Proba Stats

Dario Marcone

Lundi 15 septembre 2025

Table des matières

Chapitre 1

Probabilities : starter guide

1.1 Probabilities spaces and measures

1.1.1 Probability spaces

Définition 1

A **probability space** is a set of realisations denoted Ω, together with a probability measure on Ω. A **probability measure** on Ω is a function $\mathcal{P} : \mathcal{P}(\Omega) \to [0,1]$ such that

1. $P(\emptyset) = 0, P(\Omega) = 1$.
2. If $A_i \in \mathcal{F}, i \in \mathbb{N}$ is a sequence of events with $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$P\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i).$$

Remarque

(1) The probability that something happens is 1 and that nothing happens is 0.
(2) The probability of events that cannor occur simultaneously is the suk of the probabilities of the events.
From these properties, we can deduce the next theorems.

Théorème 1

Let P be a a probability measure on some realisation set Ω. Then,

1. $P(\emptyset) = 0, P(\Omega) = 1$;
2. *for any event A, $P(\Omega \backslash A) = 1 - P(A)$;*
3. *if two events A, B are such that $A \subset B$, $P(B) = P(A) + P(B \backslash A)$. In particular, $P(A) \leq P(B)$;*
4. *for two events A, B,*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B);$$

5. *finite σ-additivity : if $n \geq 2$, and $A_1, \ldots, A_n$ are events such that $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i);$$

6. *countable σ-additivity : if $A_1, A_2, \ldots$ are events such that $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i);$$

7. *finite σ-sub-additivity : if $n \geq 2$, and $A_1, \ldots, A_n$ are events, then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i);$$

8. *countable σ-sub-additivity : if $A_1, A_2, \ldots$ are events, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i);$$

9. *monotone convergence, increasing sequences : if $A_1, A_2, \ldots$ are events such that $A_i \subset A_{i+1}$ for all i's then*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right);$$

10. *monotoneconvergence, decreasing sequences : if $A_1, A_2, \ldots$ are events such that $A_{i+1} \subset A_i$ for all i's, then*

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Remarque It's not necessary to learn this list by heart

We will encounter two main types of probability spaces in these notes :

- **Discrete probability spaces :** in that case Ω is a finite or a countable set, and the set of events really is $\mathcal{F} = \mathcal{P}(\Omega)$.
- **Continuous probability spaces :** in that case, $\Omega = \mathbb{R}^d$ with $d \geq 1$ integer. We won't go into a formal definition of the *set of Borel sets*, and we will do as if we could take $\mathcal{F} = \mathcal{P}()$

1.1.2 Discrete probability measures

Définition 2

Let Ω be a finite countable set. A **probability mass function** on Ω is a function $p : \Omega \to [0,1]$ such that

$$\sum_{\omega\in\Omega} p(\omega) = 1$$

The **probability measure** associated to a probability mass function p is the function $P_p : \mathcal{P}(\Omega) \to [0,1]$ given by

$$P_p(A) = \sum_{\omega\in A} p(\omega).$$

1.1.3 Continuous probability measures

One cannor make sense of the probability that a drop of water falls at *precisely* one point x, but it is relatively easy to make sense of the probability that the drop falls *is a small disk* around x. This is the essence of the next definition.

Définition 3

Let $d \geq 1$. A **probability density function** on $\mathbb{R}^d$ is a Riemann integrable function $f : \mathbb{R}^d \to [0, +\infty)$ such that

$$\int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{+\infty} dx_d f(x_1, \ldots, x_d) = 1.$$

The **probability measure** associated to a probability density function f is the $[0,1]$-valued function P_f given by

$$P_f(A) = \int_{-\infty}^{+\infty} dx_1 \dots \int_{-\infty}^{+\infty} dx_d f(x_1, \dots, x_d) \mathbb{1}_A(x_1, \dots, x_d).$$

Remarque

This is the equivalent of **the second definition** but for a space where you can't delimit the element to sum : a continuous space.
The **density function** is a function that shows where the variable like to be at most. The probability is the area under this curve.
$\mathbb{1}_A = \begin{cases} 1 \text{ if } x \in A, \\ 0 \text{ otherwise.} \end{cases}$

1.1.4 Genreal probability spaces and measures

Définition 4

Let Ω be a set. A **sigma-algebra** on Ω is a set $\mathcal{F} \subset \mathcal{P}(\Omega)$ which satisfies

1. $\mathcal{F}$ contains the empty set ($\emptyset \in \mathcal{F}$).
2. $\mathcal{F}$ is stable by taking the complement ($A \in \mathcal{F} \implies \Omega \backslash A \in \mathcal{F}$).
3. $\mathcal{F}$ is stable by countabe unions (if for all i $\in \mathbb{N}$, $A_i \in \mathcal{F}$, then $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$

Remarque

$\mathcal{P}(\Omega)$ is the set of all subset of Ω.

