executing it. We can notice how this. Since the experiment is very complicated it makes sense to try and focus only on relevant outcomes. Let's list all the relevant events that must be included in our **sigma algebra**:

- E_1: module 1 contains an error$\rightarrow \overline{E}_1 \in \mathfrak{M}$ (by σ-algebra properties)
- E_2: module 2 contains an error $\rightarrow \overline{E}_2 \in \mathfrak{M}$ (by σ-algebra properties)
- Ω: any event happens $\rightarrow \emptyset \in \mathfrak{M}$ (by σ-algebra properties)
- C: program crashes $\rightarrow \overline{C} \in \mathfrak{M}$ (by σ-algebra properties)

By the properties of sigma algebras we can also state the following events must be included in $\mathfrak{M}$:

- $E_1 \cup E_2 \in \mathfrak{M}$, since sigma algebras are closed under union
- $\overline{E_1 \cup E_2} \in \mathfrak{M}$, since sigma algebras are closed under complement
- $\overline{E}_1 \cap \overline{E}_2 \in \mathfrak{M}$, by applying De Morgan laws

The same rational can be applied starting from the complement events. Therefore the sigma algebra will also contain $E_1 \cap E_2$. Notice that by simply considering the events E_1, E_2, C we can already keep track of all the possible events we may be interested in ($C \cap E_1, C \cap E_2, \ldots$).

Information Extraction

Let's now try to extract valuable information from the problem statement:

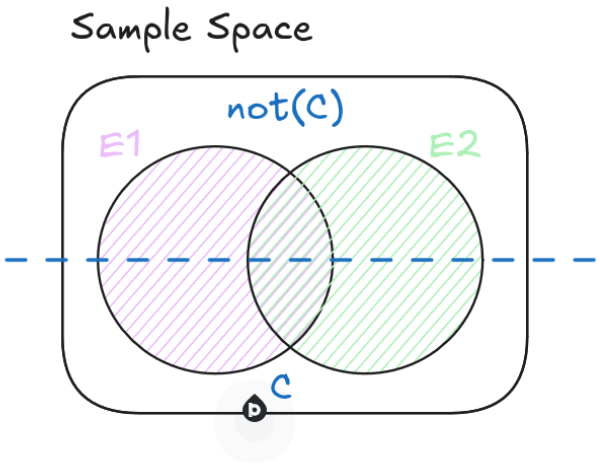
- $\mathbb{P}[\![E_1]\!] = 0.2$
- $\mathbb{P}[\![E_2]\!] = 0.4$

Since $E_1 \perp E_2$ we can use to compute the probability of their intersection: $\mathbb{P}[\![E_1 \cap E_2]\!] = \mathbb{P}[\![E_1]\!]\mathbb{P}[\![E_2]\!] = 0.08$. The problem also provides us with the following information:

- $\mathbb{P}[\![C \mid E_1 \cap \overline{E}_2]\!] = 0.5$: crash given an error on first module alone
- $\mathbb{P}[\![C \mid \overline{E}_1 \cap E_2]\!] = 0.8$: crash given an error on second module alone
- $\mathbb{P}[\![C \mid E_1 \cap E_2]\!] = 0.9$: crash given an error on both modules

Information Organization

Now that we have extracted all the relevant information from the problem statement, we can try to organize it in a more structured way. The main approach in order to compute probability is *divide and conquer*. So what we do is trying to divide our events in the smallest possible pieces, that are mutually exclusive and that cover the whole space. This amounts to finding **relevant partitions**. Consider the following picture:



One partition we can consider is the one given by the events E_1, E_2 combined and their complements:

- $B_1 = \overline{E}_1 \cap \overline{E}_2 \leftrightarrow \overline{E_1 \cup E_2}$: this represents the case in which no module has an error
- $B_2 = E_1 \cap \overline{E}_2$: this represents the case in which only the first module has an error

- $B_3 = \overline{E}_1 \cap E_2$: this represents the case in which only the second module has an error
- $B_4 = E_1 \cap E_2$: this represents the case in which both modules have an error

We can easily notice that these events (B_1, B_2, B_3, B_4) are mutually exclusive and span the whole sample space. Therefore they form a partition of the sample space. Another partition we can consider is the one given by the crash event and its complement: $\left(C, \overline{C}\right)$.

In order to divide and conquer, we can **identify** the **partition** for which we have **prior probabilities** (probabilities which are not conditioned on other events) and the ones for which we need to compute **conditional probabilities**. This step allows us to *find a logical temporal order of events*. In our case we can notice that we have prior probabilities for the events in the partition (B_1, B_2, B_3, B_4) and conditional probabilities for the events in the partition $\left(C, \overline{C}\right)$.

Now we can **compute** the **available probabilities** from the basic information we have extracted. First we can compute the probability of the *intersection* $B_4 = E_1 \cap E_2$:

$$\mathbb{P}[\![B_4]\!] = \mathbb{P}[\![E_1 \cap E_2]\!] = \mathbb{P}[\![E_1]\!]\mathbb{P}[\![E_2]\!] = 0.08$$

Now we would like to compute the probability of B_1, for doing so we need the probability of the union, that we can compute as follows:

$$\mathbb{P}[\![E_1 \cup E_2]\!] = \mathbb{P}[\![E_1]\!] + \mathbb{P}[\![E_2]\!] - \mathbb{P}[\![E_1 \cap E_2]\!] = 0.2 + 0.4 - 0.08 = 0.52$$

Thus, the probability of the event B_1 becomes: $\mathbb{P}[\![B_1]\!] = 1 - \mathbb{P}[\![E_1 \cup E_2]\!] = 1 - 0.52 = 0.48$.

To compute the probability of B_2 we need to compute the following:

- $\mathbb{P}[\![B_2]\!] = \mathbb{P}\left[\!\left[E_1 \cap \overline{E}_2\right]\!\right]$, since $E_1 \perp E_2$, then also $E_1 \perp \overline{E}_2$, therefore we can apply and compute: $\mathbb{P}\left[\!\left[E_1 \cap \overline{E}_2\right]\!\right] = \mathbb{P}[\![E_1]\!]\mathbb{P}\left[\!\left[\overline{E}_2\right]\!\right] = 0.2(1 - 0.4) = 0.12$
- the same reasoning can be applied to compute B_3, for completeness though, we can also compute both B_2, B_3 as follows (B_3 case):

$$\mathbb{P}[\![B_3]\!] = \mathbb{P}[\![E_2]\!] - \mathbb{P}[\![E_1 \cap E_2]\!] = 0.4 - 0.08 = 0.32$$

> **Remark**
>
> Since $(B_1 ... B_4)$ form a partition we can verify that the sum of their probabilities is equal to 1: $\mathbb{P}[\![B_1]\!] + \mathbb{P}[\![B_2]\!] + \mathbb{P}[\![B_3]\!] + \mathbb{P}[\![B_4]\!] = 0.48 + 0.12 + 0.32 + 0.08 = 1$.

Assuming that if we have no crashes the program works fine, we can produce the following diagram that summarizes all the information we have gathered so far. Tha's illustrate .

Now that we have all the information we have gathered we are finally ready to give a solution. Since we are asked to compute the probability of having errors in both modules given that the program has crashed. We can rewrite it in terms of our events as follows:

$$\mathbb{P}[\![E_1 \cap E_2 \mid C]\!] = \mathbb{P}[\![B_4 \mid C]\!]$$

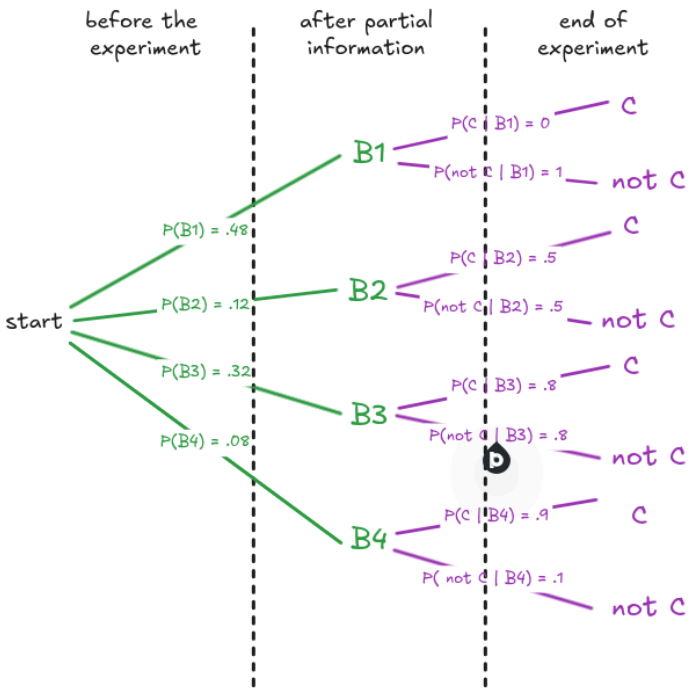


Figure 1.5: Summary of the information gathered from the problem statement and from the intermediate stages

We can retrieve the definition of conditional probability in and apply it to our case: $\mathbb{P}[\![B_4 \mid C]\!] = \frac{\mathbb{P}[\![B_4 \cap C]\!]}{\mathbb{P}[\![C]\!]}$. To do so, we need different components.

First of all we need to compute the probability of $B_4 \cap C$, this can be done by backtracking on the tree in :

$$\mathbb{P}[\![B_4 \cap C]\!] = \mathbb{P}[\![B_4]\!]\mathbb{P}[\![C \mid B_4]\!] = 0.08 \cdot 0.9 = 0.072$$

Now wee need to compute the probability of a crash, that is $\mathbb{P}[\![C]\!]$, this is a slightly more complicated matter. To do so, we need to take into consideration all the possible paths that lead from the root of the tree up to node C. Therefore we can write:

$$\begin{aligned}\mathbb{P}[\![C]\!] &= \mathbb{P}[\![B_1]\!]\mathbb{P}[\![C \mid B_1]\!] + \mathbb{P}[\![B_2]\!]\mathbb{P}[\![C \mid B_2]\!] + \mathbb{P}[\![B_3]\!]\mathbb{P}[\![C \mid B_3]\!] + \mathbb{P}[\![B_4]\!]\mathbb{P}[\![C \mid B_4]\!] \\ &= \sum_{i=1}^{4} \mathbb{P}[\![B_i]\!]\mathbb{P}[\![C \mid B_i]\!] \\ &= 0.48 \cdot 0 + 0.12 \cdot 0.5 + 0.32 \cdot 0.8 + 0.08 * 0.9 = 0.388\end{aligned}$$

Now that we have all the elements we need, we can finally compute the probability we were looking for:

$$\mathbb{P}[\![B_4 \mid C]\!] = \frac{\mathbb{P}[\![B_4 \cap C]\!]}{\mathbb{P}[\![C]\!]} = \frac{0.072}{0.388} \approx 0.1856 \ \blacksquare$$

1.5.2. Law of Total Probability

Even though we didn't explicitly mentioned, while resolving the last exercise we have actually applied probably what is one of the most important results in probability theory: the **law of total probability**.

Precisely, we did it when computing the probability of a crash $\mathbb{P}[\![C]\!]$. We can now formally state the law as follows.

Axiom 1.1 (Law of Total Probability)

Given a partition of the sample space $(B_1, B_2, ...)$ and an event A, the probability of A can be computed as follows:

$$\mathbb{P}[\![A]\!] = \sum_i \mathbb{P}[\![B_i]\!]\mathbb{P}[\![A \mid B_i]\!] \tag{1.8}$$

This is the reason why in the previous example we spent so much time trying to compute the partition (B_1, B_2, B_3, B_4). In case we have only two events we can rewrite it as follows:

$$\mathbb{P}[\![A]\!] = \mathbb{P}[\![A \mid B]\!]\mathbb{P}[\![B]\!] + \mathbb{P}[\![A \mid \overline{B}]\!]\mathbb{P}[\![\overline{B}]\!]$$

1.5.3. Bayes' Rule

Another important result that we can derive from the definition of conditional probability is **Bayes' Rule**. This rule allows us to *reverse* the conditioning of a probability. Formally we can state it as follows.

Theorem 1.4 (Bayes' Rule)

Given an event A and a partition of the sample space $B = (B_1, B_2, ..., B_k)$, the conditional probability of $B_i \in B$ given A can be computed as follows:

$$\mathbb{P}[\![B_i \mid A]\!] = \frac{\mathbb{P}[\![A \mid B_i]\!]\mathbb{P}[\![B_i]\!]}{\sum_{j=1}^{k} \mathbb{P}[\![A \mid B_j]\!]\mathbb{P}[\![B_j]\!]} \tag{1.9}$$

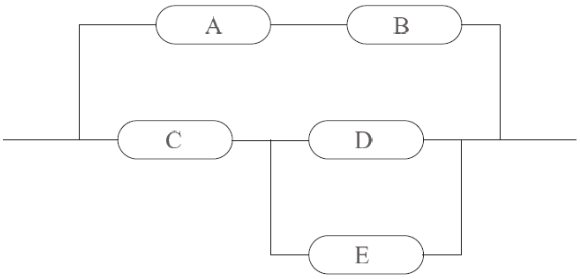
In particular, since B and $\overline{B}$ always form a partition of the sample space, we can rewrite for the case of two events as follows:

$$\mathbb{P}[\![B \mid A]\!] = \frac{\mathbb{P}[\![A \mid B]\!]\mathbb{P}[\![B]\!]}{\mathbb{P}[\![A \mid B]\!]\mathbb{P}[\![B]\!] + \mathbb{P}[\![A \mid \overline{B}]\!]\mathbb{P}[\![\overline{B}]\!]} \tag{10}$$

1.5.4. Exercise - Reliability of a System (Baron 2.20)

Problem Statement

Consider the following system of connected components.



Calculate the reliability of the following system if each component is operable with probability 0.92 independently of the other components ∎

Solution

To solve the proposed problem, we need to define the sample space. In order to be as fast as possible, we can assume it being partitioned by only two events: the system works (W) or the system fails ($\overline{W}$). To answer the question we need to compute $\mathbb{P}[\![W]\!]$, that is, the probability that the system works, which is another way to define *reliability*.

For the system to be operational it is necessary that a 'signal' is allowed to travel from the left side to the right side of the system. In order for this to happen we can see there are basically three possible paths:

- the upper part in which the components that need to function are A and B, we can call this situation B_1
- the middle part in which the components that need to function are C and either one of D and E (or both), we can call this situation B_2

So, we can say that in general the probability of full functionality of the system can be defined as $\mathbb{P}[\![B_1 \cup B_2]\!]$. Now that we have defined these events, we can try to compute their individual probabilities.

To compute $\mathbb{P}[\![B_1]\!]$ we need both components A and B to be operational. Since the components are independent we can apply and compute: $\mathbb{P}[\![B_1]\!] = \mathbb{P}[\![A]\!]\mathbb{P}[\![B]\!] = 0.92 \cdot 0.92 = 0.8464$.

For the lower part of the circuit, since we have a parallel connection between D and E, we can compute the probability of at least one of them working as follows: $\mathbb{P}[\![D \cup E]\!] = \mathbb{P}[\![D]\!] + \mathbb{P}[\![E]\!] - \mathbb{P}[\![D \cap E]\!] = 0.92 + 0.92 - (0.92 \cdot 0.92) = 0.9936$.

With this new probability defined we can also compute the probability of B_2: $\mathbb{P}[\![B_2]\!] = \mathbb{P}[\![C]\!]\mathbb{P}[\![D \cup E]\!] = 0.92 \cdot 0.9936 = 0.914112$.

Now that we have both the elements, we can compute the overall probability of the system working as follows:

$$\begin{aligned}\mathbb{P}[\![W]\!] = \mathbb{P}[\![B_1 \cup B_2]\!] &= \mathbb{P}[\![B_1]\!] + \mathbb{P}[\![B_2]\!] - \mathbb{P}[\![B_1 \cap B_2]\!] \\ &= 0.8464 + 0.914112 - (0.8464 \cdot 0.914112) = 0.9756\end{aligned}$$

Now, if we wanted we can also compute the probability of failure of the system as follows: $\mathbb{P}[\![\overline{W}]\!] = 1 - \mathbb{P}[\![W]\!] = 1 - 0.9756 = 0.0244$ ∎

2. Random Variables and Probability Distributions

This chapter will introduce us to the concept of random variables. In the first sections we are going to define the concepts of **discrete** and **continuous** random variables and their characteristic functions. Then we will explore some of the most important and widely used probability distributions.

2.1. Background

To start off, it's important to understand **what is** a random variable. To do so, let us look at the following definition.

Definition 2.1 (Random Variable)

A **random variable** is a **function** of an *outcome* defined as:

$$\begin{aligned} X &= f(\omega) \\ X : \Omega \to X(\Omega) &= \Omega_X \subseteq \mathbb{R} \end{aligned} \tag{2.1}$$

In other words, it is a *quantity that depends on a chance.*

♣

The **domain** of a random variable is the *sample space* of a random experiment Ω, while the **range**, also called the **support**, of the random variables is a subset of the real numbers $\mathbb{R}$, and it correspond to the possible values the random variable can take. Interesting cases are $\mathbb{R}, \mathbb{N}, (0, \infty), (0, 1)$.

Since random variables are functions, we may be interesed in studying their inverse. Following we give a definition of this concept.

Definition 2.2 (Generalized Inverse of a Random Variable)

Given a random variable $X : \Omega \to \Omega_X$, its **generalized inverse** is defined as:

$$X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\} \tag{2.2}$$

♣

> **Remark**
>
> The definition we gave of generalized inverse in Equation 2.2 may allow to include more than one value, thus X^{-1} is not necessarily a function.

When we defined probability, we said that given Ω and $\mathfrak{M}(\Omega)$, if $A \in \mathfrak{M}(\Omega)$, then its probability was defined as a function $\mathbb{P} : \mathfrak{M}(\Omega) \to [0, 1]$. Now, we have a random variable X defined on Ω which maps values of $\Omega \to \Omega_X \subset \mathbb{R}$. This means that if we take A and compute $X(A)$, this will get mapped to $A_X = \{x = X(\omega) : \omega \in A\} \subset \mathbb{R}$. So, formally, we can't find $\mathbb{P}[\![A_X]\!]$ because $A_X \notin \mathfrak{M}(\Omega)$, to do so, we need to define a new **sigma-algebra** which we say to be **induced by the r.v.**.

To obtain such sigma-algebra, we consider the original sigma algebra $\mathfrak{M}(\Omega)$ and we apply the mapping X to the experiment with sigma algebra $\mathfrak{M}(\Omega)$ to obtain the induced one $\mathfrak{M}(\Omega_X)$. We can now define the "new probability" as:

$$\mathbb{P}_X : \mathfrak{M}(\Omega_X) \to [0, 1]$$

If $B \in \mathfrak{M}(\Omega_X)$, that is, B is an event for the random variable X on the experiment with sample space Ω, then $\mathbb{P}_X\{B\} = \mathbb{P}[\![X^{-1}(B)]\!]$, where $X^1(B) \in \mathfrak{M}(\Omega)$. In practice we often ignore this X and we treat $\mathbb{P}$ and $\mathbb{P}_X$ as the same function. We can do this as long as we don't get confused.

Example: usage of different r.v.'s for the same experiment

Suppose we conduct an experiment where we roll two dice and record the results in order. Let's try to define Ω in this situation.

$$\Omega = \{(x_1, x_2) : x_1, x_2 \in \{1..6\}\}$$

It's easy to see that $|\Omega| = 36$. Now let's think of different random variables for this experiment:

- $X \to$ the result of the first die roll: $X((x_1, x_2)) = x_1$, where $(x_1, x_2) \in \Omega, x \in \mathbb{R}$. It's easy to see that $\Omega_X = \{1, 2, 3, 4, 5, 6\}$.
- $L \to$ the sum of both dice: $L((x_1, x_2)) = x_1 + x_2$, where $(x_1, x_2) \in \Omega, L \in \mathbb{R}$. In this case, $\Omega_L = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

Clearly, X and L are related in some way, indeed if we knew that $x = 1$, we could limit the possible values of L to $\{2, 3, 4, 5, 6, 7\}$. In informal terms, we say that X contains information about L.

Let's now try to explore the **sigma-algebra** induced by $\boldsymbol{X}$, this will be given by the **power set** $\mathfrak{M}(\Omega_X) = 2^{\Omega_X}$, that is, all possible subsets of Ω_X. For example if we take the event $B = \{2, 3, 4\} \in \mathfrak{M}(\Omega_X)$. Let's try to think about the generalized inverse of X for this event:

$$X^{-1}(\{2, 3, 4\}) = \{(2, x_2), (3, x_2), (4, x_2) : x_2 \in \{1..6\}\} \subset \Omega$$

We can notice that $|X^{-1}(\{2, 3, 4\})| = 18$; thus we can compute the probability of this event:

$$\mathbb{P}_X\{B\} = \mathbb{P}[\![X^{-1}(B)]\!] = \frac{18}{36} = \frac{1}{2}$$

Let's now try to explore the **sigma-algebra** induced by $\boldsymbol{L}$, this will be given by the **power set** $\mathfrak{M}(\Omega_L) = 2^{\Omega_L}$, that is, all possible subsets of Ω_L. In this case B is **also** an element of $\mathfrak{M}(\Omega_L)$. We can ask ourselves what is the probability of this event according to the random variable L.

First of all we'll need to compute the generalized inverse of L for this event $L^{-1}(B)$:

$$L^{-1}(\{2, 3, 4\}) = \{(1, 1), (1, 2), (2, 1), (1, 3), (2, 2), (3, 1)\} \subset \Omega$$

We can notice that $|L^{-1}(\{2, 3, 4\})| = 6$; thus we can compute the probability of this event:

$$\mathbb{P}_L\{B\} = \mathbb{P}[\![L^{-1}(B)]\!] = \frac{6}{36} = \frac{1}{6}$$

Remark

In this example we notice how it is important **not to omit** the random variable when computing probabilities, as different random variables may induce different sigma-algebras and thus different probabilities for the same event B.

The previous example illustrated how a random variable should theoretically modify the sample space and the sigma-algebra that we would use to compute our probabilities. To clarify for one last time this matter, we show what writing $\mathbb{P}_X[B]$ really translates to:

$$\mathbb{P}_X\{B\} = \mathbb{P}[\![X \in B]\!] = \mathbb{P}[\![\{\omega \in \Omega : X(\omega) \in B\}]\!] \tag{3}$$

Sometimes, if B is a singleton, we may write $\mathbb{P}_X\{X = x\}$ instead of $\mathbb{P}_X\{\{x\}\}$ to indicate the same event, where $\mathbb{P}[\![X = x]\!] \leftrightarrow \mathbb{P}[\![\omega \in \Omega : X(\omega) = x]\!]$.

2.2. Discrete and Continuous Random Variables

2.2.1. Discrete Random Variables and their Distributions

As already mentioned in the introduction, random variables can be classified in two main categories: **discrete** and **continuous**. In this section we are going to explore discrete random variables and their characteristic functions.

If a random variable X can only take a **finite** or at most **countable** number of values, we call it a **discrete random variable**.

Definition 2.3 (Distribution of Discrete R.V.)

Given a **discrete random variable** X, the *collection* of all the probabilities related to X is called the **distribution** of X.

The function

$$p_X(x) = \mathbb{P}[\![X = x]\!] \tag{2.4}$$

is called the **probability mass function**. The **cumulative distribution function** is defined as:

$$F_X(x) = \mathbb{P}[\![X \leq x]\!] = \sum_{y \leq x} P(y) \tag{2.5}$$

The set of all possible values of X is called the **support** of the distribution.

Random Variable Characterization

Suppose we toss a fair coin three times and we want to count the number of heads X that we get. The cumulative distribution function and probability mass function of such a random variable are illustrated in Figure 2.1.

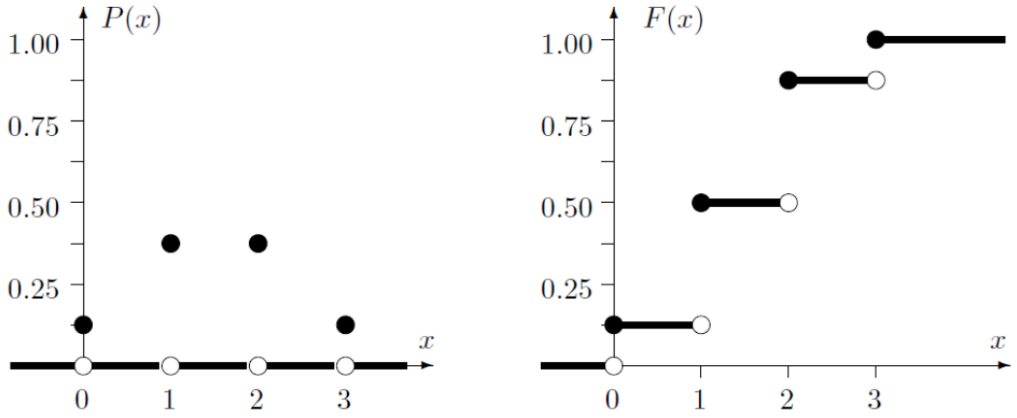


Figure 2.1: Example of Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) for a discrete random variable X representing the number of heads in three coin tosses.

