Probability course

Louis Gass

December 23, 2023

Contents

1 Some enumeration

1.1 basic rules

Enumerating, is the art of counting the number of elements in a set. Its mastery often results in pretty formulas, sometimes effortlessly. It does, however, require a little intuition, as it can be easy to get the counting wrong.

We're going to look at a few "basic" rules for learning to count. We won't go into all of them, because they're "obvious", but they already enable you to count quite a few sets.

Définition 1.1. The **cardinal** of a set A is the number of elements contained in A.

At our level, we distinguish three possibilities for the cardinal of a set.

A set is either of cardinal

- **countable**, which we separate into two categories:
  - **finite**: there is a finite number of elements in the set.
  - **infinite countable** : there is an infinite number of elements that can be listed in a sequence indexed by the natural integers; the *first*, the *second*, and so on. This applies to natural integers, rational fractions, etc.
- **uncountable**: there is an infinite number of elements, and they cannot be listed in a sequence indexed by integers. This is the case for the set of real numbers, the power set of $\mathbb{N}$, etc.

We describe some basic rules about cardinals.

If A and B are two sets in bijection, then :

$$\text{Card}(A) = \text{Card}(B)$$

This principle reduces the enumeration of A to a simpler enumeration of B. For example, consider a walker moving on a grid either up 1 square or to the right 1 square. He starts at point $(0, 0)$ and arrives at point $(2, 2)$. How many possible paths are there? There are as many as there are anagrams of the word $RURU$, i.e. 6.

Let A and B be two finite sets. Let $A \times B$ be the set of pairs:

$$A \times B = \{(a, b) \mid a \in A, b \in B\}$$

Then :

$$\text{Card}(A \times B) = \text{Card}(A) \times \text{Card}(B)$$

If I throw a dice and a coin, how many possible outcomes are there? Answer: $6 \times 2 = 12$. In probability, this is used to calculate the probabilities of two successive independent experiments. There is one white ball and two black balls in an urn. I draw a ball at random and put it back, then draw a second ball at random. What is the probability of drawing two black balls in a row? Answer:

$$\frac{2 \times 2}{3 \times 3} = \frac{4}{9}$$

Let A and B be two subsets of a finite set Ω. If A and B are disjoint then :

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B)$$

In particular, if A and B form a partition of Ω [1], so $\text{Card}(B) = \text{Card}(\Omega) - \text{Card}(A)$. This leads to the following formula for A and B which are not necessarily disjoint:

$$\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) - \text{Card}(A \cap B)$$

If I throw two dice at random, what is the probability that at least one of my dice is 6? There are 36 possible equiprobable outcomes, denoted (n, m) where n (resp. m) is the outcome of the first (resp. second) dice, with n, m being in $\{1, \ldots, 6\}$. Then the probability of rolling a 6 is :

$$\frac{6 + 6 - 1}{36} = \frac{11}{36}$$

Let Ω be a set and A a subset of Ω. Let $\mathbb{1}_A$ be the **indicator function** of the subset A, i.e. the function

$$\mathbb{1}_A : \Omega \longrightarrow \{0, 1\}$$
$$a \longrightarrow \begin{cases} 1 \text{ if } a \in A \\ 0 \text{ if } a \notin A \end{cases}$$

This function has an important theoretical purpose, so it's essential to know its definition. For example, the function

$$\mathbb{1}_{[a,b[} : \mathbb{R} \to \mathbb{R}$$

is the function equal to 1 on the semi-open interval $[a, b[$ and 0 outside it. Another example,

$$\text{Card}(A) = \sum_{e \in \Omega} \mathbb{1}_A(e).$$

Let Ω be a set. Let $\mathcal{P}(\Omega)$ be the power set of Ω, i.e. the set :

$$\mathcal{P}(\Omega) = \{A \mid A \text{ is a subset of } \Omega\}$$

Then :

$$\text{Card}(\mathcal{P}(\Omega)) = 2^{\text{Card}(\Omega)}$$

Let $\omega_1, \ldots, \omega_n$ be the elements of Ω. Then $\mathcal{P}(\Omega)$ can be put in bijection with the product set of n terms:

$$\{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\} = \{0, 1\}^{\text{Card}(\Omega)}$$

Through the mapping

$$\mathcal{P}(\Omega) \longrightarrow \{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\}$$
$$A \longrightarrow (\mathbb{1}_A(\omega_1), \mathbb{1}_A(\omega_2), \ldots, \mathbb{1}_A(\omega_n))$$

For example, if $\Omega = \{1, 2, 3, 4\}$ then the subset $A = \{2, 4\}$ is associated with $(0, 1, 0, 1)$. And, of course, we have :

$$\text{Card}(\{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\}) = 2^n = 2^{\text{Card}(\Omega)}$$

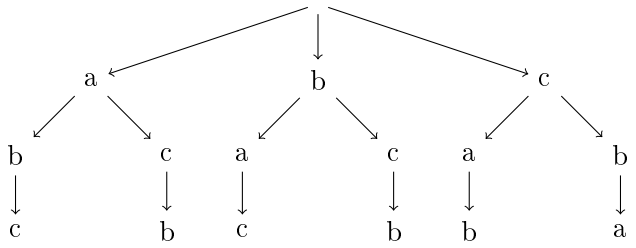
[1] The sets A and B form a partition of Ω if A and B are disjoint and $A \cup B = \Omega$. This definition generalizes to a family of subsets $(A_1, \ldots, A_n)$ of Ω. This family forms a partition of Ω if the sets $A_1, \ldots, A_n$ are two-by-two disjoint and their union is Ω. In other words, we have split Ω into n pieces $A_1, \ldots, A_n$.

1.2 Permutation

Let Ω be a set of n elements. By default, the set Ω is not *ordered* and without repetition: we can either write $\Omega = \{a, b, c\}$ or $\Omega = \{b, c, a\}$ or $\Omega = \{a, b, c, a, b, b\}$. When we write "Let $\omega_1, \ldots, \omega_n$ be the elements of Ω." this is in fact a way of ordering the set Ω: the first is ω_1, the second is ω_2, and so on. Of course, there isn't just one way of ordering Ω.

Définition 1.2. A **permutation** of the set Ω is a way of arranging the elements of Ω in an ordered way. Equivalently, it is a bijection of Ω on the set $\{1, \ldots, n\}$ (when Ω is not the empty set).

For example, there are six ways of arranging the set $\Omega = \{a, b, c\}$. We choose the first element: there are 3 possibilities among $\{a, b, c\}$. Then we choose the second: there are two possibilities among the two remaining elements to be placed. Finally, the last element is automatically placed last: so we have $3 \times 2 \times 1 = 6$ possibilities.



Théorème 1.1. *There are exactly $n!$ permutations of the set Ω, where $n!$ is the factorial of the integer n defined by :*

$$n! = 1 \times 2 \times \ldots \times n$$

By convention, $0! = 1$. There's only one way to order a set with 0 elements.

Proof. The reasoning is analogous to the previous case: there are n possibilities for choosing the first element, then $n - 1$ for the second, and so on. In total, there are $n \times (n - 1) \times \ldots \times 1 = n!$ possibilities for ordering the set Ω. □

How many anagrams of the word $PROBA$ are there? All the letters are different, so each permutation gives a different word. Therefore, $5! = 120$ anagrams of the word $PROBA$.

Same question with the word $ECOLE$. This time, you have to be careful, because swapping the two "E" will give you the same word. Let's note E_1 and E_2 the two "E". As before, there are 120 permutations for the word E_1COLE_2. However, each word is counted twice, since for an anagram of the word $ECOLE$ (e.g. $CEOEL$) we can construct CE_2OE_1L and CE_1OE_2L. There are therefore $120/2 = 60$ possibilities.

Same question with the word $MISSISSIPPI$. There are 11 letters but many of them are repeated: $1 \times M + 4 \times I + 4 \times S + 2 \times P$. So we write $MI_1S_1S_2I_2S_3S_4I_3P_1P_2I_4$. This word has 11! distinct permutations. Let's try to find the number of permutations of $MI_1SSI_2SSI_3P_1P_2I_4$ (all S are now assumed to be identical). Given a permutation of this word, we can construct exactly 4! for the word $MI_1S_1S_2I_2S_3S_4I_3P_1P_2I_4$. For example, from $SSSSMI_1I_2I_3I_4P_1P_2$ we can construct :

- $S_1S_2S_3S_4MI_1I_2I_3I_4P_1P_2$
- $S_2S_1S_4S_3MI_1I_2I_3I_4P_1P_2$

- ...

In fact we can swap all the S between them, which will give me exactly 24 distinct permutations of the word $MI_1S_1S_2I_2S_3S_4I_3P_1P_2I_4$. So there are exactly $\frac{11!}{4!}$ permutations of the word $MI_1SSI_2SSI_3P_1P_2I_4$. We can reiterate with the remaining letters. In the end, the word $MISSISSIPPI$ will have :

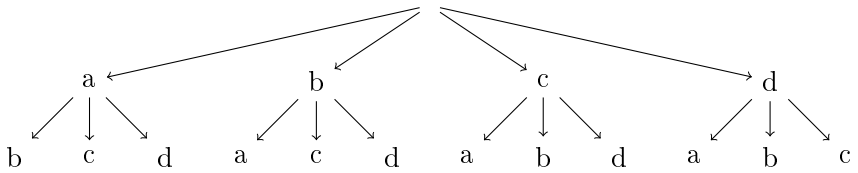
$$\frac{11!}{1!4!4!2!} = 34\,650$$

anagrams.

1.3 Arrangement

Définition 1.3. An **arrangement** of k elements taken from the n elements of a set Ω is an ordered sequence of k distinct elements of Ω.

For example (a, c, f) and (c, f, a) and (b, a, f) are distinct arrangements of 3 elements among the 6 elements $\{a, b, c, d, e, f\}$. An arrangement of n elements from a n-element Ω set is simply a permutation of Ω. For example, there are 2 possibilities for arranging 2 elements among 4: we have 4 possibilities for choosing the first, then 3 for the second.



Théorème 1.2. *There are exactly :*

$$A_n^k = n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}$$

ways of arranging k elements among n elements.

Proof. The reasoning is analogous to the previous case: there are n possibilities for choosing the first element, then $n-1$ for the second, and so on. In total, there are $n \times (n-1) \times \ldots \times (n-k+1)$ possibilities for arranging k elements among n. □

For example, in a race with 20 participants, there are :

$$20 \times 19 \times 18 = 6840$$

distinct podiums possible.

1.4 Combination

Définition 1.4. A **combination** of k elements from a n-element Ω set is a subset A of Ω containing k elements.