Now whe can deduce the following properties.

Théorème 2

Let Ω be a set and $\mathcal{F}$ a sigma-algebra on Ω. Then all of the following hold.

1. $\Omega \in \mathcal{F}$.
2. *$\mathcal{F}$ is stable by finite intersections : if $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$.*
3. *If $A, B \in \mathcal{F}$, then $A \backslash B \in \mathcal{F}$.*
4. *$\mathcal{F}$ stable by countable intersections : if $A_1, A_2, \dots \in \mathcal{F}$, then $\bigcap_{i \geq 1} A_i \in \mathcal{F}$.*
5. *$\mathcal{F}$ is stable by increasing limits : if $A_i \in \mathcal{F}$, $i \geq 1$ is such that $A_i \subset A_{i+1}$, then, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.*
6. *$\mathcal{F}$ is stable by decreasing limits : if $A_i \in \mathcal{F}, i \geq 1$ is such that $A_{i+1} \subset A_i$, then, $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Définition 5

Let Ω be a set and $\mathcal{F}$ a sigma-algebra on Ω. A **probability measure** on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0,1]$ such that

1. $P(\Omega) = 1$.
2. If $A_i \in \mathcal{F}, i \in \mathbb{N}$ is a sequence of events with $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i).$$

Remarque

The same as before but more precise.

Définition 6

A **proability space** is a triplet $\Omega, \mathcal{F}, P$ where Ω is a set (the set of realisations), $\mathcal{F}$ is a sigma-algebra on Ω (the set of events), and P is a probability measure on $(\Omega, \mathcal{F})$.

Remarque **This is the most important definintion of this "introduction"**

1.1.5 Inclusion-Exclusion

It is a generalisation of the following fact that we encounter when counting objects : to count the number of objects with property A or property B, we can count the number of objects with property A add the number of objects with property B, and correct our over-counting by removing from this the number of objects with both property A and property B (which were counted twice).

Théorème 3

Let P be a probability measure on some realisation ser Ω. Let $n \geq 1$ and $A_1, \ldots, A_n$ be events. Then,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i \leq i_1 < \ldots \leq i_k \leq n} P\left(\bigcap_{j=1}^{k} A_{i_j}\right).$$

Moreover, for $1 \leq l \leq \frac{n}{2}$ integer a,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \begin{cases} \leq \sum_{k=1}^{2l-1} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} P\left(\bigcap_{j=1}^{k} A_{i_j}\right) \\ \geq \sum_{k=1}^{2l} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} P\left(\bigcap_{j=1}^{k} A_{i_j}\right) \end{cases}.$$

1.2 Random variables and expectation

1.2.1 Random variables

Random variables are therefore functions going from the set of realisations to the real numbers ; for example, if the "experiment " is looking at all poeple born in 2000, one could make the measurement of the height of the first individual born that year.

Définition 7

A (real) **random variable** is a function from the realisation space Ω to $\mathbb{R}$. The probability that a random variable falls in a set A is

$$P(X \in A) := P\left(X^{-1}(A)\right)$$

In words : it is the probability that the realisation of the experiment is such that the measurement X takes a value A.

Example

We are throwing a dice :

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $X(\omega) = 1$ if it's even, $X(\omega) = 0$ if it's odd.
- So we are searching ω in Ω that gives $X(\omega) = 1$:

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} = \{2, 4, 6\}$$

Then we apply the probability on these results

$$P(X \in A) = P(\{2, 4, 6\}) = \frac{3}{6} = 0.5$$

We will frequently use notations similar to the following :

$$\begin{aligned} P(X = x) &\equiv P(X \in \{x\}), \\ P(X \leq x) &\equiv P(X \in (-\infty, x]), \\ P(X > x) &\equiv P(X \in (x, +\infty)). \end{aligned}$$

Définition 8

Le Ω be a set of realisation, and let P be a probability measure on Ω. Let $X : \Omega \to \mathbb{R}$ be a random variable. X is **discrete** if there is $\mathcal{D}_X \subset \Omega$ countable or finite such that $P(X \in \mathcal{D}_X) = 1$. The **law of X** is then the probability measure on $\mathbb{R}$ given by

$$P_X(A) = \sum_{x \in A \cap \mathcal{D}_X} P(X = x)$$

In words, a **discrete random variable** is a variable that can take only finitely or countabley many values with non-zero probability.