> **Remark**
>
> Boh the **pmf** and the **cdf** are said two **characterize the distribution**, that is, they contain all the information we need about a random variable. This means that we may be given each one of them, for every event we can think of it's possible to compute its probability.

It is actually possible to define another function to fully characterize a random variable, called the **survival function**.

Definition 2.4 (Survival Function of Discrete R.V.)

Given a **discrete random variable** X, the **survival function** is defined as:

$$\overline{F}_X(x) = S_X(x) = \mathbb{P}[\![X > x]\!] = 1 - F(x) \tag{2.6}$$

♣

To summarize, given a **discrete** random variable X, we have the following functions that fully characterize its distribution:

- Probability Mass Function: $p_X(x) = \mathbb{P}[\![X = x]\!] \forall x \in \mathbb{R}$
- Cumulative Distribution Function: $F_X(x) = \mathbb{P}[\![X \leq x]\!] \forall x \in \mathbb{R}$
- Survival Function: $\overline{F}_X(x) = \mathbb{P}[\![X > x]\!] = 1 - F(x) \forall x \in \mathbb{R}$

Properties of PMF, CDF and Survival Function

Following we describe some important properties of the concepts and functions we have just defined. First of all, since all the functions we have defined are related to probabilities, they must satisfy the axioms in Definition 1.6°.

Once an experiment is completed, and the outcome $\omega \in \Omega$ is known, the random variable X will take a specifica value $X(\omega) = x$, therefore the collection of events $\{[X = x] : x \in \mathbb{R}\}$ will form a **partition** of the sample space Ω and thus:

$$\sum_x p_X(x) = \sum_x \mathbb{P}[\![X = x]\!] = 1$$

Particularly, for any event A, we have that:

$$\mathbb{P}[\![X \in A]\!] = \mathbb{P}[\![A]\!] = \sum_{x \in A} p_X(x)$$

2.2.2. Continuous Random Variables and their Distributions

We say that a random variable X is **continuous** if it can take an **uncountable** number of values. In this section we are going to explore continuous random variables and their characteristic functions.

The first thing we can notice is that the definition of **probability mass function** we gave for discrete random variables cannot be used in this situation: if we think about it, if X can take an uncountable number of values, if every punctual probability was different from 0, this would violate the axiom of total probability, i.e., $\sum_x \mathbb{P}[\![X = x]\!] = \infty \neq 1$.

Therefore, a first important property we can derive for continuous random variables is:

$$p_X(x) = 0, \forall x \in \mathbb{R}$$

We need to find a different way to characterize continuous random variables. To do so, we introduce the concept of **probability density function**.

Definition 2.5 (Probability Density Function)

Given a random variable X, the **probability density function** (pdf) of X is a function $f_X(x)$ such that:

$$\int_{-\infty}^{x} f_X(x)dx = F_X(x) \iff f_X(x) = \frac{dF_X(x)}{dx} \tag{2.7}$$

♣

As already mentioned, the probability density function plays for continuous variables the same role of the probability mass function for discrete variables. Indeed $f(x) \geq 0 \forall x \in \mathbb{R}$ and

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

In particular, for any event A we have that

$$\mathbb{P}[\![X \in A]\!] = \mathbb{P}[\![A]\!] = \int_{x \in A} f_X(x)\ dx$$

Example: Typical Exercise

Suppose the lifetime, in years, of some electronic component is a continuous random variable with the following probability density function:

$$f_X(x) = \begin{cases} \frac{k}{x^3} & \text{if } x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find k, draw a graph of the cdf $F(x)$ and compute the probability for the lifetime to exceed 5 years.

Solution. To find k, we need to use the property that the total integral of the pdf must be equal to 1, that is:

$$\int_1^\infty f_X(x)\, dx = \int_1^\infty \left(\frac{k}{x^3}\right) dx = 1$$

Solving the integral we get:

$$k\int_1^\infty x^{-3}\, dx = k\left[\frac{x^{-2}}{-2}\right]_1^\infty = k\left(0 + \frac{1}{2}\right) = \frac{k}{2} = 1 \implies k = 2$$

Since we know that cdf $= \int$ pdf and we know that for $x < 1$ $F(x) = 0$, we can compute $F(x)$ for $x \geq 1$:

$$F_X(x) = \int_1^x \left(\frac{2}{t^3}\right) dt = 2\left[-\frac{t^{-2}}{2}\right]_1^x = 1 - \left(\frac{1}{x^2}\right)$$

Now that we have the distribution function we can compute the probability that the lifetime exceeds five years by looking at the complement of the cdf at $x = 5$:

$$\mathbb{P}[\![X > 5]\!] = 1 - F_X(5) = 1 - \left(1 - \frac{1}{25}\right) = \frac{1}{25} = 0.04$$

Following we also show the graph of the cdf $F(x)$ which should confirm our results:

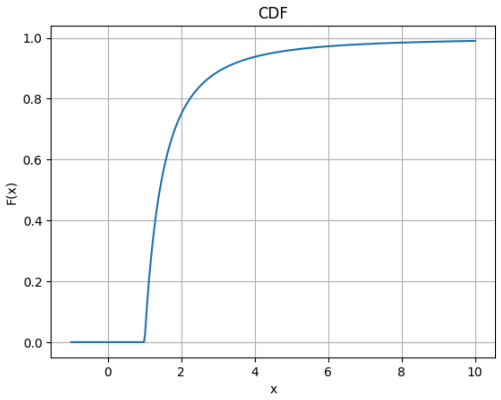


Figure 2.2: Cumulative Distribution Function (CDF) for the continuous random variable X representing the lifetime of an electronic component.

2.2.3. Discrete vs Continuous Random Variables

Following we summarize the difference between discrete and continuous random variables:

Distribution	Discrete	Continuous
Definition	(p.m.f.) $p_X(x) = \mathbb{P}[\![X = x]\!]$	(p.d.f.) $f_X(x) = F'(x)$
Probability Computation	$\mathbb{P}[\![X \in A]\!] = \sum_{x \in A} p_X(x)$	$\mathbb{P}[\![X \in A]\!] = \int_{x \in A} f_X(x)\ dx$
Cumulative Distribution Function	$F_X(x) = \mathbb{P}[\![X \leq x]\!] = \sum_{y \leq x} p_X(y)$	$F(x) = \mathbb{P}[\![X \leq x]\!] = \int_{-\infty}^{x} f(y)\ dy$
Total Probability	$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f(x)\ dx = 1$

Table 2.1: Main differences between discrete and continuous random variables.

> **Remark**
>
> In both the discrete and continuous case, the **c.d.f.** $F(x)$ is a **non decreasing** function of x, taking values in $[0, 1]$ with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. In case we were talking about the **survival function** the results would be inverted, that is, it would be a **non increasing** function of x, taking values in $[0, 1]$ with $\lim_{x \to -\infty} S(x) = 1$ and $\lim_{x \to \infty} S(x) = 0$.

2.3. Distribution of Random Vectors

Up to this moment we have only considered **one** random variable **at a time**. However in many practical situations we may be interested in studying **multiple** random variables **simultaneously**. To do so, we introduce the concept of **random vector**.

For simplicity, we focus on a vector $(\boldsymbol{X}, \boldsymbol{Y})$ of dimnesion 2, but everything can be extended to higher dimensions.

Since we are talking about more than one random variable, we need to define the concept of **joint p.m.f** and **joint p.d.f**. Given two variables X, Y we have:

$$p_{X,Y}(x, y) = \mathbb{P}[\![(X, Y) = (x, y)]\!] = \mathbb{P}[\![X = y \cap Y = y]\!] \tag{2.8}$$

In this course, we are going to focus on homogeneous randon vectors, in which all the random variables are either discrete or continuous.

Definition 2.6 (Joint Cumulative Distribution Function)

For a vector of random variables, the **joint cumulative distribution function** is defined as follows:

$$F_{X,Y}(x, y) = \mathbb{P}[\![X \leq x \cap Y \leq y]\!] \tag{2.9}$$

Given the join c.d.f we can also define the **join probability density function** as the mixed derivative of the joint c.d.f:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \tag{2.10}$$

♣

Following we report in a table the main formulas we need to use in case we are dealing with continuous or discrete random vectors.

Distribution	Discrete	Continuous
Marginal Distributions	$p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$	$f_X(x) = \int f_{X,Y}(x,y)dy$ $f_Y(y) = \int f_{X,Y}(x,y)dx$
Independence	$p_{X,Y}(x,y) = p_X(x)p_Y(y$	$f_{X,Y}(x,y) = f_X(x)f_Y(y)$
Computing Probabilities	$\mathbb{P}[\![(X,Y) \in A]\!] =$ $\sum_{(x,y)\in A}\sum p_{X,Y}(x,y)$	$\mathbb{P}[\![(X,Y) \in A]\!] =$ $\int\int_{(x,y)\in A} f_{X,Y}(x,y)\ dx\ dy$

Table 2.2: Main formulas for discrete and continuous random vectors.

Table 2.2 already introduces to the concept of **independence** between two random variables. To clarify this concept, we give the following definition.

Definition 2.7 (Independent Random Variables)

Given random variables X, Y we say that they are **independent** if:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \tag{2.11}$$

for all values of x, y. This means the events $\{X = x\}, \{Y = y\}$ are independent for all x, y. In other words, variables X and Y take their values independently of each other.

♣

Remark

Clearly Definition 2.7° is only taking into account discrete random variables, however the same definition can be extended to continuous random variables by substituting the p.m.f with the p.d.f.

Example: Study of Independency between Two R.V.'s

Consider the case in which we have X and Y, two **continuous r.v.**'s with joint probability density function given by:

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

We want to study if X and Y are independent. The answer is **no**, to see this we can look at the **support** of the random vector (X, Y), which is **not a rectangle**. Indeed, it is the unit circle, thus if we know the value of X, this will limit the possible values of Y and viceversa. Therefore, X and Y cannot be independent.

Exercise. Suppose we have the following probability density function for two continuous random variables X and Y:

$$f_{X,Y}(x,y) = \begin{cases} k & \text{if } x^2 + y^2 \leq r^2 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of k such that $f_{X,Y}(x,y)$ is a valid probability density function for (X, Y).

Solution. First of all, we know that k must be **non-negative**: $K \geq 1$. This is useful as a lower bound. The only thing that is left is to find an upper bound for k.

To upper bound this, we know that the probability of the whole sample space must be equal to 1, that is $\mathbb{P}[\![\Omega_{(X,Y)}]\!] = 1$. We also know that to compute a probability we can start from the probability density function and integrate it over the support of the random vector (X, Y):

$$\mathbb{P}[\![\Omega_{(X,Y)}]\!] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy = 1$$

The support of the random vector X, Y is the set of points inside the circle of radius r. That is, $\Omega_{(X_Y)} = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq r^\}$ which is shown in the graph below:

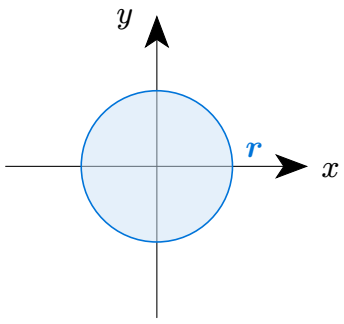


Figure 2.3: Support of the random vector (X, Y)

Therefore we can rewrite in the previous integral simply as the integral over the support, the integral everywhere else will be zero. In this case, since we are computing an integral over a circular region, we are basically searching for a value k such that the cylinder with radius r and height k has volume equal to 1.

Instead of computing the integral we can instead compute the volume of such cylinder with the classic formula $V = \pi r^2 h$; therefore we have:

$$\int\int_{\Omega_{XY}} = \pi r^2 k = 1 \implies k = \frac{1}{\pi r^2}$$

■

2.4. Conditional Probabilities

Consider the previous exercise, but now suppose the support was $\Omega_{X,Y} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, it would be just the *circumference* not the whole circle. In this case, any time we fix a value for the X variable, there are only two possible values that Y can take:

$$Y \mid X \begin{cases} -\sqrt{1-x^2} \\ \sqrt{1-x^2} \end{cases}$$

So even though this looks as a random vector of two random variables, in reality, they are not both continuous random variables. Whenever we fix one variable, say $X = x$, what we miss about the other variable is a **discrete choice**.

> **Remark**
>
> As a general rule, in order to be able to properly compute the probability density function of a random vector of dimension n (where n is the number of random variables), we need to be able to compute the 'volume' of the hyper-surface induced by the **support** of the random vector in $\mathbb{R}^n$. In the previous case, even though we had two random variables, the support was a circumference, which is a one-dimensional object, thus we could not define a proper two-dimensional probability density function. These objects for which the dimension of the support is lower than the dimensionality they are represented in are called **manifolds**.

Suppose we wanted to compute the probability of an event A for a random vector (X, Y) whose support is the manifold defined by $x^2 + y^2 = 1$. This situation is represented in Figure 2.4. In practice we have that:

$$\begin{aligned} \mathbb{P}[\![A]\!] &= \mathbb{P}[\![(x_1 \leq x \leq x_2) \cap (x^2 + y^2 = 1)]\!] \\ &= \mathbb{P}\Big[\!\Big[(x_1 \leq x \leq x_2) \cap \left[\left(y = \sqrt{1-x^2}\right) \cup \left(y = -\sqrt{1-x^2}\right)\right]\Big]\!\Big] \\ &= \mathbb{P}\Big[\!\Big[(x_1 \leq x \leq x_2) \cap \left(y = \sqrt{1-x^2}\right)\Big]\!\Big] + \mathbb{P}\Big[\!\Big[(x_1 \leq x \leq x_2) \cap \left(y = -\sqrt{1-x^2}\right)\Big]\!\Big] \end{aligned}$$

Clearly, it is not possible to compute a **double integral** over a two dimensional: if we fix a value for x there is only a possible for y. We need a different method to deal with this.

The method we need to address such a situation is **conditional probability**, that we are just about to define.

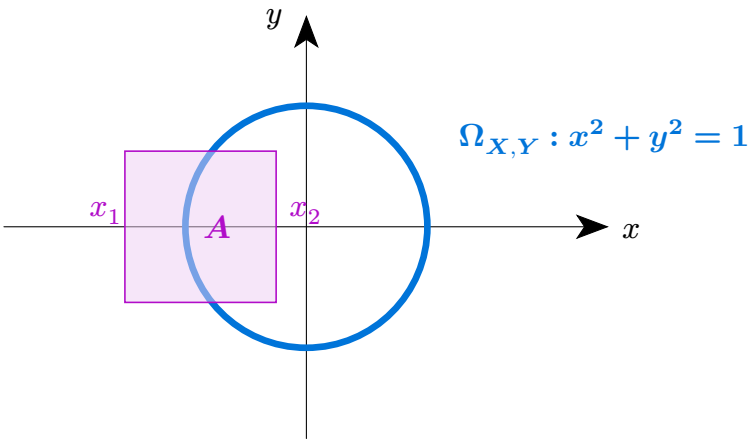


Figure 2.4: Support of the random vector (X, Y)

Definition 2.8 (Conditional Probability for Discrete R.V.'s)

Suppose we have X, Y two random variables. If these r.v.'s are **discrete**, we can define the **conditional probability mass function** of Y given X as:

$$p_{(Y|X)}(y|x) : \mathbb{P}[\![Y = y \mid X = x]\!] = \frac{\mathbb{P}[\![X = x \cap Y = y]\!]}{\mathbb{P}[\![X = x]\!]} = \frac{p_{X,Y}(xy)}{p_X(x)} \tag{2.12}$$

where the numerator of the last fraction is given by Equation 2.8 and the denominator is the **marginal p.m.f.** of X. Actually, this is not just *one* p.m.f., but we have a different one for *each value* of x. We can now define the **conditional cumulative distribution function** as:

$$F_{(Y|X)}(y|x) = \mathbb{P}[\![Y \leq y \mid X = x]\!] = \frac{\mathbb{P}[\![X = x \cap Y \leq y]\!]}{\mathbb{P}[\![X = x]\!]} \tag{2.13}$$

If we look at Equation 2.13, we can notice we can write it as a summation of probability mass functions, therefore it becomes:

$$F_{(Y|X)}(y|x) = \sum_{\hat{y} \leq y} p_{(Y|X)}(\hat{y}|x) \tag{2.14}$$

♣

Let's now take a look at the continuous case, where not surprisingly we are going to use density functions and replace summations with integrals.

Definition 2.9 (Conditional Probability for Continuous R.V.'s)

Suppose we have X, Y two random variables. If these r.v.'s are **continuous**, we can define the **conditional probability density function** of Y given X as:

$$f_{(Y|X)}(y|x) : \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{15}$$

As far as the **conditional cumulative distribution function** is concerned, we have:

$$F_{(Y|X)}(y|x) = \int_{-\infty}^{y} f_{(Y|X)}(y|x)\ dy \tag{2.16}$$

♣

⚠ Warning

It may look promising to try and directly define the conditional probability distribution function with the following formula:

$$F_{(Y|X)}(y|x) = \frac{F_{X,Y}(x,y)}{F_X(x)}$$

This is **completely wrong**, indeed if we think for a moment about the meaning of the c.d.f., we can notice that this formula does not make any sense. What this formula really does is compute the probability that $Y \leq y$ given that $X \leq x$, that is:

$$\mathbb{P}[\![Y \leq y \mid X \leq x]\!] = \frac{\mathbb{P}[\![X \leq x \cap Y \leq y]\!]}{\mathbb{P}[\![X \leq x]\!]} = \frac{F_{X,Y}(x,y)}{F_X(x)}$$

In general, for any two variables X, Y, depending on whether they are discrete or continuous, we can always factorize in the following ways:

- For *discrete* random variables we can factorize their joint probability mass function as:

$$p_{X,Y}(x,y) = p_X(x)p(Y|X)(y|x)$$

- For *continuous* random variables the joint prob. density function can be factorized as:

$$f_{X,Y}(x,y) = f_X(x)f(Y|X)(y|x)$$

Of course if the random variables are independent we'll have that the conditional mass / density functions will be equal to the marginal ones. As far as distribution functions are concerned we **cannot** factorize them as:

$$F_{X,Y}(x,y) = F_X(x)F_{Y|X}(y|x)$$

The mathematical reason behind this is that we are dealing with incompatible objects:

- the orange components use X to represent many possible values for x
- the blue component uses a fixed value of x

Theorem 2.1 (Chain Rule for Joint Probabilities)

Given a random vector of n random variables $\overline{X} = (X_1, X_2, ..., X_n)$. If the vector is made of **discrete** random variables, we can factorize the joint probability mass function as:

$$p_{\overline{X}}(\overline{x}) = p_{X_1}(x_1) \cdot p_{X_2|X_1}(x_2|x_1) \cdot p_{X_3|X_1,X_2}(x_3|x_1,x_2)...$$

We can do the same for **continuous** random variables by replacing the probability mass functions with probability density functions:

$$f_{\overline{X}}(\overline{x}) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \cdot f_{X_3|X_1,X_2}(x_3|x_1,x_2)...$$

Getting back to our examples, remember that we want to find k such that the function

$$f_{X,Y}(x,y) = \begin{cases} k \text{ if } x^2 + y^2 = 1 \\ 0 \quad \text{otherwise} \end{cases}$$

is a valid *probability density function*. Remember this is not a density in $\mathbb{R}^2$, but rather a density over the one-dimensional manifold defined by the circumference of radius 1. We can imagine to *unfold* the circumference, transforming it into a line of length equal to the circumference itself. Remember that the circumference has radius 1, therefore its length is equal to 2π. Figure 2.5 shows how this unfolding process works.

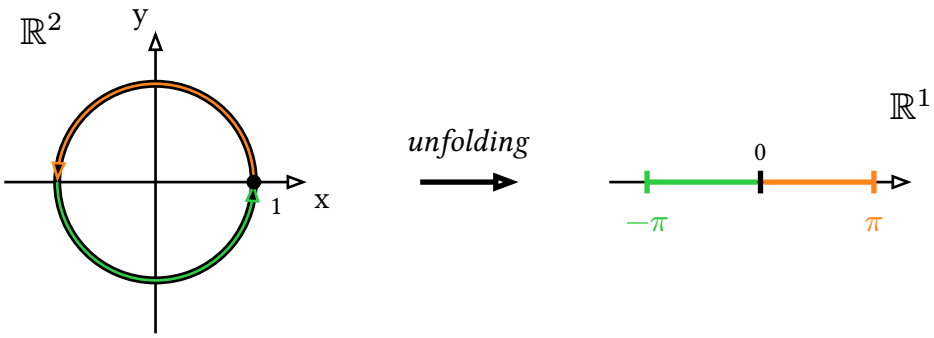


Figure 2.5: Mapping from the unit circle in $\mathbb{R}^2$ to $\mathbb{R}^1$

In this new 'unfolded' space, we can notice that for the orange part:

- X can take values in the interval $[-1, 1]$, indeed, if we go back to the original space the coordinate x can be any value between -1 and 1.
- Y can only take values in the ranger $[0, 1]$, since it is the upper part of the circle.

We can conclude that $Y|X = \sqrt{1-x^2}$. Similarly, for the green part we have that $Y|X = -\sqrt{1-x^2}$, since Y can only take values in the range $[-1, 0]$. We can now solve the original problem by solving the integral:

$$\int_{-\pi}^{\pi} k \, dx = kx \Big|_{-\pi}^{\pi} = k\pi + k\pi = 2k\pi$$

Thus since we have the constraint that the upper bound must be equal to 1, we have that $2k\pi = 1$, therefore $k = \frac{1}{2\pi}$. We can also notice that the probability of Y being larger or smaller thant 0 is completely independent of the value of X.

We can define a new random variable I which serves as an **indicator** of the sign of Y:

$$I = \begin{cases} 1 \quad \text{if } Y \geq 0 \\ 0 \quad \text{otherwise} \end{cases} \implies (X,Y) = \left(X, I\sqrt{1-x^2}\right)$$

Clearly I is a **discrete random variable**. Now it is evident that in the beginning we were not dealing with two continuous random variables, but rather with a continuous random variable X and a discrete random variable I.

2.5. Probability Density Factorization

Even though we already mentioned the concept of independency between random variables in Definition 2.7°, it's worth to revisit that definition giving a practical tool to check if two random variables are independent or not.

2.5.1. Independency via Factorization

We say that two random variables X, Y are independent if and only if it is possible to find functions h_1 and h_2 such that:

- for **discrete** random variables we have that

$$p_{X,Y}(x, y) = h_1(x)\ h_2(y) \quad \forall x, y \in \mathbb{R}^2$$

  in this case we can also say that the marginal $p_X(x) \propto h_1(x)$ and equivalently $p_Y(y) \propto h_2(y)$, that is the marginal are proportional to the functions h_1 and h_2 respectively (or equal up to a constant)

- for **continuous** random variables we have that

$$f_{X,Y}(x, y) = h_1(x)\ h_2(y) \quad \forall x, y \in \mathbb{R}^2$$

  where, again the marginals can be seen as proportional to the functions h_1 and h_2 respectively: $f_X(x) \propto h_1(x)$ and equivalently $f_Y(y) \propto h_2(y)$.

Example: Independency via factorization

Two random variable T, S have join probability density function given by:

$$f_{T,S}(t, s) = 18e^{-6t}e^{-3s}\mathbb{1}_{t,s\in R_+^2}$$

where the indicator function has exactly the function of expressing the concept that when we are considering negative values for either t or s the density is zero.

We can try to **factorize** the joint p.d.f. dividing the two members. The only 'tricky' part is the indicator function but we can notice that $(t, s) \in (\mathbb{R}_+^2 \Leftrightarrow t \in R_+ \cap s \in R_+)$; therefore we can write the following:

$$f_{T,S}(t, s) = 18e^{-6t}\mathbb{1}_{t\in R_+} \quad e^{-3s}\mathbb{1}_{s\in R_+}$$

where we can see the purple factor as $h_1(t)$, and the green factor as $h_2(s)$. Therefore, we can conclude that T and S are independent random variables.