The difference between a combination and an arrangement is that a combination is a set, and therefore *unordered*. In an enumeration problem where order is important (combination of an access code, podium, etc.), we use the notion of arrangement instead. In a problem where order doesn't matter (number of white balls drawn in a toss, probability of having 2 girls knowing you

have 5 children, etc.) it's the notion of combination that's important.

The number of combinations of k elements among a set of n elements (we call this number "n choose k") is denoted by $\binom{n}{k}$.

Théorème 1.3. *We have*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Proof. Given a combination of k elements, we can form exactly $k!$ distinct arrangements by permuting the k elements chosen. This gives the equality :

$$k!\binom{n}{k} = A_n^k = \frac{n!}{(n-k)!}$$

□

If $k > n$ we set

$$\binom{n}{k} = 0$$

In general, all enumeration formulas have two proofs: a computational proof, and a proof using only enumeration tools. It's a good thing to master both: the computational proof is generally (but not always) longer, while the enumeration proof is often prettier and more intuitive (but beware of errors in reasoning!).

Théorème 1.4. *Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$. Then :*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Proof. We can prove this by recurrence on n, but let's proceed with a combinatorial proof instead. We develop :

$$\begin{aligned}(x+y)^n &= (x+y) \times (x+y) \times (x+y) \\ &= \sum_{k=0}^{n} a_k x^k y^{n-k}\end{aligned}$$

When we expand, we end up with monomials of the form $x^k y^{n-k}$ for k varying between 0 and n. To obtain such a monomial, we choose k factors among the n and expand according to "x", and for the remaining $(n-k)$ factors we expand according to "y". All that remains is the monomial $x^k y^{n-k}$. The number of ways to choose k factors from n is exactly the binomial coefficient k from n, hence :

$$a_k = \binom{n}{k}$$

□

In particular,

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

2 Probability space

2.1 Some heuristic

Probability theory aims to quantify the notion of uncertainty. The notion of uncertainty reflects a lack of information that prevents us from predicting the outcome of an experiment with absolute certainty.

Let's take the example of a coin toss. Intuitively, we'll come up tails about half of the time. But what's the reality? The coin toss, as a physical system, can be considered deterministic. If I could describe with absolute precision the action exerted by my hand on the coin, I could deduce its exact trajectory and predict on which face the coin would fall. Let's assume that the coin's trajectory depends on a parameter $\theta \in [0, 1]$. Since we have no information on the parameter θ, we'll assume that it is chosen *randomly* according to some distribution (for example, a uniform distribution). In this case, we'd observe that for about half the values of θ the coin will land on heads, and for the other half it will land on tails.

Let's take the example of a poker player. Given his deck and the cards on the table, he must decide whether or not to continue playing. Since he doesn't know his opponents' cards, he assumes that the pack of cards used in the game has been chosen uniformly from all possible packs (52!). From this, it is possible to calculate the proportion of card packs that are favorable to him.

> Probability theory enables us to quantify a lack of information about the parameters of a given experiment. This lack of information translates into uncertainty about the outcome of the experiment. To give ourselves a framework for making probabilities is to give ourselves a *model of the uncertainty* we face.

When playing Uno, the rules of the game mean that the stack of cards at the end of a game shows many cards of the same value in succession. When shuffled by hand, some of these sequences are not broken, and the shuffled pack generally has a higher proportion of cards of the same value in sequence than if the shuffle had been "perfect". With no prior knowledge of the deck other than that it has just been shuffled following a previous game, an experienced player will model the card deck as one chosen *at random* from among the 52! possible decks, but this random choice will not be uniform across all decks. Those with cards of the same color in sequence will be considered to have a higher probability of occurrence.

In order to define a framework for doing probability we need three ingredients.

> We need to give ourselves the *universe* of possible configurations for the experiment or game under consideration.

For a dice roll, it could be the trajectory of my hand. For a card game, it's the set of all possible decks. Generally speaking, it's a big, complicated space that's difficult, if not impossible, to describe precisely.

> We need to define the notion of *information*. It's the set of events we want to consider in an experiment.

For example, we take a survey of a population. Each person is asked to specify his or her age category: between 0 and 20 years old, between 20 and 50 years old, over 50 years old. Here, the universe is all the people surveyed. Our survey allows us to quantify the probability of occurrence of events such as "The person is under 50". On the other hand, we have no access to quantities such as "The person is between 40 and 60", let alone "The person is a man". This notion of

information seems a little superfluous, as it always seems possible to "enlarge" the experiment in order to access information we don't have. There are at least three reasons for not doing so:

- This distorts/complicates the model. In the case of a dice roll, we don't really care whether the dice has made three turns on itself before hitting the ground. The essential information is only the final value of the dice.
- When we model a game, or more generally a quantity that evolves over time (the stock market price, a quantum particle, Markov chain, etc.), this notion of information takes on its full importance in understanding the game (are the rules in my favor?) and devising a strategy (should I draw a card, sell my shares?). The general framework is that of the theory of stochastic processes.
- There are certain mathematical obstructions that prevent us from considering the probability of realization of certain sets in the case where the universe Ω is uncountable, as shown by the famous Banach-Tarski paradox.

We need to define a notion of *measure of probability*. It quantifies the probability of a given event to occur.

In the case of a well-mixed pack of cards, for example, we can choose a uniform probability over all the decks of cards, meaning that each card shuffle has a probability $1/(52!)$, but if it's badly mixed, the probability may be chosen differently. Modeling an experiment therefore requires choosing the probability with which each event occurs. This measure is supposed to reflect the physical reality of the experiment as closely as possible, which can prove complicated. A statistical test can be used to check the agreement between a theoretical probabilistic model and a real experiment repeated a large number of times.

2.2 Some definitions and examples

According to the previous discussion, a probability space consists of three ingredients.

Définition 2.1. A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where :

- Ω is a set
- $\mathcal{F}$ is a σ-algebra on Ω.
- $\mathbb{P}$ is a probability measure

Let's take a closer look at the three points of the definition.

• Ω is the **universe**, i.e. the set of possible configurations of my experiment. Mathematically, it's simply a set.

• $\mathcal{F}$ represents the information we can acquire during the experiment. An element of $\mathcal{F}$ is called a **event**. An event is a subset of the Ω universe. The set $\mathcal{F}$ is therefore a set of events. Morally, it is possible to apply certain operations between events: union, intersection, difference, complementary, etc. In other words, the set $\mathcal{F}$ is stable by a number of operations.

Définition 2.2. A σ-algebra $\mathcal{F}$ on a space Ω is a subset of $\mathcal{P}(\Omega)$ (the power set of Ω) such that :

- $\Omega \in \mathcal{F}$ (the universe is an event)
- If $A \in \mathcal{F}$ then $\overline{A} \in \mathcal{F}$ (stability by passing to the complementary)
- If $(A_n)_{n\geq 0} \in \mathcal{F}$ then $\bigcup_{n\geq 0} A_n \in \mathcal{F}$ (stability by countable union)

The pair $(\Omega, \mathcal{F})$ is a **measurable space**.

- $\mathbb{P}$ is used to quantify the probability of a given event occurring. For a given event A, it associates a number between 0 and 1, denoted $\mathbb{P}(A)$, which reflects the probability of the event A occurring. To be consistent with an intuitive notion of a *measure*, $\mathbb{P}$ must verify a number of properties:

Définition 2.3. A **probability measure** $\mathbb{P}$ on a measurable space $(\Omega, \mathcal{F})$ is an application :

$$\begin{aligned} \mathbb{P} : \mathcal{F} &\longrightarrow [0,1] \\ A &\longrightarrow \mathbb{P}(A) \end{aligned}$$

Such as:

- $\mathbb{P}(\Omega) = 1$ (the universe is an event with probability 1)
- If $(A_n)_{n\geq 0}$ is a countable family of pairwise disjoint events, then :

$$\mathbb{P}\left(\bigcup_{n=0}^{+\infty} A_n\right) = \sum_{n=0}^{+\infty} \mathbb{P}(A_n)$$

If A is an event in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}(A) = 1$, we say that A is realized *almost-surely* (we sometimes write a.s.). Conversely, if A is an event in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}(A) = 0$, we'll say that A is a *negligible* event. This terminology is important and will be used many times throughout the course.

Let's look at some basic examples of σ-algebras.

Définition 2.4. The σ-algebra $\{\emptyset, \Omega\}$ is called the **trivial σ-algebra**. It is the smallest σ-algebra on Ω that we can consider. The σ-algebra $\mathcal{P}(\Omega)$ is called the **discrete σ-algebra**. It is the largest σ-algebra on Ω that can be considered.

When Ω is countable, we generally consider the discrete σ-algebra on Ω.

Définition 2.5. Let $\mathcal{C}$ be a subset of $\mathcal{P}(\Omega)$. Let $\sigma(\mathcal{C})$ be the smallest σ-algebra containing $\mathcal{C}$. It is the intersection of all σ-algebras containing $\mathcal{C}$.

For example, if Ω is countable and $\mathcal{C}$ is the set of singletons, it's easy to see that $\sigma(\mathcal{C}) = \mathcal{P}(\Omega)$.

Let $\Omega = [0,1]$. If I is an interval with ends a and b (not necessarily open or closed), then $\mu(I) = b - a$. This definition is "consistent" with the notion of a probability space:

$$\mu(\Omega) = 1$$

I_1 and I_2 are disjoint intervals such that $I_1 \cup I_2$ is an interval (e.g. $I_1 = [a, b]$ and $I_2 =]b, c]$) then :

$$\mu(I_1 \cup I_2) = \mu(I_1) + \mu(I_2)$$

Définition 2.6. Let $\mathcal{B}([0,1])$ be the σ-algebra generated by the intervals included in $[0,1]$. It is called the **Borelian σ-algebra** on $[0,1]$. Similarly, $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by the intervals in $\mathbb{R}$.

Théorème 2.1. *(Carathéodory extension) there exists a unique probability measure (still denoted μ) on $(\Omega, \mathcal{B}([0,1]))$ such that :*

- *$([0,1], \mathcal{B}([0,1]), \mu)$ is a probability space.*
- *$\mu(I) = b - a$ for any interval I of extremity a and b.*

Proof. Accepted. □

In other words, if we decide that the measure of an interval is exactly its length, then we can extend this measure to the entire σ-algebra generated by intervals: countable unions of intervals, singletons, ... and much more!

On the other hand, there are subsets of Ω that are not in $\mathcal{B}([0,1])$ (a counterexample is complicated to exhibit). In fact, it can be shown that it is impossible to consistently extend this measure to the entire σ-algebra $\mathcal{P}([0,1])$ (Vitali's counterexample). For this reason, we must restrict ourselves to the Borelian σ-algebra.

2.3 Conditional probability and independence

When we want to model two distinct quantities by a probabilistic model, it often happens that these two quantities are *correlated*. For example, an ice cream vendor's daily sells are strongly correlated with temperature. Quantifying this dependence enables the ice cream vendor to better adjust his stocks so as not to run out on hot days and lose out on cooler days. Consider the following events:

- A : the vendor sells over 100 ice creams during the day
- B: the day's temperature has exceeded 20 degrees.