The second very important family of variables are **continuous random variables**.

Définition 9

Let Ω be a set of realisation, and let P be a probability measure on Ω. Let $X : \Omega \to \mathbb{R}$ be a random variable. X is **continuous random variable** if there is a density function $f_X : \mathbb{R} \to [0, +\infty)$ such that

$$P(X \in A) = \int_{\mathbb{R}} \mathbb{1}_A f_X(x)\, dx.$$

The **law of X** is then the probability measure on $\mathbb{R}$ given by P_{f_X}.

Remarque

Here we can't say : "the probability that $X = 2$", because it will always be 0. Instead we use a density $f_X(x)$ to compute on an interval. For exemple, if X is mesuring the size of somebody, $P(170 \leq X \leq 180)$ is compute with the density f_X by an inegral, because the probability to have 170.000000cm is 0.

- Discrete : whe can say $P(X = x)$.
- Continue : values are infinite, we look at interval not precise points.

There is a similat notion of law for general random variables. Random variables allows us sometime to pass from some continuous proability to some discrete random variables.

Définition 10

Let Ω be a seet of realisation, and let P be a probability measure on Ω. Let $X : \Omega \to \mathbb{R}$. The **law of X** is the probability measure P_X on $\mathbb{R}$ given by

$$P_X(A) = P(X \in A).$$

Définition 11

Let $X : \Omega \to \mathbb{R}$ be a random variable. The **cumulative distribution function of X** is defined by

$$F_X(t) = P(X \leq t).$$

Two random variables have the same law if and only if they have the same cumulative distribution functions. Note that if X is a continuous random variable with density f_X, one has that F_X is a primitive of f_X :

$$F'_X(t) = f_X(t) \ \forall t \in \mathbb{R}.$$

Remarque

$F_x : \mathbb{R} \to [0,1]$

Example 1 : If $P(X=0) = P(X=1) = \frac{1}{2}$

$$F_X(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2} & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

Théorème 4

Let $x \to \int_a^x f(t)\,dt$ is primitive of f
If X is a continuous random variable : there exist a density function

$$f_X : \mathbb{R} \to \mathbb{R},\ P(X) = \int_{\mathbb{R}} \mathbb{1}_A(x) f_X(x)\,dx$$

So F_X is differentiable and

$$F'_X(t) = f_x(t) = \int_{-\infty}^{t} f_X(x)\,dx$$

Example

If X is a continuous random variable with density $f_X(x) = \mathbb{1}_{[0,4]}(x)$ then

$$P(X \leq t) = \int_{-\infty}^{t} f_X(x)\,dx = \frac{1}{4}\int_{-\infty}^{t} \mathbb{1}_{[0,4]}(x)\,dx = \begin{cases} 0 & \text{if } t < 0, \\ \frac{t}{4} & \text{if } t \in [0,4], \\ 1 & \text{if } t > 4. \end{cases}$$

Remarque

to

Remarque importante

To resume an important point :

- for discrete random variables we use a **probability masse function**.
- for continue random variables we use a **density function**.

1.2.2 Expectation

Définition 12

Discrete case : If Ω is a finite or countable set, $p : \Omega \to [0,1]$ is a probability mass function, and $X : \Omega \to \mathbb{R}$ is a random variable such that

$$\sum_{\omega \in \Omega} |X(\omega)| \, p(\omega) < \infty,$$

the **expectation of X under p** is

$$E_p(X) := \sum_{\omega \in \Omega} X(\omega) \, p(\omega).$$

In this case, we say that X is P_p**-integrable**.
Continuous case : If $f : \mathbb{R}^d \to \mathbb{R}$ is a probability density function, and $X : \mathbb{R}^d \to \mathbb{R}$ is a random variable such that

$$\int_{\mathbb{R}^d} dx \, |X(x)| \, f(x) < \infty,$$

the **expectation of X under** P_f is

$$E_f(x) := \int_{\mathbb{R}^d} dx X(x) \, f(x).$$

In this case, we say that X is P_f**-integrable**.
General case : If Ω is a set of realisation, and P is a probability measure on Ω, we will denote $E_p(X)$ the **expectation of X under P**.