We may now ask ourselves whether $h_1(t) = f_T(t)$ and $h_2(s) = f_S(s)$; indeed from what we said above we know they are proportional, but we can't say anything about the **propor-**

tionality constant. To solve this we can rewrite the marginals as scaled by an unknown proportionality constant. We are going to first focus on $f_T(t)$:

$$f_T(t) = k\ h_1(t) = 18ke^{-6t}\mathbb{1}_{t>0}$$

From the *law of total probability* we also know that integrating the marginal over the whole support $\mathbb{R}$ should give us 1; therefore we can write:

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} f_T(t)\ dt = \int_0^{\infty} 18ke^{-6t}\ dt &= 18k \int_0^{\infty} e^{-6t}\ dt \\ &= 18k\left[-\frac{1}{6}e^{6t}\right]_0^{\infty} = 18k\left[-\frac{0}{6} - \left(-\frac{1}{6}\right)\right] = 3k \end{aligned}$$

Therefore we can conclude that the constant k we were looking for is equal to $\frac{1}{3}$. So we can conclude that marginal density function of T is:

$$F_T(t) = 6e^{-6t}\mathbb{1}_{t>0}$$

To compute the second marginal $f_S(s)$ we don't even need to compute a second integral; that's because we know the joint p.d.f. can be written as a product of the marginals (by independency), therefore we can compute the following:

$$\begin{aligned} f_{T,S}(t,s) &= 18e^{-6t}\mathbb{1}_{t>0}\ e^{-3s}\mathbb{1}_{s>0} \\ &= f_T(t)\ f_S(s) \\ &= 6e^{-6t}\mathbb{1}_{t>0}\ 3e^{-3s}\mathbb{1}_{s>0} \end{aligned}$$

that is because since we need to get back to the original 18 in the joint p.d.f., the only possible value for the proportionality factor of $h_2(s)$ is 3. ■

By looking at the previous example we can also notice that there is another (calculus-based) method to compute the marginal probabilities, that is computing the integrals of the density functions over the other variables:

$$f_T(t) = \int_0^{\infty} f_{T,S}(t,s)\ ds \qquad f_S(s) = \int_0^{\infty} f_{T,S}(t,s)\ dt$$

2.5.2. Factorization via Marginal and Conditional Probabilities

Example: Dependent Variables Factorization

Suppose we have two R.V.'s X, Y with joint probability density function given by:

$$f_{X,Y}(x,y) = xe^{-x(y+1)}\mathbb{1}_{x,y\in R_+^2}$$

First of all, we don't even know whether this is a *valid p.d.f.*, in order to *check* this we need to compute the double integral over the whole support and require it to be equal to 1:

$$\int_0^\infty \int_0^\infty x^{e^{-x(y+1)}} \, dxdy = 1$$

Since this double integral is quite ugly we can try to **factorize** the joint p.d.f. by separating the terms depending of only on x and those depending only on y (if possible):

$$f_{X,Y}(x,y) = xe^{-x}\mathbb{1}_{\{x>0\}} \cdot e^{-xy}\mathbb{1}_{\{y>0\}} \tag{2.17}$$

For what concerns the first term, it's possible to notice it only depends on x, while the second one cannot be further simplified since it depends on on both x and y. We can conclude the two random variables are **not independent**.

It is always possible to **factorize** the joint p.d.f of two random variables X, Y using the marginal and conditional probabilities:

$$f_{X,Y}(x,y) = f_X(x) \; f_{Y \mid X}(y \mid x) \tag{2.18}$$

⚠ Warning

If the random variables are not independent, and don't have additional information, there is not shortcut to know which is the marginal and which is the conditional probability density function. So going back to the previous example, it would be totally **wrong** to see the factorization in as if the two factors were the marginal and conditional p.d.f. respectively as in Equation 2.18.

To address the previous exercise, one may recall that the **exponential distribution** has form

$$\lambda e^{-\lambda t}\mathbb{1}_{t>0}$$

and is always a valid p.d.f. for any $\lambda > 0$. With this new piece of information, we can now look again at the factorization in and notice that we could actually reorder the factors so that both members look like valid exponential marginal distributions:

$$f_{X,Y}(x,y) = e^{-x}\mathbb{1}_{x>0} \cdot xe^{-xy}\mathbb{1}_{y>0}$$

In this way, the first factor looks like an exponential distribution with $\lambda = 1$ and the second number looks like an exponential distribution with $\lambda = x$. Now we can really say that $f_{X,Y}(x,y)$ can be factorized in marginal and conditional where:

- the **marginal** p.d.f. of X is given by $e^{-x}\mathbb{1}_{x>0}$
- the **conditional** p.d.f. of Y given X is given by $xe^{-xy}\mathbb{1}_{y>0}$

To check the correctness of this factorization we can now compute check that the integral over all possible values of y yields the marginal of X, and that the joint p.d.f. divided the the marginal of X yields the conditional of Y given X, that is:

$$f_X(x) = \int_0^\infty f_{X,Y}(x,y)dy = e^{-x} \qquad \frac{f_{X,Y}(x,y)}{f_X(x)} = xe^{-xy}$$

2.6. Conditional Independency

2.6.1. Characterization of Single Random Variables

Following we try to summarize what we have seen so far about random variables and random vectors. Starting from a **single random variable**, its *distribution* is characterized by either of:

- either one between probability **mass** function or **density** function (depending on whether the random variable is discrete or continuous)
- the **cumulative distribution function** which is obtained by integrating or summing the p.m.f. or p.d.f. over the possible values of the random variable.
- the **survival function** which is obtained by computing the complement of the c.d.f.

2.6.2. Characterization of Random Vectors

In case we are dealing with a **pair** of random variables, the *distribution* (X, Y) is characterized by either of:

- the **joint** probability **mass** function or **density** function (depending on whether the random variables are discrete or continuous)
- the **joint cumulative distribution function** which is obtained by integrating or summing the joint p.m.f. or p.d.f. over the possible values of the random variables.
- the **joint survival function** which is obtained by computing the complement of the joint cumulative distribution function
- the **marginal** of X and the **conditional** of Y given X (or viceversa); when referring to marginal and conditional distribution we may referring to either one of the p.m.f. / p.d.f., c.d.f, or survival function
- the **marginals** for X and Y and the information that they are **independent**

2.6.3. Independency and Conditional Independency

In case we are told that it is possible to get the joint of X and Y from just the two marginals, then we can conclude that X and Y are independent random variables.

Similarly, if we are told that it is possible to get the joint of (X, Y, Z) from just the three marginals, then we can conclude that X, Y, Z are **mutually independent**, that is every possible pair is independent: $X \perp Y, Y \perp Z, X \perp Z$.

In case we are told that it's possible to get the joint of (X, Y, Z) from the marginal of X, the conditional of Y given X and the condition of Z given Y, this means we don't need knowledge about $Z \mid X, Y$; we can conclude Z is **conditionally independent** of X given Y. We can formalize this concept in Definition 2.10°.

Definition 2.10 (Conditional Independency)

Given three random variables (X, Y, Z), we say that Z is **conditionally independent** of X given Y if and only if:

$$f_{Z \mid Y}(z \mid y) = f_{Z \mid X,Y}(z \mid x, y) \tag{2.19}$$

which basically means that the knowledge of X does not provide any additional information about Z once we know Y.

♣

Remark

With respect of the conditional independency definition, we can rewrite the density of Z given X and Y by means of Equation 15 (conditional p.d.f.):

$$f_{Z \mid X,Y}(z \mid x, y) = \frac{f_{Z,X \mid Y}(z, x \mid y)}{f_{X \mid Y}(x \mid y)}$$

where the rational is that to compute the conditional probability of something conditioned to something else, we can always computed it dividing their joint probability by the marginal of the conditioning variable. With this in mind we can derive:

$$f_{Z \mid Y}(z \mid y) f_{X \mid Y}(x \mid y) = f_{Z,X \mid Y}(z, x \mid y)$$

2.7. Characteristics of a Distribution

In the previous chapter we discussed some conditions that are sufficient to determine all the characteristics of a distribution, we didn't mention however what these characteristics actually are. Since there is potentially a plethora of characteristics that can be defined given a distribution, we are going to focus on just some of them, specifically we are dividing them in three principal **families**:

- measures of **central tendency**: the *mean*, *median* and *mode*
- measures of **variability**: *variance*, *standard deviation*, *range*, *interquartile range*,
- measures of **position** or **shape**: *quantiles* , *skewness* and *kurtosis*

We are also going to measure the **relationship** between **two variables**, by means of tools such as *covariance* and *correlation* (potentially with many different indices).

Expected Value of a Random Variable

Definition 2.11 (Mean of a Random Variable)

Given a random variable X we define the **mean** or **expected value** as:

$$\mu_X = \mathbb{E}[\![X]\!] = \sum_{x \in \Omega_X} x \; p_X)(x) \quad \text{or} \quad \mu_X = \mathbb{E}[\![X]\!] \int_{x \in \Omega_X} x \; f_X(x) \; dx \tag{2.20}$$

basically, it is a weighted average of all possible values the random variable can take, where the weights are given by the probabilities of each value.

♣

We can see the **mean** of a random variably as the **physical center of mass** of its distribution. Indeed, if we imagine the p.d.f. as a physical object with density proportional to the value of the p.d.f. at each point, then the mean is the point where we could balance this object on the tip of a pencil.

Mode and Median

Imagine we want to guess the result of an experiment for a random variable X. To do so we may come up with different ideas.

We could pick the value with the highest probability or highest density value in case the random variable is continuous; this is called the **mode** of the random variable. It is the value that maximizes the probability of correctly guessing the outcome.

Clearly, this approach is flawed in some sense, consider for instance the experiment of throwing a die, every event has probability $\frac{1}{6}$, so we can't actually define a mode here. To see another example, suppose we have a discrete random variable X with $\Omega_X = [-n, n]$. The probability of having $x = 0$ is $\frac{1}{2}$ and the probability of having any other value is evenly split among the remaining values. The choice in this case would be to pick 0 as the value independently of the value of n. The problem is there is always a positive probability that $X = n$, so if for instance $n = 1,000,000$, we would have $\frac{1}{2}$ probability of being off by very large amounts.

Let a be a possible guess for the value of X. We can measure our error as $|a - x|$, that is, we compute the **absolute error**. Before doing the experiment the absolute error is unknown, but it can be seen as a random variable itself, thus we could try to minimize it in expectation choosing the value of a that minimizes $\mathbb{E}[\![|a - X|]\!]$.

$$m = \arg\min_{a \in \mathbb{R}} \mathbb{E}[\![|a - X|]\!]$$

The value m is actually called the **median** of the random variable X.

Properties of the Expected Value

Before looking at the properties of the expected value, it's worth to define the concept of **linear operator**. Let's look at the following definition.

Definition 2.12 (Linear Operator)

An operator $\mathscr{g}$ is a function such that for any linear transformation $\mathcal{T}$ it is the case that $\mathscr{g} \cdot \mathcal{T} = \mathcal{T} \cdot \mathscr{g}$, in other words we have that:

$$\mathscr{g}(ax + b) = a\mathscr{g}(x) + b$$

♣

Linearity

The expected value is indeed a **linear operator**. This means that for any random variable X and for any constants $a, b \in \mathbb{R}$ it is the case that:

$$\mathbb{E}[\![aX + b]\!] = a\mathbb{E}[\![X]\!] + b \tag{2.21}$$

Equation 2.21 holds only in case of linear transformations of the random variable.

> **Example: Linearity of Expected Value**
>
> A discrete random variable X with $\Omega_X = \{0, 1\}$ where the probability mass function is given by:
>
> $$p_X(X) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$
>
> The expected value of X is given by:
>
> $$\mathbb{E}[\![X]\!] = 0 \cdot (1 - p) + 1 \cdot p = p$$
>
> Similarly, if we apply a linear transformation to X, for instance $Y = 25X - 2$ we have:
>
> $$\mathbb{E}[\![Y]\!] = \mathbb{E}[\![25X - 2]\!] = 25\mathbb{E}[\![X]\!] - 2 = 25p - 2$$

Definition 2.11° suggests us that we can calculate the expected value of a random variable W where W is a linear transformation of X, i.e., $W = \mathcal{T}(X)$ without even knowing the distribution of W. We can compute the expected value for **any function of a random variable** as :

$$\mathbb{E}[\![h(X)]\!] = \begin{cases} \sum\limits_{x \in \Omega_X} h(x)\ p_X(x) & \text{if } X \text{ is discrete} \\ \int\limits_{x \in \Omega_X} h(x)\ f_X(x)\ dx & \text{if } X \text{ is continuous} \end{cases} \tag{2.22}$$

As a corollary of this property, we can also notice that the expected value of a constant is equal to the constant itself, and that given n random variables $X_1, \ldots, X_n$ we have that:

$$\mathbb{E}\left[\!\!\left[\sum_{i=1}^{n} X_i\right]\!\!\right] = \sum_{i=1}^{n} \mathbb{E}[\![X_i]\!]$$

> **Remark**
>
> As $n \to \infty$, under adequate conditions, we may have that
>
> $$\sum_{i=1}^{n} X_i \to X^*$$
>
> with X^* being another 'fresh' random variable. If this is the case we have that

$$\sum_{i=1}^{n} \mathbb{E}[\![X_i]\!] \to \mathbb{E}[\![X^*]\!]$$

Jensen's Inequality

With respect to Equation 2.22, we can notice that if the function h is **convex**, then the following inequality holds:

$$\mathbb{E}[\![h(X)]\!] \geq \mathbb{E}[\![X]\!] \tag{2.23}$$

on the other hand if the function h is **concave**, then the opposite inequality holds:

$$\mathbb{E}[\![h(X)]\!] \leq \mathbb{E}[\![X]\!] \tag{2.24}$$

Mean Squared Error Minimizer

Consider the previous example where we were trying to find the best possible guess for the outcome of some experiment. Suppose that instead of the absolute error we were trying to minimized the **mean squared error**, we would find the optimal value is $\mathbb{E}[\![X]\!]$, that is:

$$\mathbb{E}[\![X]\!] = \arg\min_{a \in \mathbb{R}} \mathbb{E}[\![(a - X)^2]\!] \tag{2.25}$$

Variance of a Random Variable

Consider Equation 2.25. If we try to substitute the optimal value $a = \mathbb{E}[\![X]\!]$ in the expression we get the following equation:

$$\min_{a \in \mathbb{R}} \mathbb{E}[\![(a - X)^2]\!] = \mathbb{E}[\![(\mathbb{E}[\![X]\!] - X)^2]\!] = \text{Var}\{X\} \tag{2.26}$$

that is, the **variance** of a random variable X can be seen as the minimum mean squared error we can get when trying to guess the outcome of the experiment. We can also write the variance of a random variable by expanding the square in Equation 2.26:

$$\text{Var}\{X\} = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]^2 \tag{2.27}$$

Properties of the Variance

Scaling Property

If we try to compute the variance of a **linear transformation** we obtain:

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

This also suggests us that the variance of a constant is equal to zero, this is by no surprise since the variance is used to measure the variability of a random variable, and a constant has no variability at all.

Law of Total Variance

Consider 2 (continuous) random variables X, Y with p.d.f. $f_{X,Y}(x, y)$. Given this random vector we can say that it is fully defined by its join p.d.f., that is $(X, Y) \sim f_{X,Y}(x, y)$, but we can also derive the marginal probabilities both for X and for Y. Given all the previous quantities we can also define the conditional probability of X given Y and viceversa:

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

Definition 2.13 (Conditional Expectation)

Given a random vector of two continuous random variables (X, Y) we can define the **conditional expectation** as the expected value of one random variable given the other:

$$\mathbb{E}[\![X \mid Y]\!] = \int_{-\infty}^{\infty} x \, f_{X \mid Y}(x \mid y) \, dx$$

Clearly, this depends on the value taken by Y.

♣

It is important to notice that $\mathbb{E}[\![X \mid Y]\!]$ is random because Y is random. For every fixed $y \in \Omega_Y$, $\mathbb{E}[\![X \mid Y = y]\!]$ is a number, but since Y is random, $\mathbb{E}[\![X \mid Y]\!]$ is a random variable. Basically we can see $\mathbb{E}[\![X \mid Y]\!]$ as a function of y, say, $g(y)$. We can try to compute the expected value of $g(y)$:

$$\begin{aligned}
\mathbb{E}[\![g(y)]\!]_Y = \mathbb{E}[\![\mathbb{E}[\![X \mid Y]\!]_X]\!]_Y &= \int_{-\infty}^{\infty} \mathbb{E}[\![X \mid Y = y]\!] \, f_Y(y) \, dy \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x \, f_{X \mid Y}(x \mid y) \, dx \right] f_Y(y) \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) \, dx \, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) \, dy \, dx = \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \right] dx \\
&= \int_{-\infty}^{\infty} x f_X(x) \, dx = \mathbb{E}[\![X]\!]
\end{aligned}$$

Definition 2.14 (Conditional Variance)

Given two random variables (X, Y) we can define the **conditional variance** as:

$$\mathrm{Var}\{X \mid Y\} = \mathbb{E}[\![X^2 \mid Y]\!] - \mathbb{E}[\![X \mid Y]\!]^2$$

♣

Again, the value of the conditional variance may depend on the value taken by Y. Similarly to the previous case, since the conditional variance can be seen as a random variable, we can try to compute the expected value of the conditional variance.

$$\mathbb{E}[\![\mathrm{Var}\{X \mid Y\}]\!] = \mathbb{E}[\![\mathbb{E}[\![X^2 \mid Y]\!]]\!] - \mathbb{E}[\![\mathbb{E}[\![X \mid Y]\!]^2]\!]$$

This is obtained by **linearity** of the expected value operator. We can notice that the first member of the difference is actually equal to $\mathbb{E}[\![X^2]\!]$, indeed the outer expectation is taken with respect to the value of Y, thus it serves to take into account all possible values for Y. For the second member however we can't do the same, since there is a square involved, we can notice the following:

$$\mathbb{E}[\![W^2]\!] - \mathbb{E}[\![W]\!]^2 = \mathrm{Var}\{W\} \implies \mathbb{E}[\![W^2]\!] = \mathrm{Var}\{W\} + \mathbb{E}[\![W]\!]^2$$

Thus we can rewrite the previous quantity as:

$$\begin{aligned}\mathbb{E}[\![\mathrm{Var}\{X \mid Y\}]\!] &= \mathbb{E}[\![X^2]\!] - [\mathrm{Var}\{\mathbb{E}[\![X \mid Y]\!]\} + \mathbb{E}[\![\mathbb{E}[\![X \mid Y]\!]]\!]^2] \\ &= \mathbb{E}[\![X^2]\!] - \mathrm{Var}\{\mathbb{E}[\![X \mid Y]\!]\} - \mathbb{E}[\![X]\!]^2\end{aligned}$$

where we used the property seen before that $\mathbb{E}[\![\mathbb{E}[\![X \mid Y]\!]]\!] = \mathbb{E}[\![X]\!]$ and we squared it. Now if we put together the terms depending only on the X we can obtain the variance back:

$$\mathbb{E}[\![\mathrm{Var}\{X \mid Y\}]\!] = \mathrm{Var}\{X\} - \mathrm{Var}\{\mathbb{E}[\![X \mid Y]\!]\}$$

To take this one step even further we can rearrange the terms to obtain the following important result called the **law of total variance**

$$\mathrm{Var}\{X\} = \mathbb{E}[\![\mathrm{Var}\{X \mid Y\}]\!] + \mathrm{Var}\{\mathbb{E}[\![X \mid Y]\!]\} \tag{2.28}$$

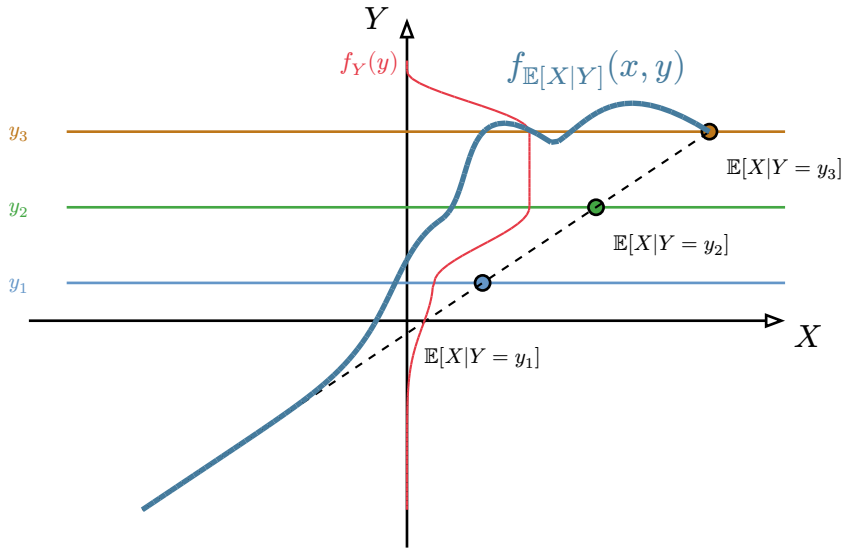


Figure 2.6: Visualization of conditional expectation distribution function

Consider Figure 2.6: whenever we fix a value for Y, which in the picture corresponds to y_1, y_2, y_3, we may find the expected value of X given that value of Y. In red we can see the density function of Y. Now since we said we can visualize the conditional expectation as a random variable itself, according to the value that Y takes, we can also visualize its density function, which is represented in blue in the picture.

Covariance and Correlation

Up to now we never really talked about the **relationship** between two random variables. In order to quantify this information it is necessary to introduce the concepts of **covariance** and **correlation**.

Definition 2.15 (Covariance)

Given two random variables (X, Y) we can define the **covariance** between them σ_{XY} and it is given by the following equation:

$$\begin{aligned} \text{Cov}\{X, Y\} &= \mathbb{E}[\![(X - \mathbb{E}[\![X]\!])(Y - \mathbb{E}[\![Y]\!])]\!] \\ &= \mathbb{E}[\![XY]\!] - \mathbb{E}[\![X]\!]\mathbb{E}[\![Y]\!] \end{aligned} \tag{2.29}$$

♣

Usually, when people talk about covariance they refer to it as the measure of association between two random variables, as if this is the only possible way of measuring association. In reality, this is a very limited index, which can only measure **linear association**.

Expected Value and Variance

Let X be a Binomial random variable with parameters n and p. Considering its probability mass function in Equation 3.4, we can compute its **expected value** as:

$$\mathbb{E}[\![X]\!] = \mathbb{E}\left[\!\!\left[\sum_{i=1}^{n} X_i\right]\!\!\right] = \sum_{i=1}^{n} \mathbb{E}[\![X_i]\!] = n \cdot p \tag{2.30}$$

where we used the linearity of expectation to move the expectation inside the sum. As far as the **variance** is concerned, we cannot use the same trick, in that the variance operator is not linear.

Consider two discrete random variables X, Y with joint probability mass function $p_{XY}(x, y)$ and with support which is represented in Figure 2.7. Suppose also that the orange point is the point $(\mathbb{E}[\![X]\!], \mathbb{E}[\![Y]\!])$.

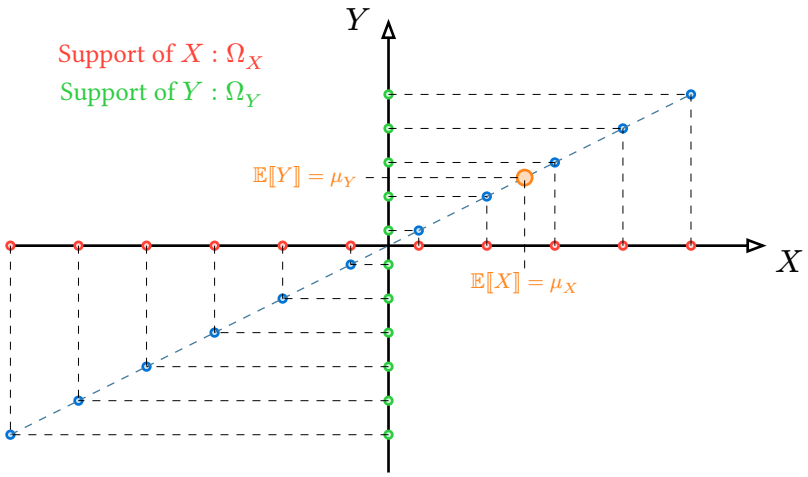


Figure 2.7: Support of two random variables X and Y that are positively correlated

Suppose now that we fix a value both for the random variable X and Y. We could try to compute the 'offset' of those values from their respective expected values, that is:

$$X - \mathbb{E}[\![X]\!] \quad \text{and} \quad Y - \mathbb{E}[\![Y]\!]$$

It is evident that, if we fix a point (x, y) in the support of the random vector (X, Y), and multiply the offsets define above and we weigh them by the joint probability of (x, y), will always get a positive quantity; that is

$$\text{Cov}\{X, Y\} \quad = \sum_{(x,y)\in\Omega_{XY}} (x - \mathbb{E}[\![X]\!])\ (y - \mathbb{E}[\![Y]\!])\ p_{XY}(x, y) > 0$$

In other word, X and Y are **positively correlated**: this means that when X is above its expected value, also Y tends to be above its expected value, and viceversa.