The salesman models his situation as follows. From a year's analysis of sells he estimates that

- $\mathbb{P}(A) = 0.5$
- $\mathbb{P}(B) = 0.4$
- $\mathbb{P}(A \cap B) = 0.3$

In the evening, the next day's weather forecast calls for temperatures of over 20. How can we quantify the probability that the seller will sell more than 100 of ice cream the next day? Let's denote $\mathbb{P}(A|B)$ this probability. Then we have :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 0.75.$$

The universe of possibilities for the following day is no longer the whole space Ω, but B, the set of configurations for which the temperature exceeds 30 degrees over the course of the day. So, in a way, we've changed the probabilistic space:

$$(\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (B, \mathcal{F}_B, \mathbb{P}(\,.\,|B))$$

Where :

$$\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$$

In this example, the notion of dependence is explained by a notion of *causality*:

$$\text{The weather's fine} \Rightarrow \text{People are buying ice cream}$$

But that's not always the case. For example, we're doing a survey of people living near electricity pylons[2]. We note:

- A: "the person falls ill more than 5 times during the year".
- B: "the person lives within 200m of an electric pylon".

Over a whole city, we notice that :

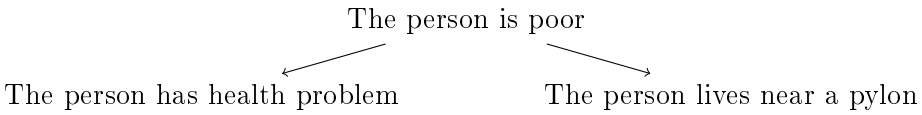
- $\mathbb{P}(A) = 0.2$
- $\mathbb{P}(B) = 0.1$
- $\mathbb{P}(A \cap B) = 0.04$

Then :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 0.4.$$

In other words, a person is twice as likely to fall ill frequently if they live near an electricity pylon. Does this mean that electricity pylons are dangerous for your health?

Not necessarily, because in reality the population living near electricity pylons is poorer on average, and poorer people have less access to health care. Here, events A and B are correlated, but not necessarily causally. In fact, these two events are correlated with a third event C: "the person is below the poverty line". Instead, we have the following plausible cause-and-effect relationships between A and B.



A more blatant example: 100% of people who drink water die... Don't drink water!

Définition 2.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and B an event of non-zero measure. Then for any event A, we call the quantity :

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

It indicates the probability of an event A occurring knowing that event B has occurred.

[2] Adapted from a real survey.

Définition 2.8. A family $\{B_1, \ldots, B_n\}$ is said to be a **complete system of events** if :

- $\forall i \in \{1, \ldots, n\},\ \mathbb{P}(B_i) \neq 0$
- $\forall i, j \in \{1, \ldots, n\},\ \mathbb{P}(B_i \cap B_j) = 0$
- $\mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = 1 \quad$ or equivalently $\quad \sum_{i=1}^{n} \mathbb{P}(B_i) = 1$

In other words, the family $\{B_1, \ldots, B_n\}$ forms a *probabilistic* partition of the universe Ω. It is possible to adapt the definition for a countable family of events.

For example, if $0 < \mathbb{P}(A) < 1$ then the family $\{A, \overline{A}\}$ forms a complete system of events. Similarly, provided the following events have non-zero probabilities, the following family :

$$\{A \cap B, \overline{A} \cap B, A \cap \overline{B}, \overline{A} \cap \overline{B}\}$$

is a complete system of events.

Théorème 2.1 (Law of total probability)**.** *Let $\{B_1, \ldots, B_n\}$ be a complete system of events, and A be any event. Then :*

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

Théorème 2.2 (Bayes' formula)**.** *Let A and B be two events of non-zero probability. Then we have the identity :*

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

If the family $\{A_1, \ldots, A_n\}$ is a complete system of events, then for $1 \leq i \leq n$ we have :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

Bayes' formula makes it possible, by knowing the *effects*, to trace back to a probability on the *causes*. It is the basis of Bayesian inference.

The notion of conditional probability leads naturally to the notion of *independence*. Let B be an event with non-zero probability and A any event. Event A is independent of event B if the probability of its occurrence does not depend on whether or not B has occurred. In other words:

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

If two dice are rolled, knowledge of the value of the first dice has no influence on the value of the second dice. The events "the first dice is a 6" and "the second is even" are independent.

Définition 2.9. Two events A and B are said to be **independent** if :

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

We then note $A \perp B$. More generally, two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are independent if

$$\forall A \in \mathcal{F},\ \forall B \in \mathcal{G}, \quad \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

and we note

$$\mathcal{F} \perp \mathcal{G}.$$

We can also speak of independence for a family of events.

Définition 2.10. Events $A_1, \ldots, A_n$ are said to be **mutually independent** if for any subset I of $\{1, \ldots, n\}$ we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Similarly, the σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are mutually independent if for any events $A_1, \ldots, A_n$ such that :

$$A_1 \in \mathcal{F}_1 \ , \ \ldots \ , \ A_n \in \mathcal{F}_n$$

The events $A_1, \ldots, A_n$ are independent.

It's easy to see that events $A_1, \ldots, A_n$ are independent if and only if the σ-algebras $\sigma(A_1), \ldots, \sigma(A_n)$ are independent. Indeed, for an event A we have :

$$\sigma(\{A\}) = \{\emptyset, A, \overline{A}, \Omega\}$$

For example, if A and B are independent events, then

$$\begin{aligned} \mathbb{P}(\overline{A} \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(\overline{A})\mathbb{P}(B) \end{aligned}$$

Théorème 2.3. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$, $\mathcal{G}_1, \ldots, \mathcal{G}_m$ be mutually independent σ-algebras. Then :*

$$\sigma\left(\bigcup_{i=1}^{n} \mathcal{F}_i\right) \perp \sigma\left(\bigcup_{j=1}^{m} \mathcal{G}_j\right)$$

Proof. Admitted. □

For example, let A_1, A_2, A_3, A_4, A_5 be mutually independent events. Then :

$$(A_1 \cup A_3) \perp (A_2 \cap (\overline{A_4} \cup A_5))$$

Generally speaking, if we separate a family of mutually independent events into two groups, then an event constructed from the first group and the operations $\cap, \cup$ and $\overline{\ }$ will be independent of an event constructed from the second group and the same operations. Mutual independence must be understood in terms of the independence of information. The information provided by knowledge of A_1 and A_3 is independent of the information generated by knowledge of A_2, A_4 and A_5.

We can also define the independence of a countable family of events.

Définition 2.11. A family $(A_n)_{n \in \mathbb{N}}$ of events is said to be mutually independent if for any *finite* subset I of $\mathbb{N}$ we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Similarly, the σ-algebras $(\mathcal{F}_n)_{n \in \mathbb{N}}$ are mutually independent if for any family of events $(A_n)_{n \in \mathbb{N}}$ such that :

$$\forall n \in \mathbb{N}, \ \ A_n \in \mathcal{F}_n$$

the events $(A_n)_{n \in \mathbb{N}}$ are mutually independent.

2.4 Some limit theorems

The aim of this section is to look at some "limit" probabilities. We recall the definition of the limit superior and limit inferior of a sequence of sets.

Définition 2.1. Let $(A_n)_{n\geq 0}$ be a sequence of sets. The limit superior of the sequence is defined as

$$\limsup_{n\geq 0} A_n = \bigcap_{n\geq 0} \bigcup_{k\geq n} A_k,$$

and the limit inferior of the sequence as

$$\liminf_{n\geq 0} A_n = \bigcup_{n\geq 0} \bigcap_{k\geq n} A_k.$$

The interpretations of these two quantities are as follows:

- An element $\omega \in \Omega$ belongs to the set $\limsup_{n\geq 0} A_n$ if and only if ω belongs to an infinite number of events in the sequence $(A_n)_{n\geq 0}$.
- An element $\omega \in \Omega$ belongs to the set $\liminf_{n\geq 0} A_n$ if and only if ω belongs to all events of the sequence $(A_n)_{n\geq 0}$ up from a certain rank.

We start with the following fundamental theorem.

Théorème 2.2 (Borel–Cantelli)**.** *Let $(A_n)_{n\geq 0}$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If*

$$\sum_{n\geq 0} \mathbb{P}(A_n) < +\infty,$$

then

$$\mathbb{P}(\limsup_{n\geq 0} A_n) = 0.$$

If the events $(A_n)_{n\geq 0}$ are mutually independent and

$$\sum_{n\geq 0} \mathbb{P}(A_n) = +\infty,$$

then

$$\mathbb{P}(\limsup_{n\geq 0} A_n) = 1.$$

Proof. The proof is not difficult, but we'll admit it. □

In other words, if the series of term $(\mathbb{P}(A_n))_{n\geq 0}$ converges, then the events in the sequence $(A_n)_{n\geq 0}$ will almost surely only occur a finite number of times. Conversely, if we add the assumption of independence of the sequence $(A_n)_{n\geq 0}$ and the series of term $(\mathbb{P}(A_n))_{n\geq 0}$ diverges, then the events of the sequence $(A_n)_{n\geq 0}$ occur infinitely often almost surely. This theorem is of fundamental importance for proving almost-sure convergences of random variables, such as the strong law of large numbers.

In the following, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $(\mathcal{F}_n)_{n\geq 0}$ a sequence of mutually independent σ-algebras included in $\mathcal{F}$.

Définition 2.2 (Asymptotic σ-algebra)**.** We define the asymptotic σ-algebra

$$\mathcal{F}_\infty = \bigcap_{n\geq 0} \sigma\left(\bigcup_{k\geq n} \mathcal{F}_k\right).$$

The asymptotic σ-algebra groups events that depend only on the "tail" of the σ-algebra sequence $(\mathcal{F}_n)_{n\geq 0}$, i.e. that do not depend on the first elements of the σ-algebra sequence. An event in $\mathcal{F}_\infty$ is called a *asymptotic* event.

Théorème 2.3 (Kolmogorov's $0-1$ law)**.** *Let A be an event of the asymototic σ-algebra $\mathcal{F}_\infty$. Then*

$$\mathbb{P}(A) \in \{0, 1\}.$$

Proof. The proof is not difficult, but we'll admit it for the purposes of this course. It relies on the fact that the σ-algebra $\mathcal{F}_\infty$ is independent of itself. □

In other words, an event in the asymptotic σ-algebra either almost surely occurs, or almost surely never occurs. There is no in-between possibility. On the other hand, it can be difficult to know in which case we are. Typical examples of asymptotic events are: "Is a sequence of independent random variables bounded? does it converge?" We'll be looking at some applications of these two theorems during this course.