Remarque

Note that the expectede value of a random variable depends only on its law

Example

Look at a fair 6-face dice roll : $\Omega = \{F1, F2, \ldots, F6\}, p(\omega) = \frac{1}{6}$ for every $\omega \in \Omega$.
Take the random variable $X(F1) = X(F3) = X(F5) = -1, X(F2) = X(F4) = X6 = 2$, then

$$E_p(X) = p(F1) \, X(F1) + p(F2) \, X(F2) + p(F3) \, X(F3) + p(F4) \, X(F4)$$

$$+ \mathrm{p}(F5) \, X(F5) + p(F6) + X(F6) = -\tfrac{1}{6} + \tfrac{1}{6} \cdot 2 - \tfrac{1}{6} + \tfrac{1}{6} \cdot 2 - \tfrac{1}{6} + \tfrac{1}{6} \cdot 2 = \tfrac{1}{2}.$$

The properties of expectation are summarized in the next Theorem.

Théorème 5

Let Ω be a realisation set, and P a probabilisty measure on Ω. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then,

1. *linearity : for any $a, b \in \mathbb{R}, E_P(aX + bY) = aE_P(X) + bE_P(Y)$;*
2. *ordering : if $P(X \geq Y) = 1, E_P(X) \geq E_P(Y)$. In particular,*
   - *if $P(x \geq 0) = 1$, then $E_P(X) \geq 0$;*

- *if* $P(a \leq X \leq b) = 1$, *then* $a \leq E_P(X) \leq b$;
- $|E_P(X)| \leq E_P(|X|)$.

1.2.3 Transfer Theorem

The transfer theorem confirm us that the intuition that if we have a random variable that takes values x_1, x_2, x_3, with probabilities p_1, p_2, p_3, and we have a function $g : \mathbb{R} \to \mathbb{R}$, then expectation of $g(X)$ should be

$$E(g(X)) = p_1 g(x_1) + p_2 g(x_2) + p_3 g(x_3).$$

Théorème 6

Let X be a random variable. Then,

1. *if X is a discrete random variable, for any $g : \mathbb{R} \to \mathbb{R}, g(X)$ is a discrete random variable, and*
$$E(g(X)) = \sum_{x \in \ Image\ (X)} g(x) P(X = x)$$
as soon as the sum converges absolutely ;
2. *if X is a continuous random variable, for any $g : \mathbb{R} \to \mathbb{R}$*
$$E(g(X)) = \int_{-\infty}^{+\infty} f_X(x) g(x) dx$$
as soon as the integral converges absolutely.

1.2.4 Random vectors

Définition 13

Let Ω be a realisation set, and P a probability measure on Ω. Let $d \geq 1$. A **random vector of dimension d** is function $X : \Omega \to \mathbb{R}^d$. We will denote

$$X(\omega) = (X_1(\omega), \ldots, X_d(\omega)).$$

The functions $X_i : \Omega \to \mathbb{R}$ are random variables. They are calles the **marginals** of X.
The **cumulative distribution function** (CDF) of a random vector $X : \Omega \to \mathbb{R}^d$ is given by $F_X : \mathbb{R}^d \to [0, 1]$,

$$F_X(t_1, \ldots, t_d) = P(X_1 \leq t_1, \ldots, X_d \leq t_d)$$

Remarque Random vectors are just a list of random variables.

Pas sur de l'utilité de celle la

Définition 14

Let Ω be a realisation set, and P a probability measure on Ω. A **complex random variable** is a function $X : \Omega \to \mathbb{C}$. The real and imaginary parts of X are then random variables.

There is then also discrete and continuous random vectors

Définition 15

Let Ω be a realisation set, and P a probability measure on Ω. Let $X : \Omega \to \mathbb{R}^d$ be a random vector. We say that

- X is a **discrete random vector** if there is a finite of countable set $\mathcal{D}_X \subset \mathbb{R}^d$ with

$P(X \in \mathcal{D}_X) = 1$;

- X is a **continuous random vector** if there is a density function $f_X : \mathbb{R}^d \to [0, +\infty)$ such that

$$P(X \in A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)\, dx$$

Example

Consider (X, Y) a uniform random vector in the unit disc :

$$f_{(X,Y)}(x, y) = \frac{1}{\pi} \mathbb{1}_{[0,1]}\left(x^2 + y^2\right).$$

The first marginal, X, of this random vector is then a continuous random variable with density given by $f_X(x) = 0$ for $|x| > 1$, and, for $|x| \leq 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)\, dy =$$