To be more specific, the one in Figure 2.7 is an example of **perfect linear relationship** meaning that, whenever we know the value of one of the two random variables, we can exactly determine the value of the other one.

> **Remark**
>
> It is possible to notice that the **magnitude** of the covariance depends on the **dispersion** of the two random variables, that is, the higher the variance of either one of the random variables, the higher the covariance will be.

Correlation Coefficient

We would actually like to be able to obtain a **standardized** way to measure the linear association between two random variables, in order to be able to compare the strength of the linear relationship between different pairs of random variables. To do so we can introduce the concept of **correlation coefficient**.

Definition 2.16 (Correlation Coefficient)

Given two random variables (X, Y) we can define the **correlation coefficient** between them ρ_{XY} as:

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}\{X, Y\}}{\sqrt{\text{Var}\{X\}}\ \sqrt{\text{Var}\{Y\}}} \tag{2.31}$$

One important property of the correlation coefficient is that it is always bounded between −1 and 1, that is:

$$-1 \leq \rho_{XY} \leq 1$$

♣

> **Warning**
>
> It is essential to notice that both covariance and correlation coefficient only measure **linear association** between two random variables. Two random variables may not be linearly correlated at all, but present some kind of non-linear dependency between them.

In our previous example of Figure 2.7, since the two random variables are perfectly positively correlated, we have that $\rho_{XY} = 1$.

To understand the previous warning consider two random variables whose support is represented in Figure 2.8.

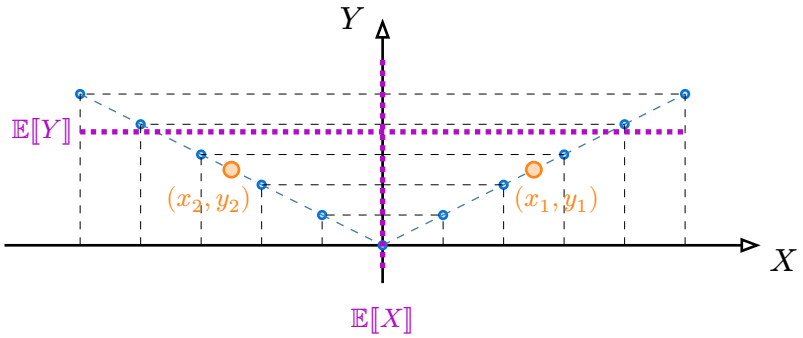


Figure 2.8: Support of two random variables X and Y that not linearly correlated

In this case we would have that the quantity $x_1 - \mathbb{E}[\![X]\!]$ would be positive, while the quantity $y_1 - \mathbb{E}[\![Y_1]\!]$ would be negative, thus the product between them would yield a negative value. The problem in this case is that if we consider the point (x_2, y_2), and consider the offset with respect to the expected values, we would find that their product is positive but with the same magnitude of the previous one. This would happen for every other pair of points in the support, thus we would obtain a linear correlation factor equal to zero.

We basically have found out that the two random variables are **uncorrelated**, meaning that there is no linear relationship between them, even though they are clearly **not independent**, indeed knowing the value of one of the two random variables allows us to know the value of the other one up to a sign.

In general, we can always say that if two random variables are independent, then they are also uncorrelated, and have covariance equal to zero. The opposite is not true in general, as we have just seen in the example above, in such case, we can only state that their relation, if any, is non-linear.

Properties of Covariance and Variance

As we have introduced a new operator, it is worth to take a look at its properties, specifically, since it is strongly related to the variance operator, not surprisingly we'll see how variance and covariance can be put together to derive some useful results.

- $\mathrm{Var}\{aX + bY + c\} = a^2\ \mathrm{Var}\{X\} + b^2\ \mathrm{Var}\{Y\} + 2ab\ \mathrm{Cov}\{X, Y\}$, here we can see that the first two members of the sum can be obtained from the scaling property of the variance, as well as the disappearance of the constant c.
- If we want to compute the covariance between linear transformations of multiple random variables we need to take into account all possible pairs of linearly transformed variable: $\mathrm{Cov}\{aX + bY, cZ + dW\} = ad\ \mathrm{Cov}\{X, W\} + bc\ \mathrm{Cov}\{Y, Z\} + ad\ \mathrm{Cov}\{X, W\} + bd\ \mathrm{Cov}\{Y, W\}$
- $\mathrm{Cov}\{X, a\} = \mathbb{E}[\![(x - \mathbb{E}[\![X]\!])(a - \mathbb{E}[\![a]\!])]\!] = \mathbb{E}[\![(x - \mathbb{E}[\![X]\!]) \cdot 0]\!] = 0$

- $\text{Cov}\{X, Y\} = \text{Cov}\{Y, X\}$, that is the covariance is **symmetric** as well as the correlation coefficient
- $\text{Cov}\{aX + b, bY + c\} = ac\ \text{Cov}\{X, Y\}$, in case we are dealing with the correlation operator we shall have that the constants a, c will cancel out with the scaling of the standard deviations in the denominator, that is $\rho(aX + b, cY + d) = \rho(X, Y)$

Chebyshev's Inequality

One last important property involving the variance of a random variable is the so called **Chebyshev's inequality**. This is a very useful result, which comes in handy when we don't know the exact distribution of a random variable, but we know its expected value μ and variance σ^2, which thinking about it are two values which is not so difficult to estimate from experimental data. In such case we can use this result.

Theorem 2.2 (Chebyshev's Inequality)

Given a random variable X with expected value $\mathbb{E}[\![X]\!] = \mu$ and variance $\text{Var}\{X\} = \sigma^2$, then for any $\varepsilon > 0$ we have that:

$$\mathbb{P}[\![|X - \mu| > \varepsilon]\!] \leq \left(\frac{\sigma}{\varepsilon}\right)^2 \tag{2.32}$$

♡

This result is very important since it allows us to bound the probability that a random variable deviates from its expected value by more than a certain amount ε, just knowing its variance.

3. Common Random Variable Distributions

In the last chapters we introduced the concept of random variables, and how to compute important quantities such as the expectation, variance and other important characteristics for those random variables, such as their c.d.f and p.d.f.

Even though from a theoretical point of view that is enough to compute everything we need about a random variable, in practice it is useful to that there are some random variables that behave in a very specific way and that we can use as building blocks to model more complex phenomena. In this chapter we will introduce some of the most common **families** of **random variables**, that is, groups of random variables that share some common characteristics and that can be used to model specific types of phenomena.

3.1. Bernoulli and Binomial Distributions

The simplest random variable distribution we can think about is the **Bernoulli distribution**.

Definition 3.1 (Bernoulli Distribution)

A random variable with two possible outcomes, 0 and 1 (usually representing *failure* and *success* respectively), is called a **Bernoulli random variable**, its distribution is a **Bernoulli distribution** and any experiment with a *binary outcome* is a Bernoulli **trial**.

The **sample space** of the random variable is given by $\Omega_X = \{0, 1\}$. The distribution is modeled by a *single parameter* p which represents the *probability of success* for the trial. Therefore the probability of a failure is $1 - p$.

♣

Probability Mass Function

Let X be a Bernoulli random variable with parameter p. The probability mass function (p.m.f.) of X is defined as follows:

$$p_X(x) = \begin{cases} q = 1 - p & \text{if } \ x = 0 \\ p & \text{if } \ x = 1 \end{cases} \tag{3.1}$$

Expected Value and Variance

Let X be a Bernoulli random variable with parameter p. Considering its probability mass function in Equation 3.1, we can compute its expected value.

$$\mathbb{E}[\![X]\!] = \sum_{x \in \Omega_X} x \cdot p_X(X) = 0 \cdot (1 - p) + 1 \cdot (p) = p \tag{3.2}$$

Given the expected value compute previously, we can also plug it into Equation 2.27 to compute the variance of a Bernoulli random variable.

$$\text{Var}\{X\} = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]^2 = p - p^2 = p(1 - p) \tag{3.3}$$

Now that we have defined all the important characteristics of a Bernoulli random variable, we can try to use is to model some more complex experiment. Suppose for example that we want to **replicate** a Bernoulli trial multiple times, say n and each of those trials is independent, this is how we get a **Binomial distribution**.

Definition 3.2 (Binomial Distribution)

A variable described as the number of successes Y in a sequence of independent Bernoulli trials $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, has **binomial distribution**. Its parameters are n, the number of trials, and p, the probability of success in each trial. ♣

Given n independent Bernoulli trials $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, we can define the random variable Y as the number of successes in those trials as $Y = \sum_{i=1}^{n} X_i$.

Probability Mass Function

Let X be a Binomial random variable with parameters n and p. The probability mass function (p.m.f.) of X is defined as follows:

$$p_X(x) = \binom{n}{x} \; p^x \; (1-p)^{n-x} \tag{3.4}$$

To get a better understanding of why the p.m.f. as defined as in Equation 3.4, we can think about the following:

- We need to have exactly x successes, which happens with probability p^x.
- We need to have exactly $n - x$ failures, which happens with probability $(1-p)^{n-x}$.
- The successes and failures can be arranged in any order, and there are $\binom{n}{x}$ ways to choose which x trials are successes out of n total trials.

Expected Value

The **expected value** of a Binomial random variable can be computed using the linearity of expectation as follows:

$$\mathbb{E}[\![X]\!] = \mathbb{E}\left[\!\!\left[\sum_{i=1}^{n} X_i\right]\!\!\right] = \sum_{i=1}^{n} \mathbb{E}[\![X_i]\!] = n \cdot p \tag{3.5}$$

where we used the fact that each X_i is a Bernoulli random variable with parameter p, therefore its expected value is p as shown in Equation 3.2. As far as the **variance** is concerned, we can compute it in the following way:

$$\text{Var}\{X\} = \text{Var}\left\{\sum_{i=1}^{n} X_i\right\} = \sum_{i=1}^{n} \text{Var}\{X_i\} = n \cdot p \cdot (1-p) = npq \tag{3.6}$$

Notice that we could use the fact that the X_i are independent to compute the variance of their sum as the sum of their variances, as the last property of the last chapter states.

R Implementation

In R we have the following functions to work with Binomial random variables at our disposal:

- `dbinom(x, n, p)` $= \mathbb{P}[\![X = x]\!]$, that is the probability mass function (p.m.f.).
- `pbinom(x, n, p)` $= \mathbb{P}[\![X \leq x]\!]$, that is the cumulative distribution function (c.d.f.).

- `qbinom(q, n, p)` $= x$ if $\mathbb{P}[\![X \leq x]\!] = q$, that is the quantile function.
- `rbinom(r, n, p)` $= \{x_1, x_2, ..., x_r\}$, that is a vector of r random samples drawn from the distribution.

> **Remark**
>
> All `R` functions that allow us to work with any common random variable distribution follow the same naming convention, where the first letter indicates the type of function (`d` for p.m.f./p.d.f., `p` for c.d.f., `q` for quantile function and `r` for random sampling), followed by the name of the distribution.

3.2. Multinomial Distribution

After introducing the Bernoulli and Binomial distributions, we can now generalize those concepts to the **Multinomial distribution**. If the experiments we are modeling are binary there are only two possible outcomes and we can model a repetition of them by means of the binomial. In case the experiments have *more than two possible outcomes*, say k we need to use the multinomial distribution.

Definition 3.3 (Multinomial Distribution)

A random variable described as the counts of each outcome in a sequence of independent trials with k possible outcomes, has **multinomial distribution**. Its parameters are n, the number of trials, and $p_1, p_2, ..., p_k$, the probabilities of each outcome in each trial, such that $\sum_{i=1}^{k} p_i = 1$.

♣

Probability Mass Function

Let X_i be the number of times outcome i occurs in n independent trials, each with k possible outcomes. The random vector $(X_1, X_2, ..., X_k)$ has **joint multinomial distribution** with probability mass function (p.m.f.) defined as follows:

$$\mathbb{P}[\![X_1 = x_1 \ X_2 = x_2 \ ... \ X_k = x_k]\!] = \frac{n!}{x_1!...x_k!} \ p_1^{x_1} \ ... \ p_k^{x_k} \tag{3.7}$$

where we implicitly have the constraints that $\sum_{i=1}^{k} x_i = n$ and we have introduced the **multinomial coefficient** $\frac{n!}{x_1!...x_k!}$ which counts the number of ways to arrange n trials with x_i occurrences of outcome i for each i. Of course we also need to have that the values $x_i \geq 0$.

It is always possible to transform a multinomial distribution into a bunch of binomial distributions by considering each outcome separately. To do so, we just need to focus on one of the k outcomes at a time, where the success is getting that specific outcome, and the failure is getting any of the other $k - 1$ outcomes.

Expected Value, Variance and Covariance

By focusing solely on the outcome i, since the experiments are Bernoulli trials with success probability p_i, we can use the results we obtained for the Binomial distribution to compute the expected value and variance of the random variable X_i as follows:

$$\mathbb{E}[\![X_i]\!] = p_1 \qquad \text{Var}\{X_i\} = n \cdot p_i \cdot (1 - p_i)$$

Since we are dealing with a vector of random variables, we can also compute the **covariance** between any two random variables X_i and X_j with $i \neq j$ as follows:

$$\text{Cov}\{X_i, X_j\} = -n \cdot p_i \cdot p_j \tag{3.8}$$

Intuitively this negative covariance makes sense, since if the count of outcome i increases, the count of outcome j must decrease, given that the total number of trials n is fixed.

R Implementation

Since the multinomial distribution is a joint distribution over multiple random variables, in `R` we only have two functions to work with it:

- `dmultinom`: the joint probability density function (p.m.f.)
- `rmultinom`: the function to generate random samples from the distribution.

We don't have a specific function for the cumulative distribution function (c.d.f.) or the quantile function; they are indeed very hard to define and manage for joint distributions.

3.3. Geometric Distribution

Another common random variable distribution is the **Geometric distribution**, it is again very much related to the Bernoulli distribution.

Definition 3.4 (Geometric Distribution)

A random variable that models the number of Bernoulli trials needed to get the first success, has **Geometric distribution**. Its parameter is p, the probability of success in each trial.

♣

Probability Mass Function

Let X be a Geometric random variable with parameter p. The probability mass function (p.m.f.) of X is defined as follows:

$$\mathbb{P}[\![X = x]\!] = (1 - p)^{x-1} \cdot p \tag{3.9}$$

where x can take any positive integer value, that is $x \in \{1, 2, 3, ...\}$. The rational behind this formulation is that the have the first success at trial x we need to have $x - 1$ failures, each of which happens with probability $1 - p$, and a success, that happens with probability p.

Expected Value and Variance

Let X be a Geometric random variable with parameter p. Considering its probability mass function in Equation 3.9. To compute its **expected value**, suppose we can write the random variable X as follows:

$$X = \sum_{i=1}^{\infty} I_i$$

Where I_i is an indicator that **at least** i trials are needed to get the first success. We can compute the expected value of X as follows:

$$\mathbb{E}[\![X]\!] = \mathbb{E}\left[\!\!\left[\sum_{i=1}^{\infty} \mathbb{E}[\![I_i]\!]\right]\!\!\right] = \sum_{i=1}^{\infty} \mathbb{P}[\![X \geq i]\!]$$

But $\mathbb{P}[\![X \geq i]\!]$ is the probability that the first $i-1$ trials are failures, so $\mathbb{P}[\![X \geq i]\!] = (1-p)^{i-1}$ therefore we have:

$$\mathbb{E}[\![X]\!] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \sum_{j=0}^{\infty} (1-p)^{j} = \frac{1}{1-(1-p)} = \frac{1}{p} \tag{3.10}$$

Notice how we use the formula for the convergence of a **geometric series** to compute the final result, this is why this distribution is called **geometric**. As far as the **variance** is concerned, we can compute it in the following way:

$$\text{Var}\{X\} = \frac{1-p}{p^2} \tag{3.11}$$

> **Remark**
>
> Notice how the expected value of a Geometric random variable has the formulation in Equation 3.10 by no surprise: the more the probability of success p increases, the less trials we expect to need in order to get the first success.

Memoryless Property

Until now we have not mentioned the cumulative distribution function (c.d.f.) of this random variable. However, it is interesting to notice that the c.d.f. of a Geometric random variable has a very special property, called the **memoryless property**.

Imagine that we have already performed at least y trials of an Bernoulli experiment without getting a success. The probability that we are going to *keep going* for at least another x trials without getting a success can be modeled with in the following way:

$$\mathbb{P}[\![X > x + y \mid X - y]\!] = \mathbb{P}[\![X > x]\!] \tag{3.12}$$

In other words, the probability of needing more than $x + y$ trials given that we have already performed y trials without success is equal to the probability of needing more than x trials from scratch. This property is called **memoryless** because the process does not care about what happened in the past, it only cares about the present situation.

R Implementation

Before understanding how `R` provides us with functions to work with this kind of random variable, it is crucial to understand that there are two different conventions to define this random variable.

Previously we defined a geometric random variable as the number of trials needed in order to observe a success. However, it is also common to define it as the number of **failures before** the success. In the first case we have $\Omega_X = \{1, 2, ...\}$, whilst in the second case we have $\Omega_X = \{0, 1, 2, ...\}$, since we can have zero failures before the first success.

The second one is exactly the convention that `R` uses, therefore all the functions we are going to introduce now are based on that definition. To switch from the second definition to the first one, it is necessary to first transform the random variable X into the random variable $X = Y + 1$. Similarly we'll have that:

$$\mathbb{P}[\![X = x]\!] = \mathbb{P}[\![Y = x - 1]\!]$$

In `R` we have the following functions to work with Geometric random variables:

- `dgeom(x-1, p)` $= \mathbb{P}[\![X = x]\!]$
- `pgeom(x-1, p)` $= \mathbb{P}[\![X \leq x]\!]$
- `qgeom(q, p)` $= x - 1$if $\mathbb{P}[\![X \leq x]\!] = q$
- `rgeom(r, p)` simulates r realizations of $X - 1$

3.4. Hyper-geometric Distribution

Another important random variable distribution is the **hypergeometric distribution**, which is used to model experiments where we draw samples without replacement from a finite population.

Definition 3.5 (Hypergeometric Distribution)

A random variable that models the number of successes in a sample of size n drawn **without replacement** from a population of size N containing M successes and $N - M$ failures has **hypergeometric distribution**.

♣

Probability Mass Function

Let X be a hypergeometric random variable with parameters N (population size), M (number of successes in the population), M (number of failures), n (the sample size). The probability mass function (p.m.f.) of X is defined as in the following equation:

$$\mathbb{P}[\![X = x]\!] = \text{hyper geom}(x, n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{3.13}$$

where x is an integer such that $\max(0, n - N + M) \leq x \leq \min(n, M)$

Expected Value and Variance

Let X be a hypergeometric random variable with p.m.f. given by hyper $\text{geom}(x, n, N, M)$, then we can de fine its expected value and variance as follows:

$$\mathbb{E}[\![X]\!] = n \cdot \frac{M}{N} \qquad \text{Var}\{X\} = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right) \tag{3.14}$$

3.5. Introduction to Stochastic Processes

In this section we will try to build a link between everything we have seen so far about random variables and basic probability theory, and the core concept of this course: **stochastic processes**. A sequence $\{X_n\}$ of random variables is a **stochastic process**. With the term "sequence" we refer to an *infinite random vector*.

If we consider a ***finite collection*** of random variables $\{X_1, X_2, ..., X_n\}$ we can characterize all we need to know about such random variables and their relationships by means of their **joint probability distribution**. Indeed starting from it we can compute all the marginal probabilities; furthermore we can also notice that $\forall i_1, i_2, ..., i_k$ and for $k \geq 1$ we can compute the joint probability of $(X_{i1}, X_{i2}, ..., X_{ik})$ by integrating (or summing) out all the other variables from the joint distribution.

If we can do this for every possible finite subset of r.v.'s from our infinite collection, that means we know the **law** (which is the distribution in this context of random processes) of the random sequence. Informally speaking, we can say that if $X = \{X_n\}_{n=1}^{\infty}$ the **law of X** is defined as the collection of all the *finite-dimensional distributions* $\forall n \in \{1, 2, 3, ...\}$. Given any subset of indices $i_1, i_2, ..., i_k$ with $k \geq 1$, the finite-dimensional distribution is defined as the joint distribution of the random variables $(X_{i1}, X_{i2}, ..., X_{ik})$.

The simplest stochastic process we can think about is a collection $X_i \overset{\text{i.i.d.}}{\sim} F_X \quad i = 1, 2, ...$ A finite subset of them is called a **sample** from distribution F_X. The reason why it is the simplest is given in the following equation:

$$\forall n, \forall i_1, i_2, ..., i_n : \mathbb{P}\left[\left(X_{i_1}, ..., X_{i_n}\right)\right] = F_{X_{i1},...X_{in}}(x_{i1}, ..., x_{in}) = \prod_{j=1}^{n} F_X(x_{ij}),$$

That is, the joint distribution of any finite subset of them can be computed as the product of their marginal distributions, since they are all **independent** and **identically distributed**.

> **Remark**
>
> If the random variables are independent but not identically distributed we need to know the **marginal distribution** for each one of the random variables. Namely, if $X_i \overset{\text{ind}}{\sim} F_{X_i}$ then we have that the joint probability of the sample is given by:
>
> $$F_{X_{i1},...X_{in}}(x_{i1}, ..., x_{in}) = \prod_{j=1}^{n} F_{X_{ij}}(x_{ij})$$
>
> Namely, we need to have knowledge about a countable number of marginal distributions.

> **Example: Sequence of independent non identically distributed random variables**
>
> Suppose we are dealing with a sequence of independent random variables which are not **identically distributed**. To keep the matter simple, let's suppose that the distribution changes according to the index of the random variable in the sequence and the basic distribution is always a Bernoulli distribution, that is: $X_i \sim \text{Bern}\left(\frac{1}{i}\right)$.

Of course, considering everything we have said so far, we can say that $\{X_n\}_{n=1}^{\infty}$ is a stochastic process.

Consider now the following object:

$$Y_n = \sum_{i=1}^{n} X_i \text{ Bin}(n, p))$$

And consider the collection $\{Y_n\}_{n=1}^{\infty}$; that one is also a **stochastic process**. Again, in this case Y_i's are surely **not identically distributed**, indeed if $n \neq m$, Y_m and Y_n have a different distribution while both being Binomial random variable. As far as independency is concerned we can take a look at the following equation:

$$Y_{n+1} = \sum_{i=1}^{n+1} X_i = Y_n + X_{n+1}$$

if we try to study the value of Y_{n+1} alone we can correctly conclude that it may take any value in {0, ..., n+1}; however if we consider $Y_{n+1} \mid Y_n = n$ we can easily see that Y_{n+1} can only take the values in {n, n+1}, thus Y_{n+1} and Y_n are **not independent**. Indeed the conditional distribution of $Y_{n+1} \mid Y_n$ is given by:

$$p_{Y_{n+1} \mid Y_n}(y_{n+1} \mid y_n) = \begin{cases} 1-p & \text{if } y_{n+1} = y_n \\ p & \text{if } y_{n+1} = y_n + 1 \\ 0 & \text{otherwise} \end{cases}$$

That is actually quite trivial to compute since we are dealing with Binomial random variables built from independent Bernoulli trials. In **general**, supposing we are working with Binomial random variables, we have that:

$$p_{Y_1, Y_2, \ldots, Y_n}(y_1, y_2, \ldots, y_n) = p_{Y_1}(y_1)\ p_{Y_2 \mid Y_1}(y_2 \mid y_1) \cdots\ p_{Y_n \mid Y_{n-1}}(y_n \mid y_{n-1})$$

Suppose that we know $Y_n = y_n$ and $Y_{n-1} = y_{n-1}$, let's see how we can use this information:

$$Y_{n+1} = Y_n + X_{n+1}$$

Basically, the first information is very useful since it tells us how many successes we had up to trial n, whilst the second information is just telling us that we can write $Y_n = y_{n-1} + X_n$, but we already know the value of Y_n so that second piece of information is not really adding anything new in case we already know Y_n.