3 Random variable

3.1 Definitions

A random variable X is an application to which a configuration, a given outcome (thus an element of Ω) associates an object, a number, a color, etc.

> A random variable models the different values that the outcome of a random experiment can take.

For example, consider the random variables associated with the following experiments.

- The roll of the dice. Consider the random variable $X : \Omega \to \{1, 2, 3, 4, 5, 6\}$ the application to which a given configuration associates the value of the dice.
- The coin toss. Consider the random variable $X : \Omega \to \{\text{"Head"}, \text{"Tail"}\}$ the application to which a given configuration associates the value of the coin.
- The dart game. Consider the random variable $X : \Omega \to \mathbb{R}_+$ the application to which a given configuration associates the distance of the dart from the center of the target.
- The throw of two dice. Consider the random variable $X : \Omega \to \{2, \dots, 12\}$ the application to which a given configuration associates the sum of the two dice.

To obtain information on the probability of a random variable belonging to a certain set of values, it must satisfy certain constraints. For example, in the case of heads or tails, we need to ensure that the subset

$$\{X = \text{"Head"}\} = \{\omega \in \Omega \mid X(\omega) = \text{"Head"}\}$$

is an event, i.e. an element of the σ-algebra $\mathcal{F}$. Indeed, the proposition "the probability that the coin lands on tails is 1/2" is written mathematically :

$$\mathbb{P}(X = \text{"Tails"}) = 1/2.$$

In the case where X has a value in a countable space E, there are no difficulties. We require that for any x in E the set

$$\{X = x\} = \{\omega \in \Omega \mid X(\omega) = x\}$$

is an event (i.e. belongs to the σ-algebra $\mathcal{F}$). Since any subset U of E can be written as a countable union of singletons, then :

$$\{X \in U\} = \bigcup_{x \in U} \{X = x\}$$

And thus

$$\forall U \subset E, \ \{X \in U\} \subset \mathcal{F}$$

If X is real-valued (or in an uncountable space), things get a little more complicated. If we ask only that the sets $\{X = x\}$ are events, we can't assert that the set :

$$\{X \in [0, 1]\} = \bigcup_{x \in [0,1]} \{X = x\}$$

is an event, since a σ-algebra is only stable by countable union, and the set $[0, 1]$ is uncountable. Conversely, we cannot[3] require that for any subset U of $\mathbb{R}$, the set :

$$\{X \in U\} = \{\omega \in \Omega \mid X(\omega) \in U\}.$$

[3] We couldn't define a proper notion of uniform random variable this way.

is an event. We need to restrict the collection of subsets we're allowed to look at. The minimum we can require is to be able to make sense of propositions like "the probability that my dart arrives within 20cm of the target is 1/2". To achieve this, the set :

$$\{X \leq 20\} = \{\omega \in \Omega \mid X(\omega) \leq 20\}$$

is an event. In general, we require that for any $x \in \mathbb{R}$, the set :

$$\{X \leq x\} = \{X \in]-\infty, x]\}$$

is an event. From this it is easy to show [4] that the following subsets of Ω are events:

$$\{X < x\} = \bigcup_{n \geq 0} \left\{ X \leq x - \frac{1}{n} \right\}$$

$$\{X > x\} = \overline{\{X \leq x\}}$$
$$\{X \geq x\} = \overline{\{X < x\}}$$
$$\{x \leq X \leq y\} = \{X \in [x, y]\} = \{X \geq x\} \cap \{X \leq y\}$$
$$\{X = x\} = \{x \leq X \leq x\}$$

In particular, the set :

$$\{X \in U\}$$

is an event when U is an interval, but also a countable union of intervals, singletons, ... and much more! In fact, the set of U for which this set is an event contains the entire σ-algebra generated by the intervals, which we call the Borelian σ-algebra on R, denoted $\mathcal{B}(\mathbb{R})$ (we've already seen this concept in the previous chapter).

Let's try to generalize. Let X be a random variable with a value in some space E. We need to decide which are the subsets U of E for which the set $\{X \in U\}$ is an event (i.e. an element of the σ-algebra $\mathcal{F}$). This is equivalent to choosing a collection of subsets of E, which we will denote $\mathcal{G}$, and such that :

$$\forall U \in \mathcal{G}, \ \ \{X \in U\} \in \mathcal{F}.$$

If $U \in \mathcal{G}$ and $(U_n)_{n \in \mathbb{N}}$ is a family of elements of the collection $\mathcal{G}$ then

$$\{X \in E\} = \Omega \in \mathcal{F},$$

$$\{X \in \overline{U}\} = \overline{\{X \in U\}} \in \mathcal{F},$$

$$\left\{ X \in \bigcup_{n=0}^{+\infty} U_n \right\} = \bigcup_{n=0}^{+\infty} \{X \in U_n\} \in \mathcal{F},$$

So it's natural to require that $\mathcal{G}$ be a σ-algebra on E.

Définition 3.1. Let X be an application from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measured space $(E, \mathcal{G})$. We say that X is a **random variable** taking values in $(E, \mathcal{G})$ if :

$$\forall U \in \mathcal{G}, \ \ \{\omega \in \Omega \mid X(\omega) \in U\} \in \mathcal{F}$$

The following theorem details the case where σ-algebra $\mathcal{G}$ is generated by a set $\mathcal{C}$ of subsets of E, stable by intersections.

[4] The set of events forms a σ-algebra. So the intersection, union and complement of events are still events.

Théorème 3.1 (Dynkin system)**.** *Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{G})$ be an application. Suppose that the σ-algebra $\mathcal{G}$ is generated by a set $\mathcal{C}$ of subsets of E, stable by intersections. This means :*

$$\mathcal{G} = \sigma(\mathcal{C})$$

$$\forall U, V \in \mathcal{C}, \ U \cap V \in \mathcal{C}$$

Then X is a random variable if and only if :

$$\forall U \in \mathcal{C}, \ \{\omega \in \Omega \mid X(\omega) \in U\} \in \mathcal{F}$$

Proof. Admitted. □

In other words, it's enough to check that $\{X \in U\}$ is an event for all $U \in \mathcal{C}$ to deduce that it's an event for all $A \in \mathcal{G}$. This allows us to clarify the definition of a random variable in the two cases we're interested in: real random variables and discrete random variables.

Définition 3.1 (Discrete case)**.** Let X be an application of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a countable space E, provided with the discrete σ-algebra. This means

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{P}(E)).$$

We say that X is a discrete random variable with values in E, if for any $e \in E$, the set $\{X = e\}$ is an event of the σ-algebra $\mathcal{F}$.

Définition 3.2 (Real case)**.** Let X be an application of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}$, provided with the Borelian σ-algebra $\mathcal{B}(\mathbb{R})$, i.e.

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

We say that X is a **real random variable** if for any $x \in \mathbb{R}$, the set $\{X \leq x\}$ is an event of the σ-algebra $\mathcal{F}$.

In practice, we rarely have to show that an application is a random variable, because the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is precisely constructed so that the quantities we're interested in (the number on the dice, the distance to the target, etc.) are random variables. In practical terms, a problem might begin with :

"Let X be a random variable that models the result of a throw of a dice..."

The aim is to get as far away as possible from this probability space, about which little is known. On the other hand, you may be asked to show that other applications dependent on X are indeed random variables.

3.2 Distribution of a random variable

When given a random variable X, it is important to know its distribution, i.e. the probability that X belongs to a certain set of values.

Définition 3.2. The **distribution** of a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{G})$ is the given for any $U : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{G})$. is the given for all $U \in \mathcal{G}$ of the quantities :

$$\mathbb{P}(X \in U) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in U\})$$

Théorème 3.2 (Dynkin system)**.** *If the σ-algebra $\mathcal{G}$ is generated by a set $\mathcal{C}$ subsets of E, stable by intersection, then the distribution of a random variable X is entirely characterized by giving for any $U \in \mathcal{C}$ the quantities :*

$$\mathbb{P}(X \in U)$$

Proof. Admitted. □

There are two fundamental examples to remember.

Définition 3.3 (Discrete case)**.** Let X be a discrete random variable, with values in a countable space E. In this case, the distribution of X is entirely characterized by the quantities :

$$\mathbb{P}(X = e)$$

For $e \in E$.

For example, if the random variable X models the number of people in a queue, then the distribution of X is characterized by giving for $k \in \mathbb{N}$ the quantities:

$$\mathbb{P}(X = k)$$

Définition 3.4. Let X be a real random variable. The distribution of X is entirely characterized by the quantities :

$$\mathbb{P}(X \leq x)$$

For $x \in \mathbb{R}$.

For example, if the random variable X models the distance of a dart from the center of the dartboard, then its distribution is characterized by giving for $x \in \mathbb{R}_+$ the quantities:

$$\mathbb{P}(X \leq x)$$

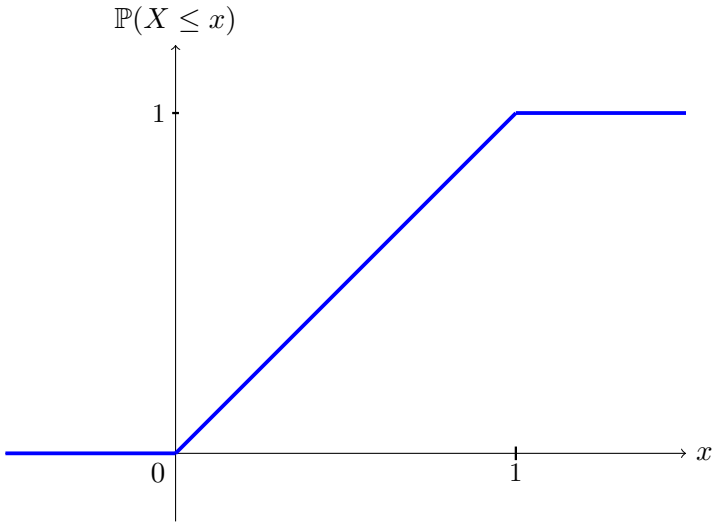
Définition 3.5. Let X be a real random variable. The function F defined by :

$$\begin{aligned} F_X : \mathbb{R} &\longrightarrow [0,1] \\ x &\longrightarrow \mathbb{P}(X \leq x) \end{aligned}$$

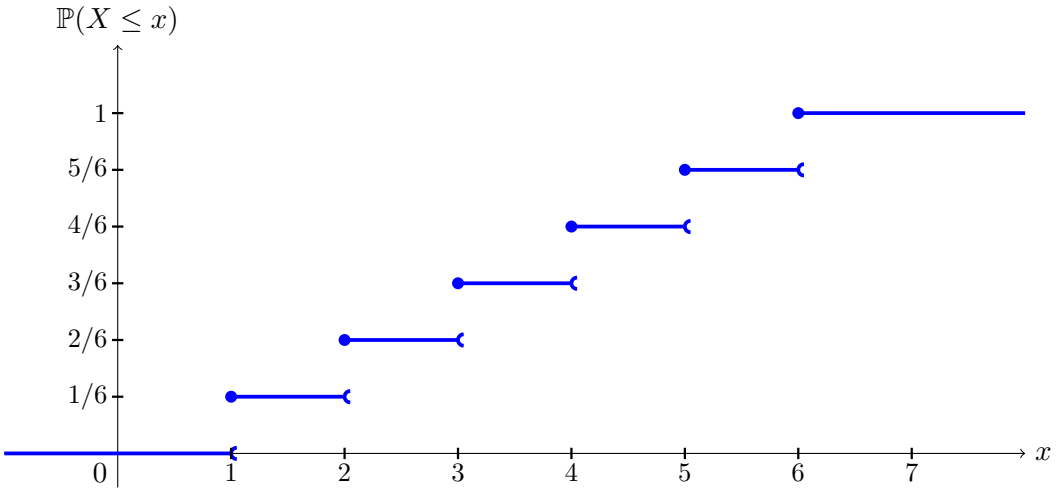
is called the **cumulative distribution function** of the variable X.