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \mathbb{1}_{[0,1-x^2]}\left(y^2\right) dy = \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy = \frac{2}{\pi}\sqrt{1-x^2}.$$

Hors sujet important

Let a random vector have two variables, then his density function will be $f_{XY}(x, y)$. So the density of the random variable X_1 will be $f_X(x) = \int_{\mathbb{R}} f_{XY}(x, y)\, dy$ and the expected value of X1 will be $E[X_1] = \int_{\mathbb{R}} x f_X(x)\, dx = \int_{\mathbb{R}} \int_{\mathbb{R}} x f_{XY}(x, y)\, dxdy$.

1.2.5 Change of variable formula

Let $d \geq 1, U \subset \mathbb{R}^d$ an open set, and $\phi : U \to \mathbb{R}^d, \phi(y) = (\phi_1(y), \ldots, \phi_d(y)), y = (y_1, \ldots, y_d)$. Then,

- we say that ϕ is **continuously differential** on U if the partial derivative $\frac{\delta \phi_i}{\delta y_j}$ exists and are continuous on U ;
- we denote $D_\phi(y)$ the **Jacobian matrix** of ϕ ;
- we denot det the determinant.

Théorème 7

Let $d \geq 1, U \subset \mathbb{R}^d$ an open set, $V \subset \mathbb{R}^d$, and $\phi : U \to V$ a continuously differentiable bijection with $detD_\phi(y) \neq 0$ for all $y \in U$. Then, for any function $f : V \to \mathbb{R}$, we have :

$$\int_U dy_1 \ldots dy_d f(\phi(y)) \left|detD_\phi(y)\right| = \int_V dx_1 \ldots dx_d f(x).$$

From this, we can deduce :

Théorème 8

Let Ω be a realisation set, P be a probability measure on Ω, and $X : \Omega \to \mathbb{R}^d$ a random vector with a density f_X. Let $U, V, \phi : U \to V$, be as in ***the theorem above****. Suppose that $P(X \in U) = 1$. We then have that $Y = \phi \odot X : \Omega \to \mathbb{R}^d$ $(Y = \phi(X))$ is a continuous random vector with density*

$$f_Y(y) = f_X\left(\phi^{-1}(y)\right) \left|detD_{\phi^{-1}}(y)\right| = \frac{1}{\left|detD_\phi\left(\phi^{-1}(y)\right)\right|} f_X\left(\phi^{-1}(y)\right).$$

In the case $d = 1$, this formula simplifies to

$$f_Y(y) = f_X\left(\phi^{-1}(y)\right)\left|\left(\phi^{-1}\right)'(y)\right| = \frac{1}{\left|\phi'\left(\phi^{-1}(y)\right)\right|} f_X\left(\phi^{-1}(y)\right).$$

Example

We take U a uniform random variable on $[0,1]$: U is a continuous random variable with density $f_U(x) = \mathbb{1}_{[0,1]}(x)$. Then for $a \in \mathbb{R}$ and $r > 0$, define $X = a + rU$. Using theorem 8 with $\phi(x) = a + rx, \phi^{-1}(x) = \frac{x-a}{r}$, we get that X is a continuous random variable with density

$$f_X(x) = f_U\left(\phi^{-1}(x)\right)\frac{1}{\phi'\left(\phi^{-1}(x)\right)} = \mathbb{1}_{[0,1]}\left(\frac{x-a}{r}\right)\frac{1}{r} = \frac{1}{r}\mathbb{1}_{[a,a+r]}(x).$$

So, X is a uniform random variable on $[a, a+r]$.