> ⚠ **Warning**
>
> This does not mean, by any means, that Y_{n+1} and Y_{n-1} are independent. Indeed the value of Y_{n+1} is very much dependent on the value of Y_{n-1}: $Y_{n+1} = Y_{n-1} + X_n + X_{n+1}$. To be more precise, we can also write the conditional probability of $Y_{n+1} \mid Y_{n-1}$ as follows:

$$p_{Y_{n+1} \mid Y_{n-1}}(y_{n+1} \mid y_{n-1}) = \begin{cases} (1-p)^2 & \text{if } y_{n+1} = y_{n-1} \\ 2(1-p)p & \text{if } y_{n+1} = y_{n-1}+1 \\ p^2 & \text{if } y_{n+1} = y_{n-1}+2 \\ 0 & \text{otherwise} \end{cases}$$

What we can say about Y_{n+1} and Y_{n-1} is that they are **conditionally independent** given Y_n, this is very useful because it allows us to simplify the computation of joint probabilities.

If we now try to look at the joint probabilities we may be interested in,we can use what we have just observed to write the following:

$$\begin{aligned} p_{Y_1,\dots,Y_n}(y_1,\dots,y_n) &= p_{Y_1}(y_1) \cdot p_{Y_2 \mid Y_1}(y_2 \mid y_1) \cdot p_{Y_3 \mid Y_2}(y_3 \mid y_2)\dots \\ &= p_{Y_1}(y_1) \prod_{i=1}^{n-1} p_{Y_{i+1} \mid Y_i}(y_{i+1} \mid y_i) \end{aligned} \tag{3.15}$$

If each X_i is the result of a coin toss we can model a Y_n as the *number of wins* in the first n throws we can model a Y_n as the *number of wins* in the first n throws. For the first toss we are going to have the following:

$$p_{Y_1}(y_1) = \begin{cases} 1-p & \text{if } y_1 = 0 \\ p & \text{if } y_1 = 1 \end{cases}$$

This is straightforward since Y_1 is just a Bernoulli random variable. For the second toss we have:

$$p_{Y_2}(y_2) = \begin{cases} (1-p)^2 & \text{if } y_2 = 0 \\ 2(1-p)p & \text{if } y_2 = 1 \\ p^2 & \text{if } y_2 = 2 \end{cases}$$

We can derive this result by noticing that $Y_2 \sim \text{Binom}(2,p)$. Given these pieces of information we can compute several different probabilities. For instance $\mathbb{P}[\![Y_1 = 1]\!]$, $\mathbb{P}[\![Y_n = 1]\!]$. But what about the probability of getting a win in the first throw and only lose in the next 6? This can be modeled by the following equation which leverages Equation 3.15:

$$\begin{aligned} \mathbb{P}[\![Y_1 = 1 \wedge Y_2 = 1 \wedge \dots \wedge Y_7 = 1]\!] &= p_{Y_1}(y_1) \cdot p_{Y_2 \mid Y_1}(1 \mid 1) \\ &\quad \cdot \dots \cdot p_{Y_7 \mid Y_6}(1 \mid 1) \\ &= p \cdot \prod_{i=1}^{7-1} p_{Y_{i+1} \mid Y_i}(1 \mid 1) \\ &= p \cdot \prod_{i=1}^{6} (1-p)^2 = p(1-p)^{12} = \frac{1}{2}\left(\frac{1}{2}\right)^{12} = \frac{1}{2^{13}} \end{aligned}$$

Definition 3.6 (Markov Process)

A sequence $X = \{X_n\}_{n=1}^{\infty}$ where each $X_n + 1$ is **conditionally independent** of $\{X_{n-1} \dots X_1\}$ given X_n is called a **Markov process** or **Markov chain**, which is a

stochastic process with some interesting properties that make it easier to study and analyze.

♣

Stochastic Processes and Common Random Variables

Let's now try to review the concept of **geometric** distribution we have already seen before but in a stochastic process fashion. Consider the random variables $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$ and consider the following:

$$W_1 = \min\{n : X_n = 1\}$$

Clearly we can see $W_1 \sim \text{Geom}(p)$. And we can sort of consider its value as the 'expected waiting time' until the first success. We can also consider the random variable Y_n as the sum of all the X_i up to time n.

We can define two random sequences as follows:

- $W = \{W_n\}$ as the sequence of **waiting times** between changes in the process (or in the value of Y_n). We also refer to them as **inter-arrival times**.
- $Y = \{Y_n\}$ as the sequence that counts the number of successes up to time n, basically it is a **counting process**.

We can actually do the same for **hyper-geometric** random variables. Suppose we have for instance N balls in an urn, of which M are successes (say red) and $N - M$ are failures. The probability of success when drawing a ball at random from the urn is $p = \frac{M}{N}$. So if we consider the first ball extraction we obtain the following:

$$X_1 \sim \text{Bern}\left(p_1 = \frac{M}{N}\right)$$

Now, if we move to the second extraction, we still have a Bernoulli random variable, but we don't know its parameter, since it depends on the result of the first extraction. We can consider two cases:

- $X_2 \mid X_1 = 1 \sim \text{Bern}\left(\frac{M-1}{N-1}\right)$, in case the first extraction was a success.
- $X_2 \mid X_1 = 0 \sim \text{Bern}\left(\frac{M}{N-1}\right)$, in case the first extraction was a failure.

If we now take a look at X_3, we can notice that this gets more complicated:

- $X_3 \mid X_2 = 1, X_1 = 1 \sim \text{Bern}\left(\frac{M-2}{N-2}\right)$, in case the first two extractions were successes.
- $X_3 \mid X_2 = 1, X_1 = 0 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$, in case the first extraction was a failure and the second a success.
- $X_3 \mid X_2 = 0, X_1 = 1 \sim \text{Bern}\left(\frac{M-1}{N-2}\right)$, in case the first extraction was a success and the second a failure.
- $X_3 \mid X_2 = 0, X_1 = 0 \sim \text{Bern}\left(\frac{M}{N-2}\right)$, in case the first two extractions were failures.

We can notice that, if we take in consideration the **sum** of the successes up to time n, we can actually define the conditional distribution based solely on the value of that sum, not the individual outcomes of each previous extraction. We say that, in general, X_{n+1} is **conditionally independent** on $\{X_n, X_{n-1}, ..., X_1\}$ given $Y_n = \sum_{i=1}^{n} X_i$.

3.6. Negative Binomial Distribution

Earlier in this chapter we have been talking about binomial distributions and how they are related to the geometric distribution, which measures the 'waiting time' until the first success in a sequence of Bernoulli trials.

Definition 3.7 (Negative Binomial Distribution)

Given a sequence of independent Bernoulli trials with success probability p, we can model the **number of trials** to obtain k **successes** with a **Negative Binomial distribution**. Suppose we have $W_i \overset{\text{i.i.d.}}{\sim} \text{Geom}(p)$. The random variable

$$N_k = \sum_{i=1}^{k} W_i \tag{3.16}$$

follows a **Negative Binomial distribution** with parameters k and p.

♣

Probability Mass Function

Let N_k be a Negative Binomial random variable with parameters k and p. The probability mass function (p.m.f.) of N_k is defined as follows:

$$p_X(x) = \binom{x-1}{k-1} (1-p)^{x-k} \, p^k \quad \forall x \geq k \tag{3.17}$$

intuitively, this formulation makes sense: to have the k-th success at trial x we need to have $k-1$ successes in the first $x-1$ trial (which can happen in $\binom{x-1}{k-1}$ ways). Every combination $k-1$ successes and $x-k$ failures happens with probability $p^{k-1}(1-p)^{x-k}$; finally we need to have a success at trial x, which happens with probability p.

Expected Value and Variance

Let N_k be a Negative Binomial random variable with parameters k and p. Considering its probability mass function in Equation 3.17, we can compute its expected value and variance as follows:

$$\mathbb{E}[\![N_k]\!] = \frac{k}{p} \qquad \text{Var}\{N_k\} = k\frac{1-p}{p^2} \tag{3.18}$$

R Implementation

Before introducing the `R` functions to work with Negative Binomial random variables, it is important to notice that there are two different conventions to define this random variable, `R` actually uses a different one with respect to the one we have just introduced, similarly to what happened with the Geometric distribution, we will need to adjust the parameters accordingly:

- `dnbinom(x - k, k, p)` $= \mathbb{P}[\![X = x]\!]$, is the probability mass function (p.m.f.).
- `pnbinom(x - k, k, p)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qnbinom(q, k, p) + k` $= x$ "if" $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rnbinom(r, k, p)` simulates r realizations of $X - k$.

To switch from the `R` definition to the one we have introduced, it is necessary to first transform the random variable X into the random variable $X = Y + k$.

3.7. Uniform Distribution

In the past few sections we have been focusing our attention on **discrete random variables**, but as we know, there are also **continuous random variables**.

Definition 3.8 (Uniform Distribution)

A random variable that has an equal probability of taking any value within a given interval $[a, b]$ has **Uniform distribution**. Its parameters are a and b, the endpoints of the interval. If the interval of values if $[0, 1]$ we say that the random variable has a **standard uniform distribution**.

♣

Probability Density Function

Let X be a Uniform random variable with parameters a and b. The probability density function (p.d.f.) of X is defined as follows:

$$f_X(x) = \frac{1}{b-a} \qquad \forall x \in [a, b] \tag{3.19}$$

Actually the above definition is not really precise, indeed when we talk about continuous random variables we cannot really talk about probability, rather we need to talk about **density**. The reason because uniform distributions are so important is that they are the building blocks for all other random variable distributions.

> **Remark**
>
> In order for Equation 3.19 to be valid, it is necessary that the value $|b - a|$ is a finite positive number so that there is no chance of dividing by zero or by infinity. The rational behind this is quite simple, if we try to choose a random number in an interval of infinite length, we cannot do it with uniform probability since the density would be zero everywhere.

Uniform Property

For any $h > 0$ and $t \in [a, b - h]$ we have that:

$$\mathbb{P}[\![t < X < t + h]\!] = \int_t^{t+h} \frac{1}{b-a}\, dx = \frac{h}{b-a}$$

is **independent of t**. The probability is only determined by the length of the interval not buy the location of the point in the interval.

Expected Value and Variance

Let X be a Uniform random variable with parameters a and b. Considering its probability density function in Equation 3.19, we can compute its expected value and variance as follows:

$$\mathbb{E}[\![X]\!] = \frac{a+b}{2} \qquad \text{Var}\{X\} = \frac{(b-a)^2}{12} \tag{3.20}$$

It is by no surprise that the expected value of a uniform random variable is the midpoint of the interval $[a, b]$. As far as the variance is concerned, we can notice that it increases quadratically with the length of the interval. If we consider the standard uniform distribution, that is $a = 0$ and $b = 1$, we have that $\mathbb{E}[\![X]\!] = \frac{1}{2}$ and $\text{Var}\{X\} = \frac{1}{12}$.

Uniform Distribution Transformation and Standardization

One very common operation when working with this kind of random variable is to transform it into a standard uniform random variable and vice-versa. Consider two random variables $X \sim \text{Uniform}(a, b)$ and $Y \sim \text{Uniform}(0, 1)$. We can transform X into Y and vice-versa as follows:

$$\begin{aligned} Y &= \frac{X-a}{b-a} \sim \text{Uniform}(0,1) \\ X &= a + (b-a)Y \sim \text{Uniform}(a,b) \end{aligned} \tag{3.21}$$

Let's now consider the following new random variable:

$$Z = \frac{X - \mathbb{E}[\![X]\!]}{\sqrt{\text{Var}\{X\}}} \sim \text{Uniform}(z_l, z_u)$$

Abd suppose we want to compute its expected value and variance. To do so we need to compute the values of z_l and z_u first:

$$x = a \quad \Rightarrow \quad z_l = \frac{a - \mathbb{E}[\![X]\!]}{\sqrt{\text{Var}\{X\}}} = \frac{a - \frac{a+b}{2}}{\sqrt{\frac{(b-a)^2}{12}}}$$

$$x = b \quad \Rightarrow \quad z_u = \frac{b - \mathbb{E}[\![X]\!]}{\sqrt{\text{Var}\{X\}}} = \frac{b - \frac{a+b}{2}}{\sqrt{\frac{(b-a)^2}{12}}}$$

Now that we have these values it is easy to notice that $\mathbb{E}[\![Z]\!] = 0, \quad \text{Var}\{Z\} = 1$.

Definition 3.9 (Standardization)

Given **any** discrete or continuous random variable X with expected value $\mathbb{E}[\![X]\!] = \mu$ and variance $\text{Var}\{X\} = \sigma^2$, we can define the **standardized** random variable Z as follows:

$$Z = \frac{X - \mu}{\sigma^2} \tag{3.22}$$

which satisfies $\mathbb{E}[\![Z]\!] = 0$ and $\text{Var}\{Z\} = 1$.

It is actually easy to see why the expected value and variance of Z are as we have just said:

$$\mathbb{E}[\![Z]\!] = \mathbb{E}\left[\!\!\left[\frac{X-\mu}{\sqrt{\sigma^2}}\right]\!\!\right] = \frac{1}{\sigma}(\mathbb{E}[\![X]\!] - \mu) = 0$$

$$\text{Var}\{Z\} = \text{Var}\left\{\frac{X-\mu}{\sqrt{\sigma^2}}\right\} = \frac{1}{\sigma^2}(\text{Var}\{X\} - 0) = 1 \quad \blacksquare$$

The set of possible values of Z and X are different. For instance, consider $X \sim \text{Bern}(p)$ with $\Omega_X = \{0, 1\}$. The standardized random variable Z can now be built as follows:

$$Z = X - \frac{p}{\sqrt{p(1-p)}}$$

If we now take a look at the possible values of Z we have: $\Omega_Z = \left\{-\frac{p}{\sqrt{p(1-p)}}, \frac{1-p}{\sqrt{p(1-p)}}\right\}$ respectively when $X = 0$ and $X = 1$. Clearly Z is **not a Bernoulli**. This tells us a very important fact about standardization.

> ⚠ **Warning**
>
> In general, a **standardized random variable** does not belong to the *same family* of the original random variable that was used to build the standardization

Following we have a theorem which tells us something very important about standardization and uniform random variables:

Theorem 3.1 (Standardization of Uniform Random Variables)

Given any uniform random variable $X \sim \text{Uniform}(a, b)$, it is **closed under linear transformation**, that is the uniformness of the random variable is preserved under any linear transformation, including **standardization**.

> ⚠ **Warning**
>
> Even though the name suggests it, the **standard uniform random variable** $Y \sim \text{Uniform}(0, 1)$ is just a special case of uniform random variable. It is **not** the result of a standardization process.

R Implementation

In `R` we have the following functions to work with Uniform random variables:

- `dunif(x, a, b)` $= f_{X(x)}$, is the probability density function (p.d.f.).
- `punif(x, a, b)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qunif(q, a, b)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `runif(r, a, b)` simulates r realizations of X.

3.8. Normal (Gaussian) Distribution

Although it is not the distribution we are going to preponderantly use in this course, the **Normal distribution** is so common and important in all probability theory that it is worth spending some time on it. If the uniform distribution serves to express the idea of 'equiprobability', the normal distribution is often used to model 'natural' phenomena.

Definition 3.10 (Normal Distribution)

A random variable that models phenomena where values tend to cluster around a central mean value with a certain variability has **Normal distribution**. Its parameters are μ (the mean) and σ^2 (the variance).

♣

When dealing with this kind of random variables we often refer to the mean as **location parameter** and to the variance as **scale parameter**.

Probability Density Function

Let X be a Normal random variable with parameters μ and σ^2. The probability density function (p.d.f.) of X is defined as follows:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \tag{3.23}$$

We can see how this formulation intuitively makes sense. The numerator of the fraction in the exponential is squared so that larger errors are more penalized (i.e., less likely) w.r.t. smaller errors. The numerator is then divided by the variance so that larger variances lead to less penalization for larger errors. Finally the whole expression is normalized by the factor $\frac{1}{\sigma\sqrt{2\pi}}$ so that the total area under the curve is equal to 1.

We can see that the value of μ serves to control the **location** of the distribution's peak, whilst the value of σ^2 serves to control the **spread** of the distribution around the mean. This is illustrated in Figure 3.1.

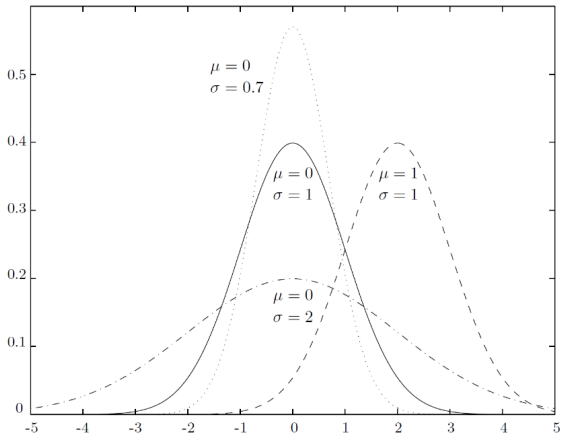


Figure 3.1: Normal distributions with different parameters

Standardization

There is actually no point in computing the expected value and variance of a Normal random variable in that they are exactly equal to the parameters used to define the distribution: $\mathbb{E}[\![X]\!] = \mu$ and $\text{Var}\{X\} = \sigma^2$.

Nevertheless, it is possible to define a **standard normal random variable** Z such as it has expected value equal to 0 and variance equal to 1. Along with this fact, it is interesting to notice that Normal random variables are **closed under linear transformation**, that is if we take any Normal random variable and we apply a linear transformation to it, the resulting random variable is still Normal. In particular we can notice the following:

$$X = aZ + b \sim \text{Normal}(a\mathbb{E}[\![Z]\!] + b, a^2 \ \text{Var}\{Z\}) \tag{3.24}$$

and by simply plugging the knowledge that $\mathbb{E}[\![Z]\!] = 0$ and $\text{Var}\{Z\} = 1$ in the above equation we have that $X \sim \text{Normal}(b, a^2)$.

Transformation from and to Standard Normal

Given any Normal random variable $X \sim \text{Normal}(\mu, \sigma^2)$ and a standard normal random variable $Z \sim \text{Normal}(0, 1)$ we can transform X into Z and vice-versa as follows:

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1) \\ X &= \mu + \sigma Z \sim \text{Normal}(\mu, \sigma^2) \end{aligned} \tag{3.25}$$

This is indeed very similar to the standardization process we have already seen in the case of uniform random variables.

R Implementation

In `R` we have the following functions to work with Normal random variables:

- `dnorm(x, mu, sigma)` $= f_{X(x)}$, is the probability density function (p.d.f.).
- `pnorm(x, mu, sigma)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qnorm(q, mu, sigma)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rnorm(r, mu, sigma)` simulates r realizations of X.

3.9. Poisson Distribution

Let's now take a look at what is probably the most important distribution for this course: the **Poisson distribution**. Let's first take a look at its definition.

Definition 3.11 (Poisson Distribution)

The numbero of "**rare**" events occurring within a fixed interval of time has **Poisson Distribution**.

♣

This definition looks a bit vague in that we still need to clarify what we mean by "rare" events. Before doing so, let's first take a look at its probability mass function.

Probability Mass Function

Let $X \sim \text{Poisson}(\lambda)$ be a Poisson random variable with parameter $\lambda > 0$. The probability mass function (p.m.f.) of X is defined as follows:

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \forall x \in \{0, 1, 2, ...\} \tag{3.26}$$

Though this formulation may look strange, it is indeed a probability mass function. Indeed it is both positive for all x and it sums to 1.

Positivity

To understand why it is positive there is not much to say, all the components of the product in Equation 3.26 are positive for any $\lambda > 0$ and any $x \in \{0, 1, 2, ...\}$.

Normalization

To understand why it sums to 1 we can consider the definition of $f(\lambda) = e^{\lambda}$ as the limit of an infinite series:

$$e^{\lambda} = \sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \iff \sum_{x=0}^{\infty}e^{-\lambda}\frac{\lambda^x}{x!} = 1$$

where the series on the right-hand side is exactly the formulation of the p.m.f. in Equation 3.26. If we try to see this the other way around, we may wonder which is the constant factor k that makes the function $\frac{\lambda^x}{x!}$ be a proper p.m.f. We can find such a constant by solving the following equation:

$$1 = k \cdot \sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \iff k = \frac{1}{\sum\limits_{x=0}^{\infty}\frac{\lambda^x}{x!}} = e^{-\lambda}$$

Expected Value and Variance

We can try to compute the **expected value** of the Poisson distribution by leveraging again the Taylor series expansion of the exponential function:

$$\begin{aligned}
\mathbb{E}[\![X]\!] &= \sum_{x=0}^{\infty} x\ p_{X(x)} = \sum_{x=0}^{\infty} x\ e^{-\lambda}\frac{\lambda^x}{x!} \\
&= 0 + \sum_{x=1}^{\infty} x\ e^{-\lambda}\frac{\lambda^x}{x!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\lambda\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}\lambda e^{\lambda} = \lambda
\end{aligned} \tag{3.27}$$

The parameter λ of the Poisson distribution is called the **rate** or **frequency** parameter, since it represents the expected (mean) number of events per fixed amount of time.

Before actually computing the variance of the Poisson distribution it is necessary to compute another quantity. Specifically by recalling Equation 2.22 we can notice the following:

$$\begin{aligned}\mathbb{E}[\![X(X-1)]\!] &= \sum_{x=0}^{\infty} x(x-1)\ p_X(x) = \sum_{x=0}^{\infty} x(x-1)\ e^{-\lambda}\frac{\lambda^x}{x!} \\ &= 0 + 0 + e^{-\lambda}\lambda^2 \sum_{x=2}^{\infty} x(x-1)\frac{\lambda^{x-2}}{x(x-1)(x-2)!} = \lambda^2\end{aligned}$$

Now, if we notice that by linearity of expectation we can also write:

$$\mathbb{E}[\![X(X-1)]\!] = \mathbb{E}[\![X^2 - X]\!] = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]$$

Therefore we can combine the results above and conclude that $\mathbb{E}[\![X^2]\!] = \lambda^2 + \lambda$. Now we are finally ready to give an expression for the **variance**:

$$\text{Var}\{X\} = \mathbb{E}[\![X^2]\!] - (\mathbb{E}[\![X]\!])^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda \tag{3.28}$$

We can also generalize what we have seen before when we were computing $\mathbb{E}[\![X(X-1)]\!]$:

$$\mathbb{E}\left[\!\!\left[\prod_{i=0}^{k-1}(X-i)\right]\!\!\right] = \lambda^k$$

R Implementation

In `R` we have the following functions to work with Poisson random variables:

- `dpois(x, lambda)` $= p_X(x)$, is the probability mass function (p.m.f.).
- `ppois(x, lambda)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qpois(q, lambda)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rpois(r, lambda)` simulates r realizations of X.

Properties of Poisson Random Variables

In this section we are going to observe some important properties of Poisson random variables, which will come in very handy in the next few chapters.

Poisson Approximation of Binomial Distribution

In this section we are going to see how the Poisson distribution can be used to approximate a Binomial distribution when the number of trials considered is large and the probability of success p of those Bernoulli trials is small. This approximation is adequate say, for $n \geq 30$ and $p \leq 0.05$ and becomes more and more accurate as n increases and p decreases.

Theorem 3.2 (Law of Rare Events)

$$\lim_{\substack{n\to\infty, p\to 0 \\ np=\lambda}} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!} \tag{3.29}$$

The convergence presented in Theorem 3.2° is called **convergence in distribution**, which is not the same as the usual convergence we are used to.

In our case, this means that, as the number of Bernoulli trials n increases and the probability of success p decreases in such a way that their product np remains constant, say equal to λ, the

distribution of the Binomial random variable $X \sim \text{Binom}(n, p)$ approaches the distribution of the Poisson random variable $Y \sim \text{Poisson}(\lambda)$.

To make this more concrete, consider a sequence of random variables $X_n \sim \text{Binom}\left(n, \frac{\lambda}{n}\right)$, where $\frac{\lambda}{n} = p$, for some adequate value of λ, that is, $p = \frac{\lambda}{n} < 1$. We can notice that $np = \lambda$ but if $n \to \infty$ then $p = \frac{\lambda}{n} \to 0$. This means that the distribution of X_n approaches the distribution of $Y \sim \text{Poisson}(\lambda)$ as n increases. If $n \to \infty$ then $X_n \xrightarrow{\text{d}} Xl \sim \text{Poisson}(\lambda)$, where "$\xrightarrow{d}$" indicates *convergence in distribution*. To better understand this, it is useful to remember that each X_n can be seen as a function of ω: $X_n(\omega)$.