By the previous theorem, the cumulative distribution function characterizes the distribution of X. This function is central to the study of real random variables.

Let X be a random variable that models the uniform drawing of a real number between 0 and 1. Its cumulative distribution function is as follows



Let X be a random variable that models a balanced throw of the dice. It is a real random variable since $\{1, 2, 3, 4, 5, 6\} \subset \mathbb{R}$. Its cumulative distribution function is as follows:



In general, when X is a discrete random variable, its cumulative distribution function is piecewise constant, and the points of discontinuity are exactly the values reached by X with non-zero probability.

Théorème 3.1. *The cumulative distribution function F_X of a random variable verifies the following properties:*

- *F_X is increasing and :*
$$\lim_{x \to -\infty} F_X(x) = 0$$
$$\lim_{x \to +\infty} F_X(x) = 1$$
- *F_X is continuous on the right: if $(x_n)_{n \geq 0}$ is a sequence decreasing that converges to a real x then :*
$$\lim_{n \to +\infty} \mathbb{P}(X \leq x_n) = \mathbb{P}(X \leq x)$$
- *F_X admits a left limit at any point: if $(y_n)_{n \geq 0}$ is a increasing sequence that converges to a real y then :*
$$\lim_{n \to +\infty} \mathbb{P}(X \leq y_n) = \mathbb{P}(X < y)$$

An function F verifying the last two points is said to be **càdlàg** (continuous on the right, limit on the left).

Proof. For the first point, let x and y be real numbers such that $x < y$. Then :

$$\{X \leq x\} \subset \{X \leq y\}$$

Then :

$$F_X(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F_X(y)$$

Furthermore, let $(x_n)_{n\in\mathbb{N}}$ be a sequence increasing towards infinity. The events $(\{X \leq x_n\})_{n\in\mathbb{N}}$ form an increasing sequence of events whose union is $\mathbb{R}$ any integer. Hence :

$$\begin{aligned}
\lim_{n\to+\infty} F_X(x_n) &= \lim_{n\to+\infty} \mathbb{P}(X \leq x_n) \\
&= \mathbb{P}\left(\bigcup_{n=0}^{\infty} \{X \leq x_n\}\right) \\
&= \mathbb{P}(X \in \mathbb{R}) \\
&= 1
\end{aligned}$$

The case in $-\infty$ is similar. Let $(x_n)_{n\in\mathbb{N}}$ be a decreasing sequence towards a real x. Then :

$$\begin{aligned}
\lim_{n\to+\infty} F_X(x_n) &= \lim_{n\to+\infty} \mathbb{P}(X \leq x_n) \\
&= \mathbb{P}\left(\bigcap_{n=0}^{\infty} \{X \leq x_n\}\right) \\
&= \mathbb{P}(X \leq x) \\
&= F_X(x)
\end{aligned}$$

Let $(y_n)_{n\in\mathbb{N}}$ be a sequence increasing towards a real y.

$$\begin{aligned}
\lim_{n\to+\infty} F_X(y_n) &= \lim_{n\to+\infty} \mathbb{P}(X \leq y_n) \\
&= \mathbb{P}\left(\bigcup_{n=0}^{\infty} \{X \leq y_n\}\right) \\
&= \mathbb{P}(X < y)
\end{aligned}$$

□

Conversely, let F be a function verifying the three points of Theorem 3.1. We define the quantile function $F^{-1} : [0,1] \to \mathbb{R}$ by

$$F^{-1}(u) = \inf\{x \in \mathbb{R} \mid F(x) \geq u\}.$$

Théorème 3.2. *Let U be a uniform distribution on $[0,1]$. The quantile function F^{-1} verifies the following properties.*

- *If F is continuous and increasing, then F^{-1} is the reciprocal bijection of F.*
- *The random variable $X = F^{-1}(U)$ has the cumulative distribution function F.*

Proof. Accepted. □

In the discrete case, we need only give ourselves a sequence of positive numbers $(\alpha_n)_{n\in\mathbb{N}}$ such that :

$$\sum_{n\geq 0} \alpha_n = 1$$

For $(x_n)_{n\geq 0}$ a sequence of real numbers, there exists a random variable X such that :

$$\forall n \geq 0,\ \mathbb{P}(X = x_n) = \alpha_n.$$

Définition 3.6. Let $x \in \mathbb{R}$. A real random variable X is said to have an **atome** at the point x if :

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

Théorème 3.3. *Let X be a real random variable. The cumulative distribution function of X exhibits a jump of discontinuity at the point x if and only if X has an atom at the point x. In this case, the jump is of size $\mathbb{P}(X = x)$.*

The theorem can be verified on the previous examples.

Proof. The cumulative distribution function of X is continuous at a point x if and only if the limits of F to the left and right of the point x coincide, which is written :

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

Now we have equality:

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x)$$

The right-hand member cancels out if and only if X is atom-free. Otherwise, the jump size is given by the difference between the two limits, which is exactly $\mathbb{P}(X = x)$. □

Définition 3.7 (Random variable with density)**.** Let X be a real random variable with cumulative distribution function F_X. The random variable X is said to have a **density** if there exists an integrable function f_X such that

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

The function f_X is called the probability density of the distribution of X.

The function F_X is a primitive of the function f_X. We deduce that F_X is continuous, so X is atom-free. We also deduce the following theorem:

Théorème 3.4. *Let X be a real random variable. If its cumulative distribution function F_X is continuous and piecewise derivable, then X is a random variable with probability density f and :*

$$f_X = (F_X)'$$

Proof. This is the link between primitive and derivative. □

Conversely, if X is a random variable that has a density, its characteristic function is continuous, so X has no atom. On the other hand, there are atom-free random variables that do not admit a density (Cantor's staircase is a counterexample).

Let a and b be real numbers with $a < b$, and X be a random variable with density f_X. Then :

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t)\mathrm{d}t$$

For example, let X be a uniform random variable on the set $[0, 1]$. For $0 \leq a \leq b \leq 1$ we have :

$$\mathbb{P}(a \leq X \leq b) = b - a = \int_a^b 1\mathrm{d}t$$

In other words, the density of the uniform distribution is the function $\mathbb{1}_{[0,1]}$.

If now $I = [x, x + \varepsilon]$ is a small interval, then we have :

$$\mathbb{P}(x \leq X \leq X + \varepsilon) = \int_x^{x+\varepsilon} f_X(t)\mathrm{d}t \simeq \varepsilon f_X(x).$$

Let X be a random variable with density f_X. The probability that X belongs to a small interval around a point $x \in \mathbb{R}$ is proportional to the size of this interval. The proportionality coefficient is exactly $f_X(x)$.

Théorème 3.5. *Let X be a random variable with density f_X. Then :*

- *f_X is positive*
- *f_X has mass 1 :*

$$\int_{-\infty}^{+\infty} f_X(t)\mathrm{d}t = 1$$

Conversely, if f is a function verifying these two properties, then there exists a random variable X of density f.

Proof. For the first point, we have for ε a small positive real :

$$0 \leq \mathbb{P}(x \leq X \leq X + \varepsilon) \simeq \varepsilon f(x)$$

And so f is necessarily positive. For the second point, we have :

$$1 = \mathbb{P}(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(t)\mathrm{d}t$$

□

3.3 Averages and dispersion of a real random variable

In this section, we consider only real random variables. We will define the notions of *mean* and *dispersion* of a real random variable. These tools give us a better understanding of the random variable we're dealing with.

3.3.1 Expectation

How can we make sense of the notion of *mean* of a random variable? An average is intuitively a real value that is supposed to reflect the entire random variable. It is therefore a way to reduce all the information generated by the random variable to a single number.

Let's start by considering a positive real random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a finite or countable set, and $\mathcal{F}$ is the discrete σ-algebra. For example $\Omega = \{a, b, c\}$, with :

$$\mathbb{P}(\{a\}) = \mathbb{P}(\{b\}) = \mathbb{P}(\{c\}) = \frac{1}{3}$$

A natural way to define its mean $\mathbb{E}[X]$ is via the formula :

$$\mathbb{E}[X] = \frac{X(a) + X(b) + X(c)}{3}$$

In general, let $\Omega = \{\omega_1, \omega_2, \ldots\}$, given the discrete σ-algebra and any probability measure $\mathbb{P}$. If the sum :

$$\mathbb{E}[X] = \sum_{n=1}^{+\infty} X(\omega_n)\mathbb{P}(\{\omega_n\})$$

converges, we say that X has finite expectation and we denote $\mathbb{E}[X]$ its mean.

Now suppose we're considering a real random variable defined on the probability space $([0,1], \mathcal{B}([0,1]), \mu)$. A random variable X is then an application of $[0,1]$ in $\mathbb{R}$ (which verifies the condition of the theorem 3.1). If X is an integrable function, then its mean is defined by :

$$\mathbb{E}[X] = \int_0^1 X(\omega)\mathrm{d}\omega.$$

In these two examples, we have a tool for averaging:

- The sum in the case where Ω is finite or countable
- The integral where $\Omega = [0,1]$.