Remarque

With polar coordinates, we then have

$$\phi^{-1}(x,y) = \left(\text{atan2}(y,x), \sqrt{x^2+y^2}\right)$$

where

$$\text{atan2}(y,x) = \begin{cases} \arctan\left(\frac{y}{x}\right) & \text{if x } > 0, \\ \arctan\left(\frac{y}{x}\right) + \pi & \text{if x } > 0, \\ \arctan\left(\frac{y}{x}\right) + \pi & \text{if x } > 0, \\ \frac{\pi}{2} & \text{if } x = 0, y > 0, \\ -\frac{\pi}{2} & \text{if } x = 0, y < 0, \\ \text{undefined} & \text{if } x = y = 0. \end{cases}$$

In this case

$$|\text{det}D_\phi(\theta, r)| = r,$$

which leads to the formula

$$dxdy = rd\theta dr$$

1.2.6 Moments, and Moment Generating Function

We already saw the cumulative distribution function (CDF), there is an other useful object sometime : the **moment generating function**

Définition 16

Let X be a random variable. Let $p > 0$. We say that X **admits a moment of order p** if

$$E\left(|X|^p\right) < \infty$$

When X admits a moment of order p, we define

- the *pth moment* of X : $E\left(X^p\right)$;
- the *pth absolute moment* of X : $E\left(|X|^p\right)$.

Définition 17

Let X be a random variable. we say that X **admits exponential moments of order** $\delta > 0$ if

$$E\left(e^{\delta|X|}\right) < \infty$$

When X admits exponential moments, we define the **moment generating function of** X by

$$M_X(t) = E\left(e^{tx}\right),\ t \in (-\delta, \delta).$$

Théorème 9

Let X, Y be two random variables. Suppose that there is $\delta > 0$ such that

$$E\left(e^{\delta|X|}\right) < \infty,\ E\left(e^{\delta|Y|}\right) < \infty.$$

Then, we have the following properties.

- *X admits moments of any integer order.*
- *M_X is analytic in a neighbourhood of 0, and for any $n \in \mathbb{N}$,*
- *M_X, M_Y characterise X, Y :*

$$M_X(t) = M_Y(t) \text{ for all } t \in (-\delta, \delta) \implies X = Y.$$

1.3 Conditional probability independence

1.3.1 Conditional probability

Définition 18

Let Ω be a set of realissations, and P a probability measure on Ω. Let $A \subset \Omega$ be an event such that $P(A) > 0$. Defin then the **probability measure P conditioned on A**, denoted $P(|A)$, by

$$P(B|A) = \frac{P(A \cap B)}{P(A)},\ \forall \text{ event } B.$$

We can then defin the **conditions expectation** of a random variable $X : \Omega \to \mathbb{R}$ by

$$E_P(X|A) = E_{P_A}(X),$$

where P_A stands for $P_A = P(|A)$.

1.3.2 Independence

Définition 19

- Two events A, B are said to be **independent** if

$$P(A \cap B) = P(A) P(B).$$

- A family of events $(A_i)_{i \in I}$ is said to be **two-by-two independent** if for any $i \neq j$, A_i and A_j are independent.
- A family of events $(A_i)_{i \in I}$ is said to be **an independent family** if for any $J \subset I$ finite,

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

Définition 20

- Two random variables X, Y are said to be **independent** if for any events $A, B \subset \mathbb{R}$,

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B).$$

Equivalentyl, X,Y are independent if for any $f, g : \mathbb{R} \to \mathbb{R}$,

$$E\left(f\left(X\right)g\left(Y\right)\right) = E\left(f\left(X\right)\right)E\left(g\left(Y\right)\right).$$

- A familiy of random variables $(X_i)_{i\in I}$ is said to be **two-by-two independent** if for any i, X_i and X_j are independent.
- A family of random variables $(X_i)_{i\in I}$ is said to be **an independent family** if for any $J \subset I$ finite, and any events $A_i \subset \mathbb{R}, i \in J$,

$$P\left(\cap_{i\in J}\{X_i \in A_i\}\right) = \prod_{i\in J} P\left(X_i \in A_i\right).$$

Equivalently, $(X_i)_{i\in I}$ is an independent family if for any $J \subset I$ finite, and any functions $f_i : \mathbb{R} \to \mathbb{R}, i \in J$,

$$E\left(\prod_{i\in J} f_i\left(X_i\right)\right) = \prod_{i\in J} E\left(f_i\left(X_i\right)\right).$$

The same definition holds with "random vectors" replacing "random variables".

Définition 21

A family $(X_{ii\in I}$ of random variables is called an **independent identically distributed** family, abbreviated **i.i.d. familiy**, if the familiy $(X_i)_{i\in I}$ is an independent family, and for any $i, j \in I, X_i = X_j$.

Théorème 10

Let $d, d' \geq 1$. Let $X : \Omega \to \mathbb{R}^d, Y : \Omega \to \mathbb{R}^{d'}$ be a random vector.