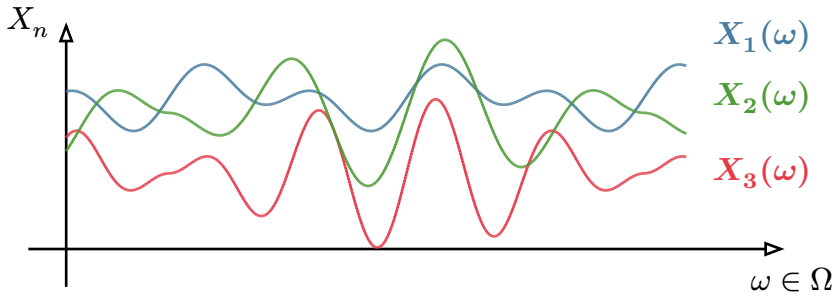


Figure 3.2: Different random variables X_n plots for the same outcome ω

Figure 3.2 shows that every time we perform the experiment, we get one outcome $\omega \in \Omega$ and each $X_n(\omega)$ gets its individual value x_n.

> ⚠ **Warning**
>
> Convergence in distribution **does not mean** that for each ω we have:
>
> $$X_n(\omega) = x_n \quad \text{and} \quad X(\omega) = x$$
>
> rather, it is interested in the **probability** of X_n taking values in certain intervals converging to the probability of X taking values in the same intervals as n goes to infinity.

Additivity of Poisson Random Variables

Another very important property of Poisson random variables is their **additivity**. Let's look at the following theorem:

Theorem 3.3 (Additivity of Poisson Random Variables)

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are two **independent** Poisson random variables, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

To view this under a more practical light, consider two disjoint periods of time π_1 and π_2, say $\pi_1 = [0, t_1), \pi_2 = [t_1, t_2]$. Suppose for each of these periods we define a Poisson random variable, for instance $X \sim \text{Pois}(\lambda)$ counts the number of rare events occurring during π_1 and $Y \sim \text{Pois}(\mu)$ counts number of rare events occurring in π_2; where X and Y represent respectively the fact that we are expecting to observe λ events during π_1 and μ events to happen during the span of π_2: $\mathbb{E}[\![X]\!] = \lambda, \mathbb{E}[\![Y]\!] = \mu$.

Let's now consider the period $\Pi = [0, t_2]$ and define the random variable W as the number or rare events occurring during Π, intuitively also this variable is a Poisson. If we also remember that the parameter of a Poisson random variable models the expected number of events occurring during the time period, we can intuitively say that it makes sense to expect to observe $\lambda + \mu$ events during the period Π. Therefore we can conclude that $W \sim \text{Pois}(\lambda + \mu)$.

> ⚠ **Warning**
>
> The additivity property of Poisson random variables **only holds** when the random variables considered are **independent**.

Let's now try to prove the theorem in a formal way. By the notion of independence we know that the following equation holds:

$$\begin{aligned}\mathbb{P}[\![X = r \wedge Y = s]\!] &= \mathbb{P}[\![X = r]\!]\mathbb{P}[\![Y = s]\!] \\ &= \lambda^r \frac{e^{-\lambda}}{r!}\ \mu^s \frac{e^{-\mu}}{s!}\end{aligned}$$

Now we actually need to consider all possible ways of obtaining $W = n$, that is, we need to consider all the pairs (r, s) such that $r + s = n$. Therefore we can write:

$$\begin{aligned}\mathbb{P}[\![X + Y = n]\!] &= \sum_{r=0}^{n} \mathbb{P}[\![X = r \wedge Y = n - r]\!] \\ &= \sum_{r=0}^{n} \frac{\lambda^r e^{-\lambda}}{r!} \frac{\mu^{n-r} e^{-\mu}}{(n-r)!}\end{aligned}$$

We can multiply and divide everything inside the sum by $n!$, this is useful since now we can bring out of the summation all the terms that do not depend on r and obtain the following:

$$\begin{aligned}\mathbb{P}[\![X + Y = n]\!] &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{r=0}^{n} \binom{n}{r} \lambda^r \mu^{n-r} \\ &= \frac{(\lambda + \mu)^n\ e^{-(\lambda+\mu)}}{n!}\end{aligned}$$

which is exactly the p.m.f. of a Poisson random variable with parameter $\lambda + \mu$. This can be easily generalized to the sum of k independent Poisson random variables by *mathematical induction*. It is actually possible to generalize this property even further: we can even consider the case in which there is a **infinite countable** number of independent Poisson random variables.

Theorem 3.4 (Generalized Additivity of Poisson Random Variables)

Let $X_j \sim \text{Pois}(\lambda_j)$ for $j = 1, 2, \ldots$ be a sequence of independent random variables. If we have that $\sum_{j=1}^{\infty} \lambda_j = \lambda < \infty$, i.e., the series converges, then we have that:

$$\mathbb{P}\left[\left[S = \sum_{j=2}^{\infty} X_j < \infty \right]\right] = 1 \quad \text{and} \quad S \sim \text{Pois}(\lambda)$$

that is, the infinite sum of independent Poisson r.v.'s is still a Poisson r.v. If, on the other hand, the series $\sum \lambda_j = \infty$ then also the probability that the infinite sum diverges is equal to 1.

♡

The type of convergence used by this theorem is called **almost sure convergence**, which is stronger than convergence in distribution. To better understand this, suppose we have a sequence $X_i \sim \text{Pois}(\lambda_i)$. We can define **partial sums**: $S_1 = X_1$, $S_2 = X_1 + X_2$, $S_3 = X_1 + X_2 + X_3$, and so on until S_n. Along with these random variables we can also define the partial sums of their parameters: $\mu_1 = \lambda_1$, $\mu_2 = \lambda_1 + \lambda_2$, $\mu_3 = \lambda_1 + \lambda_2 + \lambda_3$, and so on until μ_n, so that $S_n \sim \text{Pois}(\mu_n)$.

If we have that $\mu_n \underset{\infty}{\to} \mu < \infty$, then we have that $S_n \xrightarrow{\mathbb{P}} S \sim \text{Pois}(\mu)$, where "$\xrightarrow{\mathbb{P}}$" indicates **convergence in probability**.

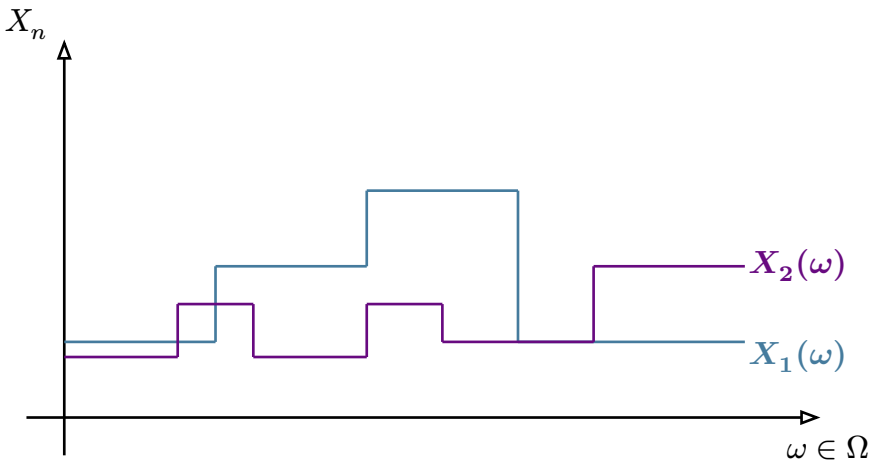


Figure 3.3: Two random variables X_1 and X_2 for the same outcome ω

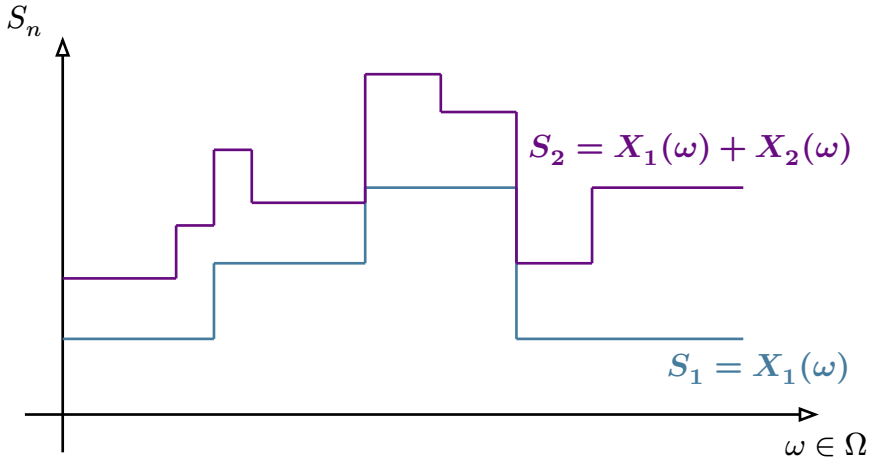


Figure 3.4: Partial sums S_1 and S_2 for the same outcome ω

Looking at the figures above and trying to consider a fixed value of ω, in Figure 3.3 the observed values of the sequence may not converge, but in Figure 3.4, for each ω the observed values $X_n(\omega) = x_n$ always converge, if n is large enough. The convergence works every time

we perform the experiment. This happens thanks to the strong dependence that exists between the partial sums.

Poisson - Multinomial Relationship

Up to this point we have talked about sums, basically we have seen that if we know the values of the X's then we can easily compute the value of their sums. Now we would like to invert this relation. Suppose we know the value of a sum of Poisson random variables, we'd like to be able to infer something about the values of the individual Poisson r.v.'s that are being summed up. Let's look at the following theorem.

Theorem 3.5 (Poisson - Multinomial Relationship)

Let $S_n = X_1 + ... + X_n$ be the sum of n independent Poisson random variables each with parameter λ_i and let $\lambda = \lambda_1 + ... + \lambda_n$. The **conditional distribution** of the vector $\boldsymbol{X} = (X_1, ..., X_n)$ given the value of S_n is **multinomial** with its parameter being $\boldsymbol{p} = (\lambda_1/\lambda, ..., \lambda_n/\lambda)$.

Intuitively, this makes sense, because if we know that the total number of events observed is k, then the only uncertainty the remains is about how these k events are distributed among the different X_i's. This is exactly what a multinomial distribution models.

To see this in a more formal way, suppose we have $r_1 + r_2 + ... + r_n = s$, then we can write:

$$\begin{aligned}\mathbb{P}[\![X_1 = r_1, \ ..., \ \ X_n = r_n \mid S_n = s]\!] &= \frac{\mathbb{P}[\![X_1 = r_1, \ ..., \ \ X_n = r_n \ , S_n = s]\!]}{\mathbb{P}[\![S_n = s]\!]} \\ &= \frac{\Pi_{j=1}^{n}\left(\lambda_j^{r_j}\frac{e^{-\lambda_j}}{r_{j!}}\right)}{\lambda^s\frac{e^{\lambda}}{s!}} = \frac{s!}{\pi_{j=1}^{n} r_{j!}}\left(\frac{\lambda_1}{\lambda}\right)^{r_1} ... \left(\frac{\lambda_n}{\lambda}\right)^{r_n}\end{aligned}$$

where the first equality comes from the definition of conditional probability in Equation 2.13 and the second equality is obtained by noticing that $S_n = s$ is a redundant condition once we have all the values of the X_i's and by multiplying the p.m.f.'s of the individual Poisson random variables. As far as the third equality is concerned the λ^s has been replaced by a product of $\lambda^{r_1} \cdot ... \cdot \lambda^{r_n} = \lambda$, and the other simplifications are highligted in green.

Remark

Notice that, in case $n = 2$, the multinomial distribution reduces to a *Binomial distribution*. Given $S_2 = s$, if $X_1 = r$ and $X_2 = s - r$, we have that:

$$\begin{aligned}\mathbb{P}[\![X_1 = r, \ X_2 = s - r \mid S_2 = s]\!] &= \mathbb{P}[\![X_1 = r \mid S_2 = s]\!] \\ &= \binom{s}{r} p^r (1-p)^{s-r}\end{aligned}$$

where $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

> **Remark**
>
> In a very similar fashion it is possible to do the opposite, that is, let $S \sim \text{Pois}(\lambda)$ and assume that, conditionally on S, X has a $\text{Binom}(S, p)$ distribution. Then X and $Y = S - X$ are **independent** Poisson random variables with parameters $\lambda_1 = \lambda p$ and $\lambda_2 = \lambda(1-p)$.

To see why this last remark is true, we can produce the following derivation:

$$\begin{aligned}\mathbb{P}[\![X = r,\, S - X = k]\!] &= \mathbb{P}[\![S = k + r]\!]\mathbb{P}[\![X = r \mid S = k + r]\!] \\ &= \frac{\lambda^{k+r} e^{-\lambda}}{(k+r)!}\binom{k+r}{r} p^r (1-p)^k \\ &= \frac{(\lambda p)^r e^{-\lambda p}}{r!}\ \frac{(\lambda(1-p))^k e^{-\lambda(1-p)}}{k!}\end{aligned}$$

Here, instead of starting from the conditional, and writing it as the ratio between the joint and the marginal, we have started from the joint of X, Y being equal to r, k writing it as the product of the marginal of S and the conditional of X given S. The marginal of S can be found by noticing that $S \sim \text{Pois}(\lambda)$. The conditional of X given S is a Binomial distribution, indeed we want to estimate the probability of having exactly r successes out of $k + r$ trials, where the probability of success is p.

Notice how $\mathbb{P}[\![X = r \mid S = n]\!] = \binom{n}{r} p^r (1-p)^{n-r} \overset{\substack{n=k+r \\ n-r=k}}{\Longrightarrow} \binom{k+r}{r} p^r (1-p)^{k+r-r}$. And we have written $e^{-\lambda}$ as $e^{-\lambda(p+1-p)} = e^{-\lambda p}\ e^{-\lambda(1-p)}$ which has been later split into the two partes in the last equality. We can notice that the two factors are exactly the p.m.f.'s of two independent Poisson random variables with parameters λp and $\lambda(1-p)$ respectively.

In the end, the important takeaway is that, with this type of random variables, if we have the marginals, we can also derive the conditionals and vice-versa. This property may look trivial but it is actually quite unique in the whole world of probability distributions.

3.10. Exponential Distribution

Another very important distribution that is often used to model 'natural' phenomena is the **Exponential distribution**. Specifically it is often used to model **time**: waiting times, inter-arrival times, hardware lifetime, failure times and so on.

We are not going to spend time on giving the definition of this distribution, since it is pretty much all contained in the few lines above. Similarly to Poisson random variables, exponential random variables are also characterized by a **rate** parameter λ which models the expected number of events occurring per unit of time.

Probability Density Function

Since this distribution models time, it is only possible to define it in a continuous fashion. Let $X \sim \text{Exp}(\lambda)$ be an exponential random variable with parameter $\lambda > 0$. The **probability density function** of X is defined as:

$$f_X(x) = \lambda e^{-\lambda x} \quad \forall x > 0 \tag{3.30}$$

By means of this p.d.f. we can also write the **cumulative distribution function** as follows:

$$F_X(x) = \mathbb{P}[\![X \leq x]\!] = \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x} \tag{3.31}$$

Sometimes it may be really useful to deal with the **survival function**, which is defined as in Equation 2.6, therefore in this specific case we have:

$$S_X(x) = \mathbb{P}[\![X > x]\!] = 1 - F_X(x) = e^{-\lambda x} \tag{3.32}$$

Expected Value and Variance

Since we know the expected value of a Poisson random variable with parameter λ is exactly λ, we can leverage this, and obtain that, since λ here models the amout of 'rare' events occurring per unit of time, the **expected** waiting time for the occurrence of one such event is:

$$\mathbb{E}[\![X]\!] = \frac{1}{\lambda} \tag{3.33}$$

To see this in a more formal way, we can compute the expected value as follows:

$$\begin{aligned}
\mathbb{E}[\![X]\!] &= \int_0^\infty x \ \lambda e^{-\lambda x} \ dx \quad \text{integrating by parts: } \begin{cases} u = x, \ du = dx \\ dv = \lambda e^{-\lambda x}, \ v = -e^{-\lambda x} \end{cases} \\
&= \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} \ dx = 0 - \left[\left(\frac{e^{-\lambda x}}{\lambda}\right)\right]_0^\infty = \frac{1}{\lambda}
\end{aligned}$$

As far as the variance is concerned we need to first compute the value of $\mathbb{E}[\![X^2]\!]$:

$$\begin{aligned}
\mathbb{E}[\![X^2]\!] &= \int_0^\infty x^2 \ \lambda e^{-\lambda x} \ dx \quad \text{integrating by parts: } \begin{cases} u = x^2, \ du = 2x \\ dv = \lambda e^{-\lambda x}, \ v = -e^{-\lambda x} \end{cases} \\
&= \left[-x^2 e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x} \ dx = 0 + \frac{2}{\lambda} \int_0^\infty \lambda x e^{-\lambda x} \ dx = \frac{2}{\lambda^2}
\end{aligned}$$

Now we can compute the **variance** as follows:

$$\text{Var}\{x\} = \mathbb{E}[\![X^2]\!] - \mathbb{E}[\![X]\!]^2 = \left(\frac{2}{\lambda^2}\right) - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2} \tag{3.34}$$

R Implementation

As usual, in `R` we have the following functions to work with exponential random variables:

- `dexp(x, lambda)` $= f_X(x)$, is the probability density function (p.d.f.).
- `pexp(x, lambda)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function (c.d.f.).
- `qexp(q, lambda)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, is the quantile function.
- `rexp(r, lambda)` simulates r realizations of X.

Poisson - Exponential Relationship

As we have already hinted, there is a very strong relationship between Poisson and Exponential random variables. This section is dedicated to exploring it in detail.

Consider a sequence or *rare* events, where the number N_t of occurrences during a period of time of length t is modeled as a Poisson random variable with parameter λ proportional to t. In other words we can write $N_1 \sim \text{Pois}(\lambda), N_t \sim \text{Pois}(\lambda t)$.

Consider now the event $A =$ "the time T until the next event (arrival) is greater than t". This is basically equivalent to saying that "during a period of time of length t no events occur": we can write the event as $A = \{N_t = 0\}$. If we try to *compute the probability of* A we get the following:

$$\mathbb{P}[\![A]\!] = \mathbb{P}[\![T > t]\!] = \mathbb{P}[\![N_t = 0]\!] = e^{-\lambda t}$$

This is because, if we take a look at the p.m.f. of a Poisson random variable in Equation 3.26, we can see that our parameter when considering the time period of length t is exactly λt, therefore if we set $x = 0$ we get exactly the following equality:

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda t}$$

But this is exactly equal to the **survival function** of an Exponential random variable with parameter λt as shown in Equation 3.32. This property is also known as the **inevitability of exponential distribution**.

Properties of Exponential Random Variables

We have defined quite a bit of properties about Poisson random variables; not surprisingly, since they are so tightly related to Exponential ones, many of these properties can be translated to Exponential random variables as well.

Memoryless Property

One of the most important properties of Exponential random variables is their **memoryless property**. We are going to present it in the following theorem.

Theorem 3.6 (Memoryless Property for Exponential)

Suppose that an exponential random variable T represents a waiting time. Regardless of the event $T > t$, when the total waiting time exceeds t, the remaining waiting time still has exponential distribution with the same parameter λ.

$$\mathbb{P}[\![T > t + x \mid T > t]\!] = \mathbb{P}[\![T > x]\!] \quad \text{for } t, x > 0$$

where t represents the portion of waiting time that has already elapsed, and x represents the additional remaining time.

This can be proved by recognizing the survival function of the exponential distribution:

$$\mathbb{P}[\![T > t + x \mid T > t]\!] = \frac{\mathbb{P}[\![T > t + x \cap T > t]\!]}{\mathbb{P}[\![T > t]\!]} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}}$$
$$= e^{-\lambda x} = \mathbb{P}[\![T > x]\!]$$

> **Remark**
>
> Just like the **geometric distribution** is the only discrete memoryless distribution, the **exponential distribution** is the only continuous memoryless distribution.

This is by no surprise, since we have already seen that the Binomial distribution is related to the Geometric distribution in a very similar way as the Poisson distribution is related to the Exponential distribution.

Minimization Property

When we talked about the Poisson distribution, we have mentioned the case in which we may have different independent Poisson random variables and we were interested in their sum, in which case the sum was still a Poisson random variable, and on the other hand we have also seen that if we had the sum of independent Poisson random variables, conditionally on the value of the sum, the individual random variables were multinomially distributed.

Since exponential random variables do not consider the number of events occurring but rather times until the occurrence of such events, it makes sense to consider the **minimum** of a set of independent exponential random variables.

Theorem 3.7 (Minimization Property)

Let $X_j \sim \text{Exp}(\lambda_j)$ for $j = 1, 2, ..., n$ be a collection of **independent** exponential random variables, then we have that:

$$L_n = \min\{X_1, X_2, ..., X_n\} \sim \text{Exp}(\lambda)$$

where $\lambda = \sum_{j=1}^{n} \lambda_j$ is the parameter of the resulting exponential random variable.

♡

Indeed, we can take a look at the cumulative distribution function of L_n:

$$\begin{aligned} F_{L_n}(x) &= \mathbb{P}[\![L_n \leq x]\!] = 1 - \mathbb{P}[\![L_n > x]\!] \\ &= 1 - \mathbb{P}[\![X_1 > x,\ X_2 > x,\ ...,\ X_n > x]\!] \\ &= 1 - \prod_{j=1}^{n} \mathbb{P}[\![X_j > x]\!] \quad \text{by independence} \\ &= 1 - \prod_{j=1}^{n} e^{-\lambda_j x} = 1 - e^{\sum_{j=1}^{n} -\lambda_j x} = 1 - e^{-\lambda x} \end{aligned}$$

where we went from the first to the second line by noticing that the minimum of n r.v.'s is greater than x if and only if all the individual r.v.'s are greater than x. In the end we have

obtained exactly the survival function of an exponential random variable with parameter λ, as we were supposed to.

Remark

It is important to notice that the minimum L_n of independent exponential random variables is **not independent** of $\{X_1, X_2, ..., X_n\}$. Indeed $L_n = X_k$ for some $k \in \{1, ..., n\}$ and:

$$\mathbb{P}[\![L_n = X_k]\!] = \frac{\lambda_k}{\lambda} = \frac{\lambda_k}{\sum_{j=1}^{n} \lambda_j}$$

Indeed, by the law of total probability for continuous random variables we have that:

$$\begin{aligned}
\mathbb{P}[\![L_n = X_k]\!] &= \int_0^\infty \mathbb{P}\left[\!\!\left[\bigcap_{j \neq k} (X_k < X_j) \mid X_k = x \right]\!\!\right] f_{X_k}(x)\, dx \\
&= \int_0^\infty \mathbb{P}\left[\!\!\left[\bigcap_{j \neq k} X_j > x \right]\!\!\right] f_{X_k}(x)\, dx = \int_0^\infty \left(\prod_{j \neq k} e^{-\lambda_j x} \right) \left(\lambda_k e^{-\lambda_k x} \right) dx \\
&= \lambda_k \int_0^\infty e^{-(\lambda_1, ..., \lambda_n)x}\, dx = \frac{\lambda_k}{\sum_{j=1}^{n} \lambda_j}
\end{aligned}$$

where we switched from the first to the second equality by noticing that the all the X_j's are independent from the conditioning event $X_k = x$. It is clear that $\sum_{k=1}^{n} \mathbb{P}[\![L_n = X_k]\!] = 1$.

The above result can actually be generalized to the case in which we have a countably infinite number of independent exponential random variables. If the same of the parameters converges to a finite value, say λ, when we have that:

$$\mathbb{P}[\![L_n = x]\!] = \frac{\lambda_k}{\lambda} \quad \text{where } \lambda = \sum_{j=1}^{\infty} \lambda_j < \infty$$

Warning

The same reasoning **cannot** be applied to the **maximum** of two or more independent exponential random variables. Indeed:

$$\begin{aligned}
\mathbb{P}[\![\max\{X_1, X_2\} \leq x]\!] &= \mathbb{P}[\![X_1 \leq x,\, X_2 \leq x]\!] = \mathbb{P}[\![X_1 \leq x]\!]\, \mathbb{P}[\![X_2 \leq x]\!] \\
&= \left(1 - e^{-\lambda_1 x}\right) \left(1 - e^{-\lambda_2 x}\right)
\end{aligned}$$

which cannot be simplified to the c.d.f. of an exponential random variable. Intuitively, this happens because we cannot reduce the c.d.f to a survival function as we did in the original case with the minimum.