We're going to define the notion of **integral** (or **expectation**) on any probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let's start by defining the integral of "easy" functions. Let A be an event. It is natural to pose :

$$\mathbb{E}\left[\mathbb{1}_A\right] = \mathbb{P}(A)$$

We ask that the expectation be *linear*. If $A_1, \ldots, A_n$ are events and $\lambda_1, \ldots, \lambda_n$ are real numbers, then :

$$\mathbb{E}\left[\lambda_1 \mathbb{1}_{A_1} + \ldots + \lambda_n \mathbb{1}_{A_n}\right] = \lambda_1 \mathbb{P}(A_1) + \ldots + \lambda_n \mathbb{P}(A_n)$$

It's easy to see that this definition is consistent. Suppose there are events $B_1, \ldots, B_m$ and real $\alpha_1, \ldots \alpha_m$ such that :

$$\lambda_1 \mathbb{1}_{A_1} + \ldots + \lambda_n \mathbb{1}_{A_n} = \alpha_1 \mathbb{1}_{B_1} + \ldots + \alpha_m \mathbb{1}_{B_m}$$

In this case:

$$\lambda_1 \mathbb{P}(A_1) + \ldots + \lambda_n \mathbb{P}(A_n) = \alpha_1 \mathbb{P}(B_1) + \ldots + \alpha_m \mathbb{P}(B_m)$$

For example:

$$\mathbb{1}_{[0,1/2]} + 2\mathbb{1}_{[1/2,1]} = \mathbb{1}_{[0,1]} + \mathbb{1}_{[1/2,1]}$$

And :

$$\frac{1}{2} + 2 * \frac{1}{2} = 1 + \frac{1}{2}$$

A random variable X written in the form :

$$X = \lambda_1 \mathbb{1}_{A_1} + \ldots + \lambda_n \mathbb{1}_{A_n}$$

is called **simple**. These are the building blocks for the definition of our integral (just as staircase functions are the building blocks for the Riemann integral).

Définition 3.3. Let X be a **positive** real random variable on a probability space. If X is a simple random variable with :

$$X = \lambda_1 \mathbb{1}_{A_1} + \ldots + \lambda_n \mathbb{1}_{A_n}$$

We pose :

$$\mathbb{E}[X] = \lambda_1 \mathbb{P}(A_1) + \ldots + \lambda_n \mathbb{P}(A_n)$$

In the general case, we pose :

$$\mathbb{E}[X] = \sup \{\mathbb{E}[Y] \mid Y \text{ simple and } Y \leq X\}$$

We say that X is integrable (or of finite expectation) if :

$$\mathbb{E}[X] < +\infty$$

Définition 3.4. Let X be a real random variable **quelconque** (not necessarily positive). We decompose X into :

$$X = X \mathbb{1}_{X \geq 0} + X \mathbb{1}_{X < 0}$$

If the (positive) random variables $X \mathbb{1}_{X \geq 0}$ and $(-X \mathbb{1}_{X < 0})$ have finite expectations, we say that X has a finite expectation and define its mean by :

$$\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}_{X \geq 0}] - \mathbb{E}[X \mathbb{1}_{X < 0}]$$

In measure theory, this integral is called the Lebesgue integral. In the discrete case, it coincides with the sum as seen in the introduction. In the real case, it coincides with the Riemann integral for piecewise continuous functions, but allows even more functions to be integrated. We can, for example, integrate highly irregular functions like $\mathbb{1}_{\mathbb{Q}}$. Dominated convergence theorems are also more powerful and simpler to state.

The expectation of a positive random variable may be infinite. For example, in the case where $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \mu)$ the function :

$$X : \omega \mapsto \frac{1}{\omega}$$

has an infinite integral.

Figure 1: Left: approximation of the Riemann integral by staircase functions. Right: Lebesgue integral approximated by simple functions.

Théorème 3.6. *Let X and Y be two integrable random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and λ a real number. Then,*

- *Linearity: For $\lambda \in \mathbb{R}$, we have*

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

- *Monotonicity: If, for almost all $\omega \in \Omega$, we have $X(\omega) \leq Y(\omega)$, then*

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

- *If X follows a Bernoulli distribution with parameter p, then*

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

Proof. We have seen that the expectation is linear for simple random variables. We will assume that this property remains true when passing to the limit. The monotonicity of expectation can be proved in a similar way. □

Unfortunately, this definition of the average is not practical at all. It seems to depend heavily on the probability space Ω. Fortunately, we have the following fundamental theorem that allows us to overcome the problem.

Théorème 3.7 (Transfer Theorem)**.** *Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and g be a function from $\mathbb{R}$ to $\mathbb{R}$. Then, the quantity $\mathbb{E}[g(X)]$ depends only on the function g and the distribution of X.*

Proof. We use an approximation argument. If U is a measurable subset of $\mathbb{R}$, and $g = \mathbb{1}_U$, then:

$$\mathbb{E}[g(X)] = \mathbb{E}[\mathbb{1}_{\{X \in U\}}] = \mathbb{P}(X \in U).$$

This quantity depends only on the distribution of X. It's still true when g is a sum of indicator functions, due to linearity of expectation. We use an approximation argument by simple functions for arbitrary functions g. □

We will specify this theorem in the case where X is a discrete or continuous random variable.

Théorème 3.8. *Let X be a discrete random variable, and let $(x_n)_{n\in\mathbb{N}}$ be the values it can take. Then, X has finite expectation if and only if:*

$$\sum_{n=0}^{\infty} |x_n|\mathbb{P}(X = x_n) < +\infty$$

In this case:

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} x_n \mathbb{P}(X = x_n)$$

Moreover, if g is a function from $\mathbb{R}$ to $\mathbb{R}$ (provided the sum converges absolutely):

$$\mathbb{E}[g(X)] = \sum_{n=0}^{\infty} g(x_n)\mathbb{P}(X = x_n)$$

Proof. You only need to adapt the beginning of the proof of Theorem 3.7 to the case where the random variable X is not necessarily positive. □

Théorème 3.9. *Let X be a real random variable with density f_X. Then, X has finite expectation if and only if:*

$$\int_{\mathbb{R}} |x| f_X(x)\mathrm{d}x < +\infty$$

In this case:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\mathrm{d}x$$

Furthermore, if g is a function from $\mathbb{R}$ to $\mathbb{R}$ (provided the integral converges absolutely):

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)\mathrm{d}x.$$

Proof. If $g = \mathbb{1}_{[a,b]}$:

$$\mathbb{E}[g(X)] = \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x = \int_{\mathbb{R}} g(x)f(x)\mathrm{d}x$$

By linearity, the formula is still true if g is a sum of indicator functions. Using an approximation result, we deduce the general case. □

An important theorem for bounding certain probabilities is the following.

Théorème 3.10 (Markov's Inequality)**.** *Let X be a positive random variable with finite expectation, and α a strictly positive real number. Then*

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

More generally, if g is a positive and strictly increasing function, then

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

Proof. Let $A = \{X \geq \alpha\}$. We observe that, for all $\omega \in \Omega$, $\alpha \mathbb{1}_A(\omega) \leq X(\omega)$. Taking the expectation, we obtain:

$$\alpha \mathbb{E}[\mathbb{1}_A] \leq \mathbb{E}[X]$$

$$\mathbb{P}(A) \leq \frac{\mathbb{E}[X]}{\alpha}$$

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

For the second inequality, we observe that $A = \{X \geq \alpha\} = \{g(X) \geq g(\alpha)\}$. We can apply the inequality to the random variable $g(X)$ and the real number $g(\alpha)$. □

3.3.2 Median

An INSEE survey revealed that the average salary of a French person is 2,263 euros net per month. However, half of the population earns a monthly net salary less than 1,683 euros. How can we explain such a difference (580 euros)?

The average of a sample is sensitive to extreme values in the sample. That's why it's not always suitable for studying a statistic.

In our case, a small portion of people with very high income inflates the average salaries. Let's imagine a village with 1,000 inhabitants, where the average salary is 2,000 euros per month. If Bill Gates (monthly salary: 500,000 euros) decided to move to this village, the average salary would increase by 500 euros, a 25% increase. That's why it's more reasonable here to use an estimator that is robust against extreme values.

Définition 3.8. The median of a random variable is a number m such that:

$$P(X \leq m) \geq 1/2 \quad \text{and} \quad P(X \geq m) \geq 1/2$$

It can be easily shown that there is always at least one median, but it may not be unique. For example, if X follows a Bernoulli distribution, any number between 0 and 1 is a median.

In the previous example, the median salary is 1,683 euros net per month. The advantage of a median is that it is insensitive to extreme values. If salaries were capped at 3,000 euros net per month, the median salary would remain unchanged. A significant difference between the average salary and the median salary shows that the majority of the population lives poorly, while a small portion of the population accumulates wealth.

In the case of nuclear decay, the median is also called half-life. The probability that an atom decays before its half-life is exactly $1/2$.

3.3.3 Variance

Variance is a measure of the *spread* of a random variable.

The spread quantifies how much a random variable (or a sample) deviates from its average value. A random variable with zero spread is constant. In finance, spread, or volatility, is related to risk. A high variance indicates that a stock's value tends to fluctuate a lot, while a low variance indicates relative stability.

For another example, consider looking at the scores of an exam you just graded. If the variance is low, it means your students are all or mostly answering the easy questions and struggling with

the difficult ones. It's harder to rank the students in this case. Conversely, a high variance suggests that the exam was well-balanced, making it easier to rank the candidates. This is often desired in competitive exams.

Définition 3.9. Let X be a random variable with finite expectation. We denote $\mathrm{Var}(X)$ as the **variance** of X, defined as:

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

If this quantity is finite, we say X has finite variance. In that case, we denote by

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

its standard deviation.

Variance measures the deviation from the mean. It's one of the most common ways to measure the spread of a random variable.

Théorème 3.11. *Let X be a random variable with finite variance, and λ a real number. We have:*

- $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X + \lambda) = \mathrm{Var}(X)$
- *X is a.s. constant if and only if* $\mathrm{Var}(X) = 0$
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Proof. For the first point:

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

For the second point:

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

For the third point, if X is constant, then X is equal to c. So:

$$\mathrm{Var}(X) = \mathbb{E}[(c - E[c])^2] = \mathbb{E}[(c - c)^2] = E[0] = 0.$$

The converse follows from the fact that a positive random variable with zero mean is equal to zero almost surely. Hence $X = \mathbb{E}[X]$ and is thus constant a.s.. For the last point, we expand the expression of the variance:

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

□

The following formulas for variance are also important in discrete and continuous cases.

Théorème 3.12. *If X is a discrete random variable with values in $\{x_1, x_2, \ldots\}$ and finite variance, then:*

$$\mathrm{Var}(X) = \left(\sum_{n=1}^{+\infty} x_n^2 P(X = x_n)\right) - \left(\sum_{n=1}^{+\infty} x_n P(X = x_n)\right)^2$$

Théorème 3.13. *If X is a random variable with density f_X and finite variance, then:*

$$\mathrm{Var}(X) = \left(\int_{\mathbb{R}} t^2 f_X(t)\,dt\right) - \left(\int_{\mathbb{R}} t f_X(t)\,dt\right)^2$$

We also have the following theorem, which is important for the rest of the course.

Théorème 3.14 (Chebyshev's Inequality)**.** *Let X be a random variable with finite variance and β a nonzero real number. By using Markov's inequality, we can show that:*

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

3.4 Common probability distributions

Let's review some common probability distributions frequently encountered in probability theory.

3.4.1 Discrete distributions

• Bernoulli distribution $\mathcal{B}(p)$: This distribution is associated with an experiment having two possible outcomes: 0 or 1. A random variable X follows a Bernoulli distribution with parameter p if:

$$\mathbb{P}(X = 0) = 1 - p \qquad \text{and} \qquad \mathbb{P}(X = 1) = p$$

If A is an event, the indicator random variable $\mathbb{1}_A$ follows a Bernoulli distribution with parameter $\mathbb{P}(A)$.

$$\mathbb{E}[X] = p \qquad \mathrm{Var}(X) = p(1-p)$$

• Binomial distribution $\mathcal{B}(n,p)$: This distribution is associated with the repetition of n independent and identically distributed random variables following a Bernoulli distribution with parameter p. A random variable X follows a binomial distribution with parameters n and p if:

$$\text{For } k \in \{0,\dots,n\},\ \ \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

If you toss a biased coin n times (with a probability of landing on heads being p), the random variable X counting the number of heads follows a binomial distribution with parameters n and p.

$$\mathbb{E}[X] = np \qquad \mathrm{Var}(X) = np(1-p)$$

• Uniform distribution $\mathbb{U}(a,b)$: This is the uniform distribution over the integers $a, a+1, \dots, b$. We denote $n = b - a + 1$. A random variable X follows a uniform distribution with parameters a and b if:

$$\text{For } k \in \{0,\dots,n\},\ \ \mathbb{P}(X = k) = \frac{1}{n}$$

We have:

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathrm{Var}(X) = \frac{n^2 - 1}{12}$$

• Geometric distribution $\mathcal{G}(p)$: This distribution is associated with the distribution of the first occurrence of a repeated experiment. A random variable X follows a geometric distribution with parameter p if:

$$\forall k \in \mathbb{N}^*, \ \mathbb{P}(X = k) = (1-p)^{k-1}p^k$$

We have:

$$\mathbb{E}[X] = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

If you toss a biased coin (with a probability of landing on heads being p), the random variable X giving the first occurrence of heads follows a geometric distribution with parameter p.

• Poisson distribution $\mathcal{P}(\lambda)$: This distribution is associated with the number of people in a queue. A random variable X follows a Poisson distribution with parameter λ if:

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

In a queue, we assume that the time between the arrival of two customers follows a memoryless (exponential) distribution with parameter λ. In this case, the number of people in the queue at time 1 follows a Poisson distribution with parameter λ (at time t, the parameter is λt).

$$\mathbb{E}[X] = \lambda \qquad \mathrm{Var}(X) = \lambda$$

3.4.2 Continuous distributions

• Uniform distribution $\mathcal{U}([a,b])$: This is the uniform distribution over a bounded interval $[a,b]$. X follows a uniform distribution over $[a,b]$ if:

$$\forall c,d \text{ such that } a \leq c \leq d \leq b, \quad \mathbb{P}(c < X < d) = \frac{d-c}{b-a}.$$

The density is given by

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

You can consider random variables that are uniform over any set with finite measure (bounded interval, circle, sphere, torus, etc).

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

• Exponential distribution $\mathcal{E}(\lambda)$: This is the only family of memoryless distributions. Its density is given by:

$$f_X(t) = \lambda e^{-\lambda t}.$$

We have:

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

• Normal distribution $\mathcal{N}(m,\sigma^2)$: This distribution naturally appears in the central limit theorem. It often models random fluctuations of a parameter around its average value (temperature, white noise, stock prices, etc). Its density is given by:

$$f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-m)^2}{2\sigma^2}}.$$

3.5 Pair of random variables

We are often required to consider multiple random variables simultaneously. For instance, we can calculate their sum, examine if they are correlated, etc. We provide the definition of a pair of random variables, but this definition can be extended to a countable set of random variables. Let X be a random variable:

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{G})$$

and let Y be another random variable defined on the same sample space:

$$Y : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (F, \mathcal{H})$$

Then it is possible to define the product random variable:

$$(X, Y) : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E \times F, \mathcal{G} \otimes \mathcal{H})$$

Définition 3.5. Let $(E, \mathcal{G})$ and $(F, \mathcal{H})$ be two measurable spaces. The product space:

$$E \times F = \{(x, y) \,|\, x \in E, y \in F\}$$

can naturally be equipped with a sigma-algebra, called the product sigma-algebra and denoted by $\mathcal{G} \otimes \mathcal{H}$. It is generated by the sets:

$$\mathcal{G} \otimes \mathcal{H} = \sigma\{(A \times B) \,|\, A \in \mathcal{G}, B \in \mathcal{H}\}$$

The law of a pair (X, Y) of random variables defined as above is the specification of probabilities:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(\{X \in A\} \cap \{Y \in B\})$$

for all $A \in \mathcal{G}$ and $B \in \mathcal{H}$.

When we know the law of the pair (X, Y), it is possible to determine the laws of X and Y. This is referred to as the **marginal** laws of the pair (X, Y).

Théorème 3.3 (Marginal Distribution)**.** *Let X be a random variable:*

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{G})$$

and let Y be another random variable defined on the same sample space:

$$Y : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (F, \mathcal{H})$$

If we know the law of the pair (X, Y), we can derive the law of X using the formula:

$$\forall A \in \mathcal{G}, \;\; \mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in F)$$

Proof. Trivial. □

The lemma of monotonous classes remains true. Let's detail the two cases of interest.

Théorème 3.15 (Discrete Case). *Let X and Y be two discrete random variables defined on the same probability space. The law of the pair (X, Y) is given by the probabilities:*

$$\mathbb{P}(X = x_k, Y = y_j)$$

where x_k (respectively, y_j) ranges over the values taken by X (respectively, Y).

The marginal laws of X and Y are then given by:

$$\mathbb{P}(X = x_k) = \sum_j \mathbb{P}(X = x_k, Y = y_j) \quad \text{and} \quad \mathbb{P}(Y = y_j) = \sum_k \mathbb{P}(X = x_k, Y = y_j).$$

Théorème 3.16 (Continuous Case). *Let X and Y be two real random variables defined on the same probability space. The law of the pair (X, Y) is given by the probabilities:*

$$\mathbb{P}(X \leq x, Y \leq y)$$

where x and y range over the real numbers.

We can specify the particular case where the pair (X, Y) has a joint density.

Théorème 3.17. *Let X and Y be two real random variables defined on the same probability space. We say that the pair (X, Y) has a joint density $f_{(X,Y)}$ if, for all intervals I and J in $\mathbb{R}$, we have:*

$$\mathbb{P}(X \in I, Y \in J) = \int_{I \times J} f_{(X,Y)}(x, y)\, dx\, dy$$

In this case, the random variables X and Y have marginal densities given by:

$$f_X(x) \mapsto \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy \quad \text{and} \quad f_Y(y) \mapsto \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dx.$$

3.6 Independent random variables

Let X be a random variable:

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{G})$$

and let Y be another random variable defined on the same sample space:

$$Y : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (F, \mathcal{H})$$

We define the independence of two random variables, X and Y.

Définition 3.6. Two random variables X and Y defined as above are independent if:

$$\forall A \in \mathcal{G}, \forall B \in \mathcal{H}, \ \ \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

In other words, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.

Let's specify the discrete and real cases.

Définition 3.10 (Discrete Case)**.** Let X and Y be two discrete random variables defined on the same probability space. The random variables X and Y are independent if and only if:

$$\mathbb{P}(X = x_k, Y = y_j) = \mathbb{P}(X = x_k)\mathbb{P}(Y = y_j)$$

where x_k (respectively, y_j) ranges over the values taken by X (respectively, Y).

Définition 3.11 (Continuous Case)**.** Let X and Y be two real random variables defined on the same probability space. The random variables X and Y are independent if and only if:

$$\forall x, y \in \mathbb{R},\ \ \mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y).$$

If X and Y are two random variables with densities $f_X(x)$ and $f_Y(y)$, they are independent if and only if the pair (X, Y) has a joint density given by:

$$f_{(X,Y)}(x, y) \mapsto f_X(x) f_Y(y).$$

Thus, if two random variables X and Y are independent, then we know the law of the pair (X, Y): it is simply the law given as the "product law" of X and Y. Independence can be understood in terms of "information." If X is a random variable with values in $(E, \mathcal{G})$, we can define the sigma-algebra $\sigma(X)$:

$$\sigma(X) = \{\{X \in A\} \mid A \in \mathcal{G}\}$$

This is the set of events generated by the random variable X. In this case, the random variables X and Y are independent if and only if the sigma-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.

In general, it is possible to talk about a family $(X_n)_{n\in\mathbb{N}}$ of mutually independent random variables. They are independent if and only if the sigma-algebras $(\sigma(X_n))_{n\in\mathbb{N}}$ are mutually independent. For example, if X, Y, and Z are the results of rolling three dice, then $X + Y$ is independent of Z.

Théorème 3.18. *Let X and Y be two independent random variables with finite expectations. Then:*

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

And more generally, for any measurable functions f and g from $\mathbb{R}$ to $\mathbb{R}$:

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

If X and Y have finite variances:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Proof. For the first point, we can easily prove it for indicator functions: if $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$ with A and B as intervals, then:

$$\begin{aligned} E[f(X)g(Y)] &= E[\mathbb{1}_{\{X\in A\}}\mathbb{1}_{\{Y\in B\}}] \\ &= \mathbb{P}(X \in A, Y \in B) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \\ &= \mathbb{E}[f(X)]\mathbb{E}[g(Y)] \end{aligned}$$

By linearity and passing to the limit, we can derive the general result. For the second point:

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

We denote:

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

the covariance of X and Y. In the case where X and Y are independent, the covariance of the two terms is zero, according to the first part of the theorem, and we can deduce the proposition. □

Théorème 3.19. *Let $X_1, \ldots, X_n$ be mutually independent random variables. Then:*

$$\mathbb{E}[X_1 \ldots X_n] = \mathbb{E}[X_1] \ldots \mathbb{E}[X_n]$$

And:

$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

Proof. Similar to the proof of the previous theorem. □

An important application of independence is the sum of independent real random variables.