3.11. Gamma Distribution

When dealing with discrete random variables, we have seen how, given a Bernoulli trial and a series of independent repetitions of it, we can define more and more complex distributions, arriving at the negative binomial distribution, which models the number of trials until a fixed number of successes is observed, which in practice is the sum of independent geometric random variables.

Now we are going to se how, starting from a bunch of independent exponential random variables, we can define a more general distribution, called the **Gamma distribution**, which models the behavior of their sum.

Definition 3.12 (Gamma Distribution)

Let $X_1, X_2, ..., X_n$ be independent exponential random variables with parameter λ. The **Gamma distribution** with parameters $\alpha = n$, $\lambda = \lambda$ serves to model the distribution of the sum $Y = X_1 + X_2 + ... + X_n$. We write $Y \sim \text{Gamma}(\alpha, \lambda)$.

♣

The parameter α is often called the **shape parameter**, while λ is called the **rate parameter**. We can notice that when $\alpha = 1$ we have that the Gamma distribution reduces to the Exponential distribution.

Probability Density Function

The **probability density function** of a Gamma random variable with parameters α, λ is defined as follows:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \, x^{\alpha-1} e^{-\lambda x} \quad \forall x > 0 \tag{3.35}$$

Intuitively the factor $x^{\alpha-1}$ is used to gradually decrease the speed of the exponential decay, this is mainly the reason why, for the shape parameter $\alpha = 1$ the distribution reduces to the exponential one. Figure 3.5 shows how the α parameter affects the shape of the p.d.f. of a Gamma random variable.

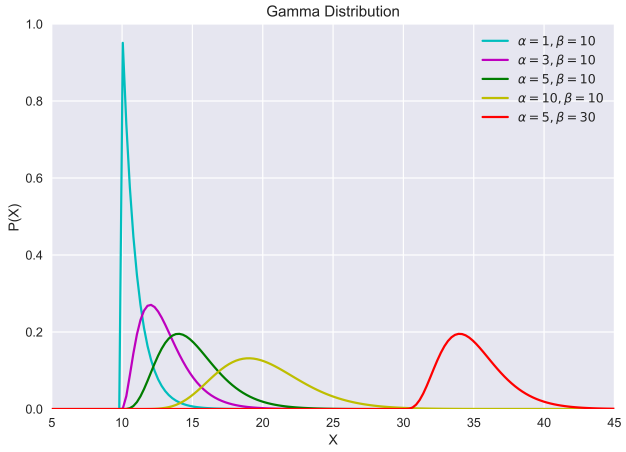


Figure 3.5: Probability density function of a Gamma random variable for different values of the shape parameter

Gamma Function

In the above definition we have used the **Gamma function** $\Gamma(\alpha)$ which can be seen as a generalization of the factorial function to real numbers. Suppose we are to write an algorithm that computes the factorial of a number n: in this case we must proceed **recursively**, since the factorial is defined by induction.

For the Gamma function we are going to proceed similarly, the only difference is that we are going to define it for any $\alpha \in \mathbb{R}$. One may therefore think that we can define it as follows:

$$\Gamma(\alpha) = \alpha \cdot \Gamma(\alpha - 1)$$

The problem is the **starting point** (the base case of induction). This was easy in case of the factorial in that we stopped when reaching $0! = 1$. Here we need to proceed differently, before defining the function formally it is important to start from the mathematica (probabilistic, actually) motivation behind it. Consider a random variable with the following probability density function:

$$f_X(x) = k\ x^{\alpha-1} e^{-\lambda x} \quad \forall \alpha, x > 0$$

where k is a normalizing constant that makes the area under the curve equal to 1. To find the value of k we need to solve the following integral.

$$1 = k \int_0^\infty x^{\alpha-1}\ e^{-\lambda x}\ dx$$

To solve this we proceed by parts, identifying the following: $\begin{cases} u = x^{\alpha-1},\ du = (\alpha-1)x^{\alpha-2} \\ dv = e^{-\lambda x} dx,\ v = -\frac{e^{-\lambda x}}{\lambda} \end{cases}$. If $\alpha = n$, integration by parts takes us to the following:

$$\frac{1}{k} = \int_0^\infty x^{n-1}\ e^{-\lambda x}\ dx = \left[x^{n-1} \frac{e^{-\lambda x}}{-\lambda} \right]_0^\infty - \int_0^\infty (n-1)x^{n-2}\ \frac{e^{-\lambda x}}{(-\lambda)}$$

$$= 0 + \frac{n-1}{\lambda} \int_0^\infty x^{n-2}\ e^{-\lambda x}\ dx$$

$$= (n-1)\frac{n-2}{\lambda^2} \int_0^\infty x^{n-3}\ e^{-\lambda x}\ dx = \ldots = \frac{(n-1)!}{\lambda^n}$$

Therefore we can conclude that $k = \frac{\lambda^n}{(n-1)!}$. This is not the normalization factor we find in Equation 3.35, that's because this is only valid in case α is an integer.

When lambda is not an integer, the exponent of the x's that get derived inside the integration by parts will not reach 0 and we will not be able to stop the process, instead they may reach negative values. In general we have that the equation

$$\frac{1}{k} = \int_0^\infty x^{\alpha-1}\ e^{-\lambda x}\ dx$$

has not a closed (analytical) form solution for any α. To solve this, a named function has been defined, called the **Gamma function**, which has actually been computed by setting $\boldsymbol{\lambda = 1}$ and is defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\ dt$.

Expected Value and Variance

Following we are going to provide the formulas to compute the expected value and the variance of a Gamma random variable with parameters α, λ.

$$\mathbb{E}[\![X]\!] = \frac{\alpha}{\lambda} \qquad \text{Var}\{X\} = \frac{\alpha}{\lambda^2} \tag{3.36}$$

Additivity of Exponential Random Variables

Now that we have defined the Gamma distribution it is important to mention that, if we have n independent random variables $X_j \sim \text{Exp}(\lambda)$ for $j = 1, 2, ..., n$, then their sum $S_n = X_1 + X_2 + ... + X_n$ is a Gamma random variable with parameters $\alpha = n$ and $\lambda = \lambda$.

R Implementation

As usual, in `R` we have the following functions to work with Gamma random variables:

- `dgamma(x, shape, rate)` $= f_X(x)$, is the probability density function
- `pgamma(x, shape, rate)` $= \mathbb{P}[\![X \leq x]\!]$, is the cumulative distribution function
- `qgamma(q, shape, rate)` $= x = F^{-1}(q)$, i.e., $\mathbb{P}[\![X \leq x]\!] = q$, the quantile function
- `rgamma(r, shape, rate)` simulates r realizations of X.

4. Introduction to Stochastic Processes

After introducing the basic elements we are going to work with, we can present the really interesting objects that are going to make use of the previously introduced building blocks: **stochastic process**. This chapter provides an introduction to stochastic processes and some of their properties.

4.1. Central Limit Theorem and Law of Large Numbers

Before actually defining stochastic processes, it is fundamental to introduce two key results in probability theory that are essential for understanding the behavior of stochastic processes: the **Central Limit Theorem** and the **Law of Large Numbers**. These theorems provide insights into how random variables behave when aggregated over time or across many trials.

4.1.1. Sequences of Sums of Random Variables

Let us consider a sequence of random variables $X_1, X_2, X_3, \dots$ and the **sum** of their first n elements. As we have seen in the previous chapter, this is again a random variable:

$$S_n = X_1 + X_2 + \dots + X_n$$

We should already be familiar with the following facts:

- if all random variables X_i have a **common mean** $\mu = \mathbb{E}[\![X_i]\!]$, then the expected value of the sum is given by: $\mathbb{E}[\![S_n]\!] = n\mu$
- if all random variables X_i are **independent** and have a **common variance** $\sigma^2 = \text{Var}\{X_i\}$, then the variance of the sum is given by: $\text{Var}\{S_n\} = n\sigma^2$

These two facts combined give us an idea that the distribution of these incremental sums S_n is normal (because it comes from the sum of normal random variables) with mean $n\mu$ and variance $n\sigma^2$: $S_n \sim \mathcal{N}(0, n)$.

Additionally, for each n, we know that these variables are **not independent**, in fact, $S_{n+1} = S_n + X_{n+1}$, so we have that:

$$\mathbb{E}[\![S_{n+1} \mid S_n]\!] = S_n, \text{Var}\{X_{n+1} \mid S_n\} = 1 \implies S_{n+1} \mid S_n \sim \mathcal{N}(S_n, 1)$$

The dependence between the variables creates a very different structure of the observed sequences (or **paths**) compared to what we would observe if the variables were independent. We can visualize this by simulating some realizations from the sequence of sums S_n, $n = 1, 2, \dots$ and comparing them with realizations of a sequence $W_n \sim \mathcal{N}(0, n)$.

To start off we try to produce a single realization of the first $N = 1000$ variables in the initial i.i.d. $X_n \sim \mathcal{N}(0, 1)$ sequence:

```r
set.seed(42)  # set random number generator seed
N = 1000  # number of random variables to simulate
X = rnorm(N)   # simulate N i.i.d. standard normal r.v.'s
```

We can observe the results of this simulation by the command `head(X)` which provides the first few values of the simulated vector `X`. Alternatively it is possible to visualize the

scatter plot of the simulated values. We prefer to visualize the results using a scatter plot to remind us of *individual realizations*, however, to represent the idea of a single random function $X(n) = X_n$ changing value through time, we can also use a line plot `plot(x, type='l', xlab='n', ylab='expression(X[n])')`. This comparison is shown in Figure 4.1.

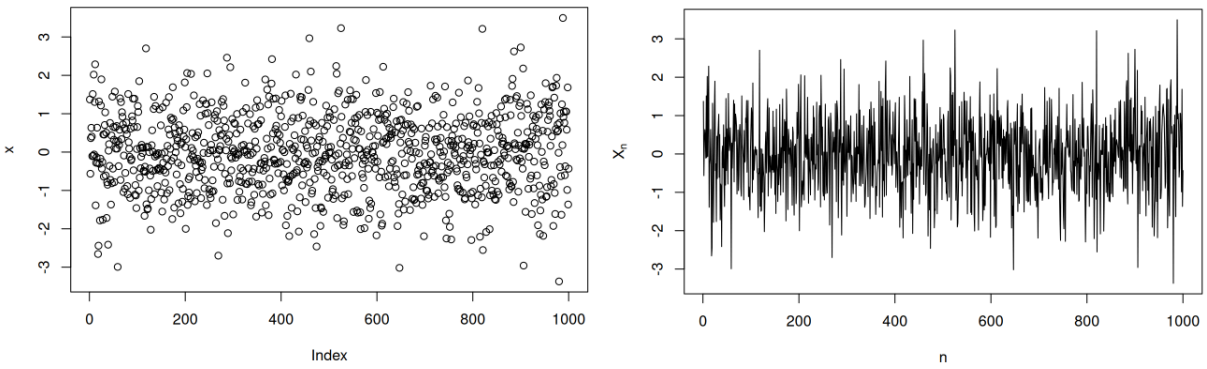


Figure 4.1: Left: Scatter plot of a single realization of $X_n \sim \mathcal{N}(0,1)$. Right: Line plot of the same realization of $X_n \sim \mathcal{N}(0,1)$. Both plots represent the same experiment run 1000 times.

It is important to remember that random variables are functions of the results of an experiment. The sequence we just produced can be seen as the initial 1000 outcomes of an infinite sequence corresponding to a single observed experiment result $\omega \in \Omega$. We may try to repeat the experiment multiple times, say 20, obtaining each time a different observed sequence, or path, of this process. This is done through the function `replicate` in `R` in the following code whose results are illustrated in Figure 4.2:

```R
N = 1000
m = 20  # number of realizations
x_paths = replicate(m, rnorm(N))
plot(x_paths[,1], type='l',xlim=c(0,N), ylim=range(x_paths),
      xlab='n', ylab=expression(X[n]))  # plot the first realization

for (i in 2:m) # plot the remaining on the existent canvas
  lines(x_paths[,i], col=i)
```

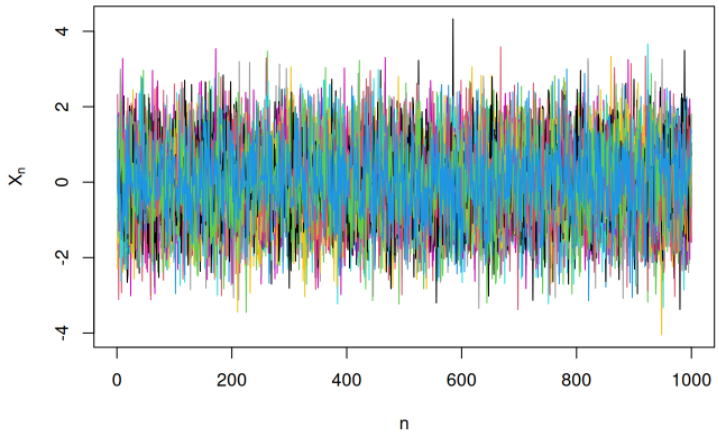


Figure 4.2: 20 realizations of the sequence $X_n \sim \mathcal{N}(0,1)$.

From Figure 4.2 we can notice how all the paths exhibit a similar behavior. This is due to the independence of the individual X_n variables. These sequences of i.i.d. standard normal random variables are a mathematical representation of what is called **white noise** in *time series analysis.*

Now that we have defined the realizations of all the random variables we are going to work with, we can proceed to create the sums S_n. To start off, let's create a single realization of the sums S_n for $n = 1, ..., N$:

```r
set.seed(42)
N = 1000
S = rnorm(1)
for (n in 2:N)
  S = c(S1, rnorm(1) + S[n-1])
```

```r
set.seed(42)
N = 1000
S = cumsum(rnorm(N))
```

These codes are actually identical and produce the same resulting vector `S` which contains a single realization of the cumulativ sum $S_n = \sum_{i=1}^{n} X_i$ for $n = 1, ..., N$. The only difference is that the code on the right uses much more efficient functions and should be preferred when working at large scale.

Following we illustrate the results of simulating a single realization of the sums S_n via the command `plot(S, type='l', xlim=c(0,N), ylim=max(abs(S))*c(-1,1))` in Figure 4.3.

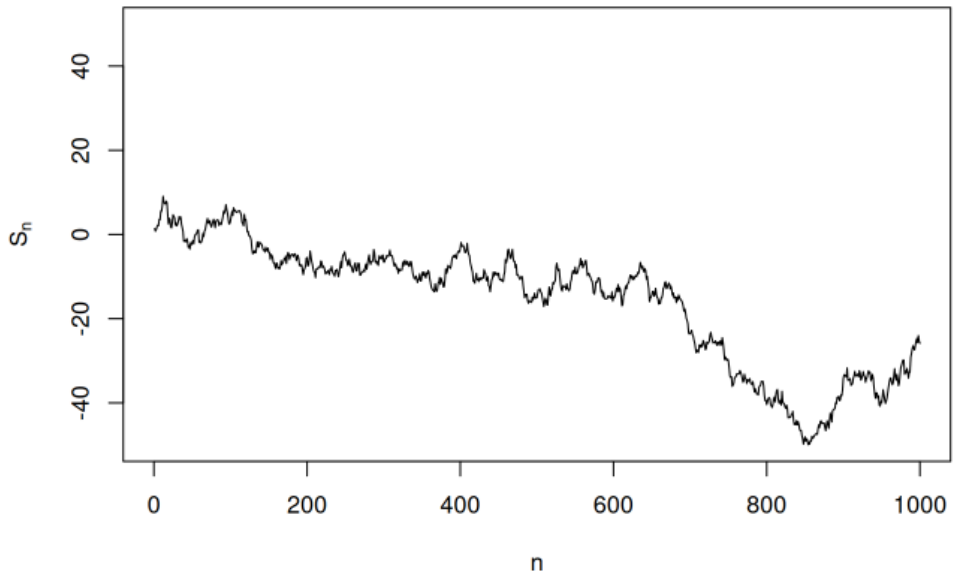


Figure 4.3: Single realization of the sum $S_n = \sum_{i=1}^{n} X_i$ for $n = 1, ..., N$.

In a very similar way as before, we can produce multiple realizations of the sums S_n by using the `replicate` function in `R`. This is shown in the following code:

```r
set.seed(42)
N = 1000
m = 20  # number of realizations
S_paths = replicate(m, cumsum(rnorm(N)))

```

```
# plot the first realization to create the canvas
plot(S_paths[,1], type='l',xlim=c(0,N), ylim=range(S_paths),
      xlab='n', ylab=expression(S[n]))

# plot the remaining realizations on the existent canvas
for (i in 2:m)
  lines(S_paths[,i], col=i)
```

The results of this code are shown in Figure 4.4.

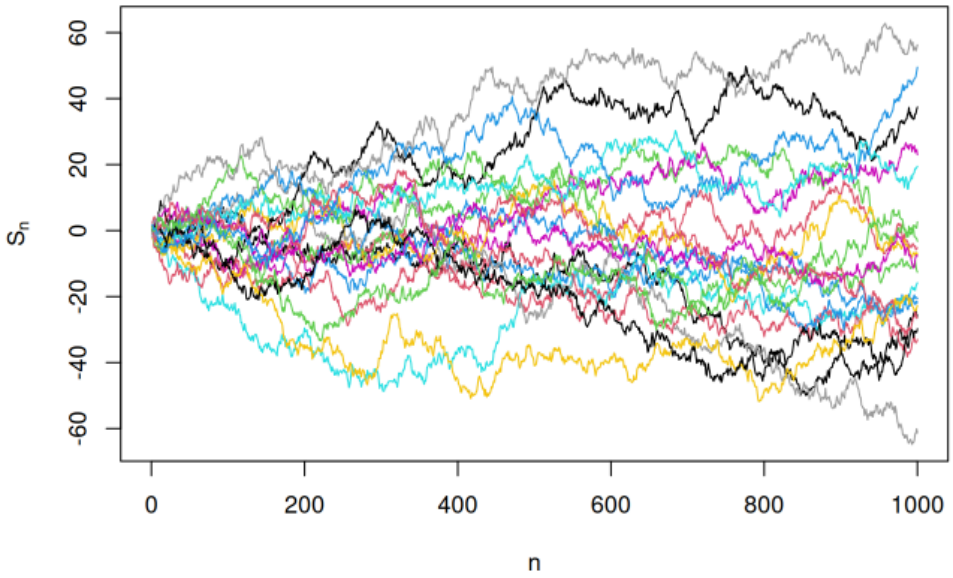


Figure 4.4: 20 realizations of the sum $S_n = \sum_{i=1}^{n} X_i$ for $n = 1, ..., N$.

We can notice that all these paths exhibit a similar "wandering" behavior, indeed each step only depends on the previous one plus a new independent standard normal random value, which will make the trajectory move up or down with equal probability. All values are centered around the same mean (0) but as the value of n increases we see that the paths tend to spread out more and more.

This means that we are actually able to describe the distribution of S_n for each fixed n characterizing it as a normal random variable with mean 0 and variance n: $S_n \sim \mathcal{N}(0, n)$. Even though we noticed that X_{n+1} is independent of S_n, the newly obtained sum S_{n+1} is **not independent** of S_n, in fact we have that $S_{n+1} \mid S_n \sim \mathcal{N}(S_n, 1)$. This may not look obvious at first, but it is a crucial point. To realize this, let's have a look at what would happen if we tried to simulate N realizations of n independent random variables each with distribution $\mathcal{N}(0, n)$:

```
set.seed(42)
N=1000
W = rnorm(N, 0, sqrt(c(1:N)))
plot(W, type='l', xlab("n", ylab=expression(W[n])))
```

This code produces a single realization of N independent random variables $W_n \sim \mathcal{N}(0, n)$ for $n = 1, ..., N$. The results of this simulation are shown in Figure 4.5.

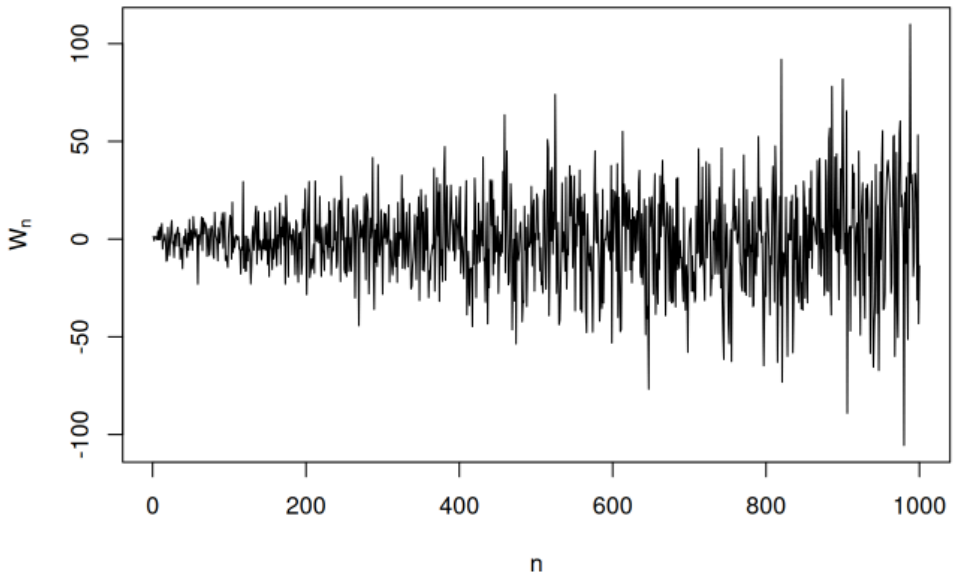


Figure 4.5: Realization of N independent random variables $W_n \sim \mathcal{N}(0, n)$ for $n = 1, ..., N$.

We can see how a single realization of this process produces a much more erratic behavior than any of the individual paths of dependent sums in Figure 4.4. Because the marginal distributions coincide, we can still see how the observed values w_n tend to get further away from the mean as n grows, but they are able to span the whole range of values in a single realization, while we needed $m = 20$ realizations of the dependent processes to produce a similar coverage of the value range.

This shows the importance of the dependence between the S_n variable. Singularly, for each n, we have that S_n and W_n have the same marginal distribution, but within a single path the W_n are free to vary independently, while the S_n are constrained by their previous values.

4.1.2. Law of Large Numbers

We can notice that as we make n grow to infinity, both S_n and W_n will tend to diverge to infinity as well, because their variance grows without bound. Since the variance tends to grow linearly with n we can try to normalize these variables by dividing each sum by n. This produces a new sequence of random variables:

$$\overline{X}_n = \frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4.1}$$

Basically for each n we are computing the **sample mean** of of the sample $X_1, ..., X_n$. Let's look at the code to produce directly the same number of realizations of this new sequence of random variables as the previous example in Figure 4.4:

```r
# reproduce same experimental setup
set.seed(42)
N = 1000
m = 20

sm_paths <- replicate(m, cumsum(rnorm(N)) / (1:N))
```

```
7
8  # plot the first realization to create the canvas
9  plot(sm_paths[,1], type='l', xlim=c(0,N), ylim=range(sm_paths),
10       xlab='n', ylab=expression(bar(X)[n]))
11
12 for (i in 2:m)
13   lines(sm_paths[,i], col=i)
```

Following we illustrate the results of this code in Figure 4.6.

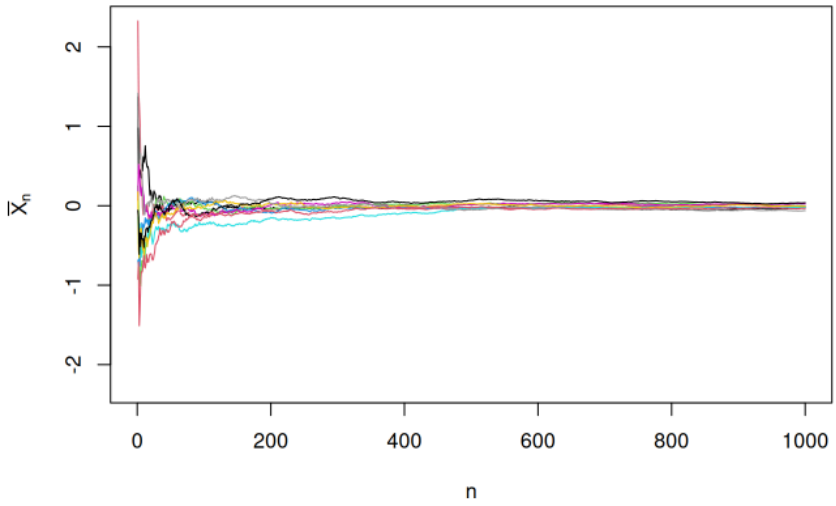


Figure 4.6: 20 realizations of the sample mean sequence $\overline{X}_n = \frac{1}{n} S_n$

We can notice how now we have managed to control the variability of the sequence, but perhaps we did it too much. In fact all realizations seem to converge to have variance 0 quite fast. This is by no surprise since we have the following result

$$\text{Var}\{\overline{X}_n\} = \frac{1}{n} \underset{n\to\infty}{\longrightarrow} 0$$

Basically as n grows the mean remains fixed to 0, but the variance shrinks to 0 as well, making all realizations converge to the same constant value (the mean).

Theorem 4.1 (Law of Large Numbers)

Let $X_1, X_2, ..., X_n$ be a sequence of independent, identically distributed random variables with finite mean $\mu = \mathbb{E}[\![X_i]\!]$ and finite variance $\sigma^2 = \text{Var}\{X_i\}$. Then, the sample mean sequence defined in Equation 4.1 converges in probability to the true mean μ as n approaches infinity:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \underset{n\to\infty}{\longrightarrow} \mu$$

4.1.3. Central Limit Theorem

The Law of Large Number is a very important result that guarantees that the sample mean converges to the true mean as the sample size increases. However, it does not preserve any

type of randomness in the limit. We would now like to control the variance of the sequence without making it vanish completely; in order to do so, we can try to divide the sums by a smaller factor than n, for instance $\sqrt{n}$ (which is exactly the standard deviation of S_n).

> **Remark**
>
> Dividing each sum by $\sqrt{n}$ is equivalent to **standardizing** the sums:
>
> $$Z_n = \frac{S_n - \mathbb{E}[\![S_n]\!]}{\sqrt{\mathrm{Var}\{S_n\}}} = \frac{1}{\sqrt{n}} S_n$$

Following we illustrate the code to produce multiple realizations of this standardized sequence:

```r
set.seed(42)
N = 1000
m = 20

z_paths -> replicate(m, cumsum(rnorm(N)) / sqrt(1:N))

# plots
plot(z_paths[,1], type='l', xlim=c(0,N), ylim=range(z_paths),
     xlab='n', ylab=expression(Z[n])
for (i in 2:m
  lines(z_paths[,i], col=i)
```

Figure 4.7 shows the results of this simulation, where we can notice how this time we have generated a sequence of identically distributed standard normal random variables.

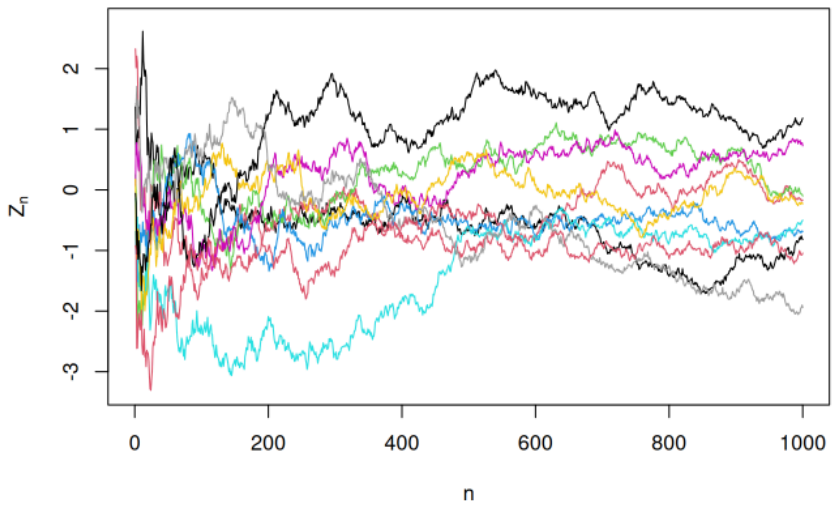


Figure 4.7: 20 realizations of the standardized sequence $Z_n = \frac{1}{\sqrt{n}} S_n$

Even though all variables Z_n are identically distributed as standard normal random variables, they are **not independent** as we can notice from the regularity of the individual paths. We can actually visualize the marginal distribution of some of the Z_n variables using a histogram,

for different values of n: $n = 1, n = \frac{N}{2}, n = N$. To effectively visualize the histograms we are going to simulate a larger number of realizations, say $m = 1000$. This is shown in Figure 4.8.

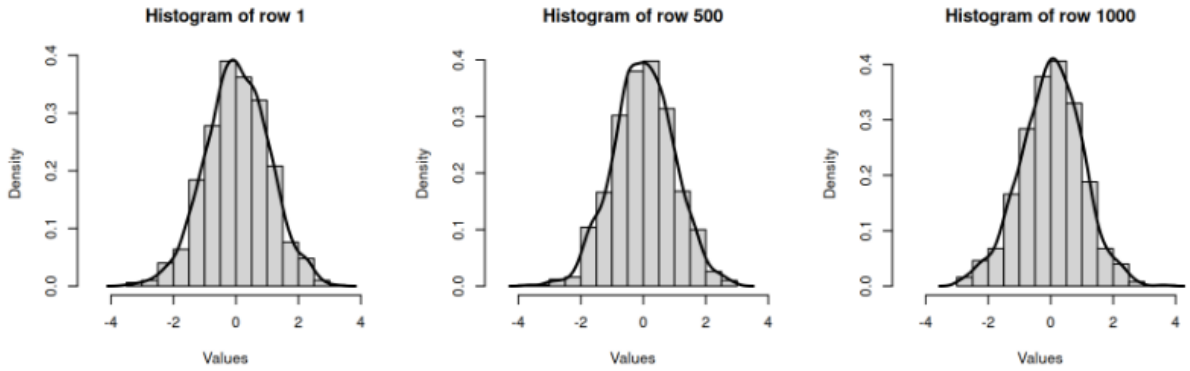


Figure 4.8: Histograms of the marginal distributions of Z_n for different values of n. From left to right: $n = 1, n = \frac{N}{2}, n = N$.

Theorem 4.2 (Central Limit Theorem)

Let $X_1, X_2, ..., X_n$ be a sequence of independent, identically distributed random variables with finite mean $\mu = \mathbb{E}[\![X_i]\!]$ and finite variance $\sigma^2 = \mathrm{Var}\{X_i\}$ amd let

$$S_n = \sum_{i=1}^{n} X_i = X_1 + X_2 + ... + X_n$$

As $n \to \infty$, the **standardized sum** defined as

$$Z_n = \frac{S_n - \mathbb{E}[\![S_n]\!]}{\sqrt{\mathrm{Var}\{S_n\}}} = \frac{S_n - n\mu}{\sigma}\sqrt{n}$$

converges in distribution to a **standard normal random variable**, that is:

$$F_{Z_n}(z) = \mathbb{P}\left[\!\!\left[\frac{S_n - n\mu}{\sigma}\sqrt{n} \leq z \right]\!\!\right] \longrightarrow \Phi(z) \qquad \forall z \tag{4.2}$$

This is quite straightforward to understand in this case, since we have started from standard normal random variables X_i and summed them up to obtain normal random variables S_n. However, the **central limit theorem** holds in much more general settings, in fact it holds even when the original distribution of the sample X_n is not normal, but anything else. In such case the distribution of Z_n will no longer be exactly normal but will **converge** to it. This holds also in case of the Law of Large Numbers: regardless of the original distribution of the X_n variables, the sample mean will always converge to the true mean if the X_n sample has a common mean μ.

4.2. Introduction to Stochastic Processes

After providing two of the most important results in probability theory, it's finally time to introduce the long awaited **stochastic processes**.

Definition 4.1 (Stochastic Process)

A **stochastic process** is a random variable that also depends on **time**. It is therefore a function of *two arguments* $X(t, \omega)$, where:

- $t \in \mathcal{T}$ is time, with $\mathcal{T}$ being a set of possible times, usually $[0, \infty)$, $(-\infty, \infty)$, $\{0, 1, 2, 3...\}$ or $\{..., -2, -1, 0, 1, 2, ...\}$
- $\omega \in \Omega$ is the outcome of a random experiment, with Ω being the whole sample space

In this context, the values of $X(t, \omega)$ are called the **states** of the process.

♣

The one we have just introduced in Definition 4.1° is actually one of the two possible ways we have at our disposal for defining a stochastic process. At any fixed time t we have a random variable $X_{t(\omega)}$, a function of a random outcome; on the other hand, if we fix the outcome w, we obtain a function of time $X_{\omega(t)}$. This function is called a **realization** or **sample path** or a **trajectory** of the process $X = \{X(t) : t \in \mathcal{T}\}$.

Definition 4.2 (Stochastic Process - Alternative Definition)

For a given sample space $\mathcal{S}$ of some experiment, a **random process** is any rule that associates a time-dependent function with each outcome in $\mathcal{S}$. Any such function that may result is a **sample function** of the random process. The collection of all possible sample function is called the **ensemble** of the random process.

♣

A stochastic process may be classified according to the nature, either of the outcomes of the experiment or of the time parameter in the following way:

- $X(t, \omega)$ is **discrete-state** if the variable $X_{t(\omega)}$ is **discrete** for each time t, and it is **continuous-state** if $X_{t(\omega)}$ is **continuous** for each time t.
- $X(t, \omega)$ is a **discrete-time** process if the set of times $\mathcal{T}$ is discrete, that is, it consists of separate isolated points. It is a **continuous-time** process if the set of times $\mathcal{T}$ is a connected, possibly unbounded, interval of the real line.

Example: Example of Stochastic Process

A communication system uses phase-shift keying to transmit information. A quaternary phase-shift keying system can transmit four distinct symbols (often used to encode two bits at a time). The four symbols are distinguished by varying the phase at which they are transmitted; specifically for $k \in \{1, 2, 3, 4\}$, the k-th symbol is transmitted for T seconds with the following wave form:

$$x_{k(t)} = \cos\left(2\pi f_0 t + \frac{\pi}{4} + k\frac{\pi}{2}\right), \quad 0 \leq t \leq T$$

for some determined frequency f_0. Considering the transmission of a single randomly selected symbol, and letting $X(t)$ denoting the corresponding transmitted wave we obtain the graph in Figure 4.9.

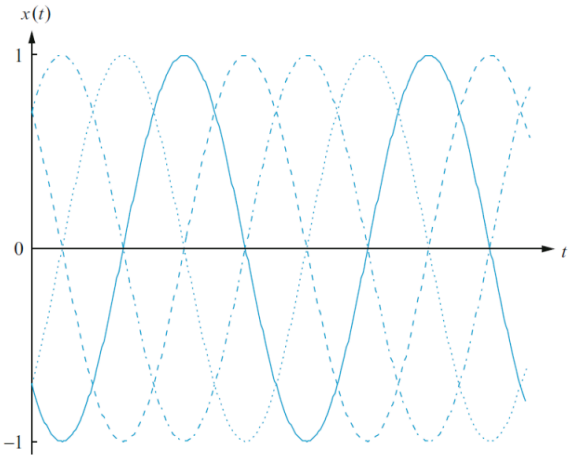


Figure 4.9: Ensamble of a continuous-time, continuous-state stochastic process

Notice how this scenario is also quite strange: whenever we fix a time, the support of the random variable $X(t)$ may change, indeed for some times t the possible values of $X(t)$ are only two, while for other times they are four. This is because at some times the different waveforms may overlap.

At any fixed time point t_0, the ensemble of a random process $X(t)$ forms a probability distribution: $X(t_0)$ is a random variable with support determined by the ensemble. Just like when we discussed vectors of random variables, we did not limit ourselves to describing only *marginal distribution*, they would not be enough: to fully characterize the behavior of a stochastic process we need to consider also the **joint distributions** of $X(t_1), ..., X(t_r)$ for all finite sets of time points $t_1 < ... < t_r$. The collection of all such joint distributions constitutes the **finite-dimensional distributions** of the process.

For the purpose of this course, we are going to deal mainly with continuous-time stochastic processes.

4.2.1. Mean and Variance Functions

In this section we are going to introduce some of the most important functions that can be used to describe the behavior of a stochastic process. Suppose we have a random process $X = \{X(t) : t \in \mathcal{T}\}$. For each fixed $t \in \mathcal{T}$, $X(t)$ is a random variable, therefore we can define its expected value and variance.

If we keep into account all the random variables we can obtain by varying the fixed time t, we can then define some functions that describe how the mean and the variance of the random process change over time.

Definition 4.3 (Mean and Variance Functions)

The **mean function** of a random process $X(t)$ is given by

$$\mu_X(t) = \mathbb{E}[\![X(t)]\!] \tag{4.3}$$

where $\mathbb{E}[\![X(t)]\!]$ is the expected value of the random variable $X(t)$ for the fixed time point t. Similarly we can define the **variance** and **standard deviation functions** of the process as:

$$\begin{aligned} \sigma^2_{X(t)} &= \mathrm{Var}\{X(t)\} = \mathbb{E}\Big[\!\Big[(X(t) - \mu_X(t))^2\Big]\!\Big] \\ &= \mathbb{E}[\![X^2(t)]\!] - [\mu_X(t)]^2 \end{aligned} \tag{4.4}$$

The **standard deviation function** is simply given by $\sigma_X(t) = \sqrt{\mathrm{Var}\{X(t)\}}$. ♣

Remark

The mean, variance and standard deviation functions are **deterministic** (non random) functions of time t. This is because normal mean and variances are not random variables, they are simply plain numbers.

Example: Mean and Variance Functions of a Random Process

An ideal signal has the form $v_0 + a\cos(\omega_0 t + \theta_0)$ but amplitude variations may occur. We can model this situation with the random process:

$$X(t) = v_0 + A\cos(\omega_0 t + \theta_0)$$

where A is a random variable whose distribution describes the amplitude variation. This is often modeled as a Rayleigh random variable with parameter σ_A which has mean and variance:

$$\mathbb{E}[\![A]\!] = \sigma\sqrt{\frac{\pi}{2}} \qquad \mathrm{Var}\{A\} = \frac{4-\pi}{2}\sigma^2$$

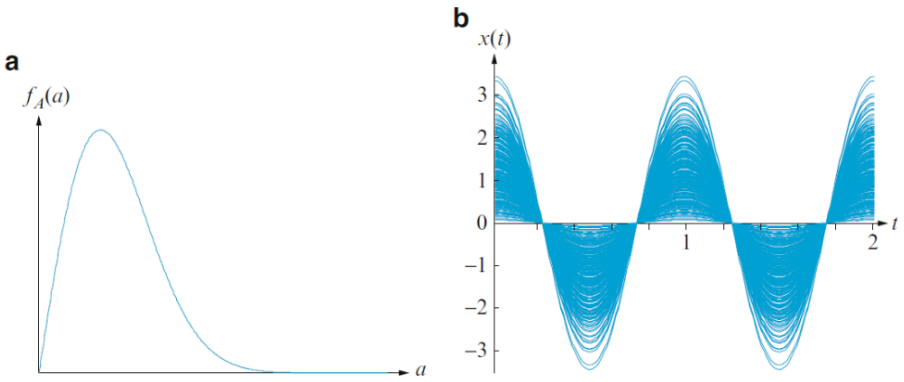


Figure 4.10: **a)**: Reyleigh p.d.f. for $\sigma_A = 1$; **b)**: Ensable of the random process $X(t)$ for the parameters $v_0 = 0, \omega_0 = 2\pi, \theta_0 = 0$.

Both the p.d.f. of the Reyleigh random variable A and the ensemble of the random process $X(t)$ are shown in Figure 4.10. If we fixed t we can apply the properties of expected value and variance to find the mean and variance functions of the process X.

Since, given any fixed t, we have that $\cos(\omega_0 t + \theta_0)$ is a **constant**, we can write:

$$\mu_X(t) = \mathbb{E}[\![X(t)]\!] = \mathbb{E}[\![v_0 + A\cos(\omega_0 t + \theta_0)]\!] = v_0 + \mathbb{E}[\![A]\!]\cos(\omega_0 t + \theta_0)$$
$$\sigma_X^2(t) = \mathrm{Var}\{X(t)\} = \mathrm{Var}\{A\cos(\omega_0 t + \theta_0)\} = \mathrm{Var}\{A\}\cos^2(\omega_0 t + \theta_0)$$

We can notice that, once the constants are fixed, both the mean and variance functions are **deterministic functions** that oscillate over time.

4.2.2. Auto-covariance Function

The mean and variance functions provide information about the behavior of the ensemble at each single point in time. However, they do not provide any information about the **dependence** between the random variables at different time points. Not surprisingly if we pick two distinct time point t, s we'll have that the random variables $X(t)$ and $X(s)$ are generally **related**. To capture this dependence we can use the **auto-covariance function** of the process.

Definition 4.4 (Auto-covariance Function)

The **auto-covariance function** of a random process $X(t)$ is defined as:

$$C_{XX}(t,s) = \mathrm{Cov}\{X(t), X(s)\} = \mathbb{E}[\![(X(t) - \mu_X(t))(X(s) - \mu_X(s))]\!] \tag{4.5}$$

♣

The auto-covariance function is typically denoted $\sigma_X(t,s)$ and, when $t = s$, we recover the variance function: $\sigma_X(t,t) = \sigma_X^2(t)$. Following we outline some properties of the auto-covariance function:

- $C_{XX}(t,s) = C_{XX}(s,t)$, that is the auto-covariance function is **symmetric**
- $C_{XX}(t,s) = \mathbb{E}[\![X(t)X(s)]\!] - \mu_X(t)\mu_X(s)$
- $\sigma_X^2(t) = \mathrm{Var}\{X(t)\} = \mathrm{Cov}\{X(t), X(t)\} = C_{XX}(t,t) = \mathbb{E}[\![X^2(t)]\!] - \mu_X^2(t)$

This function has the same interpretation of the covariance between two random variables. The problem is that covariance does not only contain information about the strength of the dependence between two variables, but also about their variance. To solve this problem we can introduce the **auto-correlation function** of the process.

Remark

If we take as a reference Figure 4.10 (b), we can ,according to the selection of time points t and s, observe different auto-covariance functions:

- if we pick $t = 0.5, s = 1.5$, we are going to have a positive auto-covariance since the values of the process at those times tend to vary in the same direction.

- if we pick $t = 0.5, s = 1.0$, we are going to have a negative auto-covariance since the values of the process at those times tend to vary in opposite directions.
- if we pick t, s such that all the ensemble paths cross the x axis at those times, we are going to have an auto-covariance close (if not equal) to zero. But this **does not mean** that $X(t) \perp X(s)$

4.2.3. Auto-correlation Function

For the purpose of this course, we are going to refer to auto-correlation in a very similar way as we did for random variables, just we consider a function of two time points instead of two random variables.

Definition 4.5 (Auto-correlation Function)

Given a random process $X(t)$, and two time points $t, s \in \mathcal{T}$, the **auto-correlation function** of the process is given by the following equation:

$$\rho_X(t, s) = \sigma_X \frac{t, s}{\sigma_X(t)\sigma_X(s)} \tag{4.6}$$

and its interpretation is analogous to that of the correlation between random variables. In particular $-1 \leq \rho_X(t, s) \leq 1$ indicates the magnitude and the direction of the association between $X(t) \wedge X(s)$.

♣

It is worth to mention that this is not the only possible interpretation of auto-correlation. Another common interpretation comes from signal processing. The alternative definition is given by $R_{XX}(t, s) = \mathbb{E}[\![X(t)X(s)]\!]$ and it is equivalent to $\rho_X(t, s)$ only when the mean and variance function are respectively equal to 0 and 1. To see why this is true, let X, Y be two random variables; the correlation between X and Y is given by:

$$\rho_{X,Y} = \frac{\text{Cov}\{X, Y\}}{\sqrt{\text{Var}\{X\}\ \text{Var}\{Y\}}} = \frac{\mathbb{E}[\![XY]\!] - \mathbb{E}[\![X]\!]\mathbb{E}[\![Y]\!]}{\sqrt{\text{Var}\{X\}\ \text{Var}\{Y\}}}$$

This can easily be retrieved via Equation 2.29 and Equation 2.31. Let's try to understand when it is the case that $\rho_{X,Y} = \mathbb{E}[\![XY]\!]$: first of all it is necessary that either one of $\mathbb{E}[\![X]\!]$ or $\mathbb{E}[\![Y]\!]$ is equal to zero; secondly the product of the variances in the denominator must be equal to 1.

Let now $\widetilde{X}$, $\widetilde{Y}$ be two random variables such as $\mu_{\widetilde{X}} = \mathbb{E}\left[\!\left[\widetilde{X}\right]\!\right], \sigma_{\widetilde{X}} = \text{std}\left(\widetilde{X}\right)$ and $\mu_{\widetilde{Y}} = \mathbb{E}\left[\!\left[\widetilde{Y}\right]\!\right], \sigma_{\widetilde{Y}} = \text{std}\left(\widetilde{Y}\right)$.

If we put $X = \frac{\widetilde{X} - \mu_{\widetilde{X}}}{\sigma_{\widetilde{X}}}$ and $Y = \frac{\widetilde{Y} - \mu_{\widetilde{Y}}}{\sigma_{\widetilde{Y}}}$, that is, we consider the standardized versions of $\widetilde{X}$ and $\widetilde{Y}$, we have that the two conditions mentioned before are satisfied, indeed we have $\mathbb{E}[\![X]\!] = \mathbb{E}[\![Y]\!] = 1$ and $\text{Var}\{X\} = \text{Var}\{Y\} = 1$; therefore $\rho_{X,Y} = \mathbb{E}[\![XY]\!]$.

4.2.4. Stationarity and Weak Stationarity

After defining some of the most important functions that characterize the behavior of a stochastic process, we now introduce a concept which tells us about the "**stability**" of the random process itself. We are referring to **stationarity** and **weak stationarity**.

Definition 4.6 (Stationary Stochastic Process)

A random process $X(t)$ is said to be **strictly stationary** if all of its statistical properties are **invariant** with respect to time. More precisely, $X(t)$ is stationary if, for any time points $t_1, ..., t_r$ and any value τ, the joint distribution of $X(t_1), ..., X(t_r)$ is the same as the joint distribution of $X(t_1 + \tau), ..., X(t_r + \tau)$.

♣

To verify strict stationarity it is therefore necessary to check that **all** finite-dimensional distributions of the process are invariant w.r.t. time shifts. This definition implies that all statistical properties of the process are also unchanged, since we require all joint distributions to be invariant over time.

Stationarity is very important: granting that the statistical properties of a process will not change over time allows us to make inferences about the future behavior of the process based on its past behavior. The problem of this definition is that it requires the invariance for any time frame considered, even infinitely large ones. This is often too restrictive for practical applications, where we usually deal with **finite time frames**.

Strict stationarity is also very hard to verify in practice, since it requires the knowledge of all finite-dimensional distributions of the process, along with some knowledge about the characterizing functions (mean, variance, auto-covariance, etc.). To overcome this, we introduce a weaker form of stationarity.

Definition 4.7 (Wide Sense Stationarity)

A random process $X(t)$ is said to be **weakly stationary** (or **stationary in the wide sense**) if the following two conditions hold:

- The mean function of $X(t)$, $\mu_X(t)$ is a constant, i.e. there exists a constant μ such that $\mu_X(t) = \mu$ for all $t \in \mathcal{T}$.
- The auto-covariance function of $X(t)$, $C_{XX}(t, s)$ depends only on the time difference $s - t$

♣

While the first condition of Definition 4.7° is quite straightforward to understand, the second one may require some further explanation. The idea is that we require that the degree of association between $X(t)$ and $X(s)$, measured by the auto-covariance, depends only on the distance between the times s, t but not on the positino of those times on an absolute scale.

4.3. Markov Processes

In this course we are going to focus on a very specific kind of stochastic process called **Markov process**. In this section we are going to introduce the basic definitions and properties of these processes.

Definition 4.8 (Markov Process)

A stochastic process $X(t)$ is **Markov** if for any $t_1 < ... < t_n < t$ and any sets $A; A_1, ..., A_n$ we have that:

$$\begin{aligned} &\mathbb{P}[\![X(t) \in A \mid X(t_1) \in A_1, \ ..., \ X(t_n) \in A_n]\!] \\ &\qquad = \mathbb{P}[\![X(t) \in A \mid X(t_n) \in A_n]\!] \end{aligned} \tag{4.7}$$

♣

To put this in practical terms, the Markov property states that, knowing the present, there is no additional information from the past that can be used to better predict the future:

$$\mathbb{P}[\![\text{future} \mid \text{past}, \ \text{present}]\!] = \mathbb{P}[\![\text{future} \mid \text{present}]\!]$$

For the future development of a Markov process, only its present state is important, and it does not matter how the process arrived to this state.

Figures Index