Théorème 3.20. *Let X and Y be two independent random variables with values in $\mathbb{N}$, and $Z = X + Y$. Then:*

$$\mathbb{P}(Z=n) = \sum_{k=0}^{n} \mathbb{P}(X=k)\mathbb{P}(Y=n-k)$$

Let X and Y be two real independent random variables with densities. Then, $Z = X+Y$ has a density given by:

$$f_Z(z) = \int_{\mathbb{R}} f_X(u) f_Y(z-u)\, du$$

Proof. In the discrete case, it's the formula of total probabilities. In the case with densities, we have:

$$\mathbb{P}(X+Y \leq z) = \mathbb{P}((X,Y) \in A_z)$$

with A_z as the set:

$$A_z = \{(x,y) \in \mathbb{R}^2 \,|\, x+y \leq z\}$$

Thus:

$$\begin{aligned}\mathbb{P}(X+Y \leq z) &= \int_{x+y\leq z} f_X(x) f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{z} \int_{\mathbb{R}} f_X(u) f_Y(v-u)\, du\, dv\end{aligned}$$

We performed the change of variables $(u,v) = (x, x+y)$. The density is obtained by differentiating with respect to the variable z. □

4 Limit theorems in probability

4.1 Modes of convergence of random variables

Let $(X_n)_{n\geq 0}$ be a sequence of random variables and X a random variable, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We aim to give meaning to the convergence of the sequence $(X_n)_{n\geq 0}$ to X. To achieve this, we distinguish between 4 modes of convergence.

Définition 4.1 (Almost Sure Convergence)**.** We say that the sequence of random variables $(X_n)_{n\geq 0}$ converges almost surely to X if, for almost every $\omega \in \Omega$,

$$\lim_{n\to+\infty} X_n(\omega) = X(\omega).$$

We write

$$X_n \xrightarrow[n\to+\infty]{a.s} X.$$

This is the pointwise convergence of random variables. It appears in the law of large numbers. Intuitively, if we simulate the sequence $X_n(\omega) - X(\omega)$ for a fixed ω, then this sequence converges to 0. It is also found in the strong law of large numbers, and its main application is in the Monte Carlo method. A simple case of almost sure convergence occurs when dealing with an increasing and bounded sequence of random variables. In that case, it converges almost surely to a limiting random variable.

Définition 4.2 (Convergence in L^p)**.** We say that $(X_n)_{n\geq 0}$ converges in L^p if

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

We write

$$X_n \xrightarrow[n\to+\infty]{L^p} X.$$

When $p = 2$, it is called quadratic convergence. This mode of convergence is useful for certain theoretical applications, particularly for proving the convergence of the mean, variance, etc.

Définition 4.3 (Convergence in Probability)**.** We say that the sequence of random variables $(X_n)_{n\geq 0}$ converges in probability to X if, for every $\varepsilon > 0$,

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

We write

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X.$$

Intuitively, a sequence of random variables converges in probability to its limit when the probability that the sequence takes values far from its limit tends to 0. For example, if $(X_n)_{n\geq 0}$ is a sequence of random variables following a Bernoulli distribution with parameter $1/n$, then it converges in probability to the constant random variable equal to 0, because

$$\forall \varepsilon > 0, \qquad \mathbb{P}(|X_n - 0| > \varepsilon) \leq \mathbb{P}(X_n = 1) = \frac{1}{n} \quad \underset{n\to+\infty}{\longrightarrow} 0.$$

Définition 4.4 (Convergence in Distribution)**.** We say that the sequence of random variables $(X_n)_{n\geq 0}$ (with distribution functions $(F_n)_{n\geq 0}$) converges in distribution to X (with distribution function F) if, at every continuity point x of the function F, we have

$$\lim_{n\to+\infty} F_n(x) = F(x).$$

We write

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X.$$

Convergence in distribution is unique among the 4 modes of convergence. It depends only on the distribution of the sequence and not on the probability space on which it is defined. For example, if X is a symmetric distribution, the distribution of the sequence

$$(X, -X, X, -X, \ldots)$$

is constant and therefore converges in distribution to X, but generally does not converge for other modes of convergence. When simulating an instance of the sequence, one typically does not observe particular convergence. However, it becomes apparent when repeating the experiment a large number of times and, for example, creating a histogram. It is a weak mode of convergence but is easy to prove because it depends only on the sequence of distributions of the sequence $(X_n)_{n\geq 0}$, not on their interdependence. Instead of X, one can directly put a distribution. For example, we can write

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

The following implications hold for convergences:

$$\text{Almost Sure Convergence} \Rightarrow \text{Convergence in Probability} \Rightarrow \text{Convergence in Distribution}$$
$$\text{Convergence in } L^p \Rightarrow \text{Convergence in Probability} \Rightarrow \text{Convergence in Distribution}$$

The converses are generally false, except for some special cases.

4.2 Law of large numbers

The law of large numbers establishes the connection between the empirical mean and the theoretical mean. For example, in a dice roll, I have a one in six chance of rolling a 5. The law of large numbers asserts that, on average, after a large number of experiments, I would have rolled a 5 approximately one in six times.

The law of large numbers allows validating (or invalidating) a probabilistic model. For instance, to check whether a coin is biased or not, one can simply toss it a large number of times. If the average number of heads is significantly different from 0.5, it can be concluded that the coin is biased.

In the following, $(X_n)_{n\geq 1}$ is a sequence of real random variables, independent and identically distributed, meaning they all have the same distribution. Let X be a random variable following this distribution. Sometimes, it is abbreviated by saying that $(X_n)_{n\geq 1}$ is an i.i.d sequence with the same distribution as X. We define the sequence of *empirical means* $(\overline{X}_n)_{n\geq 1}$ as

$$\forall n \geq 0, \qquad \overline{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

Théorème 4.1 (Law of Large Numbers)**.** *If the expectation of X is finite, then*

$$\overline{X}_n \xrightarrow[n\to+\infty]{a.s} \mathbb{E}[X].$$

Proof. Admitted. □

For example, suppose I toss a biased coin infinitely many times, where the probability of getting "Heads" is p. We can define A_n as the event "The coin lands on "Heads" on the n-th toss," and

$$X_n = \mathbf{1}_{A_n}.$$

The random variable X_n is equal to 1 if the coin lands on heads and 0 otherwise. Here, the sequence $(X_n)_{n\geq 1}$ is indeed an i.i.d sequence following a Bernoulli distribution with parameter p. In this case,

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} = \frac{\text{number of "Heads" in the first } n \text{ tosses}}{n}.$$

It corresponds to the proportion of "Heads" in the first n tosses. The law of large numbers then asserts that this proportion converges to the expectation of X_n, which is the probability of the event "The coin lands on "Heads"":

$$X_n \xrightarrow[n\to+\infty]{a.s} p.$$

An application of the law of large numbers is the Monte Carlo method for estimating the value of an integral. This method is particularly useful in dimensions $d \geq 1$.

Théorème 4.2 (Monte Carlo). *Let $f : [0,1]^d \to \mathbb{R}$ be an integrable function, and $(U_n)_{n\geq 1}$ be an i.i.d sequence of uniform random variables on $[0,1]^d$. Then*

$$\frac{f(U_1) + \ldots + f(U_n)}{n} \xrightarrow[n\to+\infty]{} \int_{[0,1]^d} f(x)\mathrm{d}x.$$

Proof. We define $X_n = f(U_n)$. The sequence $(X_n)_{n\geq 1}$ is a sequence of real random variables i.i.d with expectation

$$\mathbb{E}[f(U)] = \int_{[0,1]^d} f(x)\mathrm{d}x,$$

according to the transfer theorem. The conclusion follows from the law of large numbers applied to the sequence $(X_n)_{n\geq 1}$. □

4.3 Central limit theorem

The Central Limit Theorem (CLT) allows us to specify the fluctuations of the empirical mean around the theoretical mean. It helps construct confidence intervals when estimating a certain parameter. For example, if I toss a coin 10000 times and get heads 4500 times, can I conclude that the coin is fair? This is the purpose of the CLT. As before, we define $(X_n)_{n\geq 1}$ as a sequence of i.i.d random variables with the same distribution as X, and

$$\forall n \geq 0, \qquad \overline{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

As a reminder, a real random variable Z follows a normal distribution $\mathcal{N}(m, \sigma^2)$ if it has a density given by

$$f_Z(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Théorème 4.3. *Assuming that the variance of X is finite, then*

$$\frac{\overline{X}_n - \mathbb{E}[\overline{X}_n]}{\sqrt{\mathrm{Var}(\overline{X}_n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Proof. Admitted. □

Observation of the CLT is made using a histogram. If we denote $\sigma^2 = \mathrm{Var}(X)$, then

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}[X] \qquad \text{and} \qquad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

If Z follows a standard normal distribution (i.e., $\mathcal{N}(0,1)$), then

$$\sigma Z + m \sim \mathcal{N}(m, \sigma^2).$$

In this case, the CLT can be written in different ways.

Equivalently, the CLT can be expressed as

$$\sqrt{n}\left(\overline{X_n} - \mathbb{E}[X]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

$$\frac{(X_1 + \ldots + X_n) - n\mathbb{E}[X]}{\sqrt{n}\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Or informally,

$$\overline{X_n} \simeq \mathcal{N}\left(\mathbb{E}[X], \frac{\sigma^2}{n}\right).$$

The CLT asserts that the error between the theoretical and empirical mean is of the order of $\sqrt{\mathrm{Var}(\overline{X}_n)} = \sigma/\sqrt{n}$. This allows the calculation of asymptotic confidence intervals. Let's see an example to estimate the mean of a random variable using the observation of a sample of size n, in the case where the variance σ is known.

Théorème 4.4 (Asymptotic Confidence Interval at level α)**.** *Let Z be a standard normal random variable, and $0 < \alpha < 1$. We define the number $q_{\alpha/2}$ such that*

$$\mathbb{P}(-q_{\alpha/2} \leq Z \leq q_{\alpha/2}) = 1 - \alpha.$$

Then

$$\lim_{n\to+\infty} \mathbb{P}\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q_{\alpha/2} \;\leq\; \mathbb{E}[X] \;\leq\; \overline{X}_n + \frac{\sigma}{\sqrt{n}} q_{\alpha/2}\right) = 1 - \alpha.$$

Proof. According to the CLT, we have

$$\lim_{n\to+\infty} \mathbb{P}\left(-q_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \mathbb{E}[X]}{\sigma} \leq q_{\alpha/2}\right) = 1 - \alpha.$$

The conclusion follows. □

We have $\mathbb{P}(-2 \leq Z \leq 2) \simeq 0.95$, and when the sequence $(X_n)_{n\geq 0}$ is an i.i.d sequence of Bernoulli random variables, we have the equality

$$\sigma \leq 1/2.$$

In this case, the commonly used confidence interval is:

$$\lim_{n\to+\infty} \mathbb{P}\left(\overline{X}_n - \frac{1}{\sqrt{n}} \;\leq\; \mathbb{E}[X] \;\leq\; \overline{X}_n + \frac{1}{\sqrt{n}}\right) \gtrsim 0.95.$$

This is the "classic" confidence interval, which is sometimes summarized by the saying "The error made in estimating the proportion of a population is of order of the inverse square root of the sample size."