- *If X, Y **are continuous random vector** : X and Y are independent if and only if the random vector $(X, Y) : \Omega \to \mathbb{R}^{d+d'}$ has density*

$$f_{(X,Y)}\left(x, y\right) = f_X\left(x\right) f_Y\left(y\right),$$

where $f_X : \mathbb{R}^d \to \mathbb{R}$ is a density for X, and $f_Y : \mathbb{R}^{d'} \to \mathbb{R}$ is a density for Y.

- *If X, Y **are discrete random vectors** : X and Y are independent if and only if for any $x \in \mathbb{R}^d, y \in \mathbb{R}^{d'}$,*

$$P\left(X = x, Y = y\right) = P\left(X = x\right) P\left(Y = y\right).$$

- *If X **is discete and Y is continuous** : X and Y are independent if and only if for any $x \in \mathbb{R}^d, A \subset \mathbb{R}^{d'}$,*

$$P\left(X = x, Y \in A\right) = P\left(X = x\right) \int_A f_Y\left(y\right) dy$$

where $f_Y : \mathbb{R}^{d'} \to \mathbb{R}$ is a density for Y.

1.3.3 Bayes law, formula of total probability

Bayes Law :

Théorème 11

Let Ω be a set of realisations, and let P be a probability measure on Ω. Let $A, B \subset \Omega$ be two events such that $P\left(A\right), P\left(B\right) > 0$. Then

$$P\left(A|B\right) = \frac{P\left(B|A\right) P\left(A\right)}{P\left(B\right)}.$$

Théorème 12

Let Ω be a set of realisations, and let P be a probability measure on Ω. Let I be a finite or countable set. Let $A_i, i \in I$ be a collection of events such that

- *if $i \neq j$, then $A_i \cap A_j = \emptyset$;*
- $\cup_{i \in I} A_i = \Omega^a$.

Suppose moreover that $P(A_i) > 0$ for all $i \in I$. Then for any event B,

$$P(B) = \sum_{i \in I} P(B \cap A_i) = \sum_{i \in I} P(B|A_i) P(A_i).$$

In the same fashion, for every random variable X

$$E(X) = \sum_{i \in I} E(X|A_i) P(A_i).$$

Example

We throw some dices :

- A_1 : we throw an even dice
- A_2 : we throw an odd dice
- B : the result is ≤ 4

Then

$$P(B) = P(B|A_1) P(A_1) + P(B|A_2) P(A_2).$$

1.3.4 Almost sure properties

Définition 22

An event A is said to occur **almost-surely** if

$$P(A) = 1.$$

1.4 Correlation

1.4.1 Variance, Covariance

Variance is a way to quantify "how far frome its mean is typically my variable". If every value than X can take is note far from the mean of every value of X, then the variance will be small.

Définition 23

Let Ω be a realisation set and P a probability measure on Ω. Let $X : \Omega \to \mathbb{R}$ be a random variable. The **variance of X** is given by

$$\mathrm{Var}_P(X) := E_P\left((X - E_P(X))^2\right).$$

Alternatively, $\mathrm{Var}_P(X) = E_P\left(X^2\right) - E_P(X)^2$.

The inside of the expected value in the definition on Var $(X - E_P(X))$ is called the standard deviation

Définition 24

The **standard deviation** of a random variable X, often denoted as σ_X, is the square root of its variance :

$$\sigma_X = \sqrt{\mathrm{Var}\,(X)}.$$

Définition 25

Let Ω be a realisation set and P a probability measure on Ω. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables. The **covariance between X and Y** is given by

$$\mathrm{Cov}_P\,(X,Y) := E_P\,(XY) - E_P\,(X)\,E_P\,(Y)\,.$$

When $\mathrm{Cov}_P\,(X,Y) = 0$, we say that X and Y are **uncorrelated**.

Remarque

The covariance between X and Y is a measure of how much "typical large values of X" and "typical large values of Y" are influencing each other. Two independent event have a covariance of 0 (the opposite isn't true !).
But there is one case where uncorrelated implies independent : it's with Bernoulli random variables

Théorème 13

Let x, Y be two random variables such that

$$P(X \in \{0,1\}) \equiv P(Y \in \{0,1\}) = 1.$$

*Suche variable are called **Bernoulli random variables**. Then, X and Y are independent if and only if* $Cov\,(X,Y) = 0$.

Théorème 14

Let X, Y, Y_1, Y_2 *be random variables and* $a, b \in \mathbb{R}$. *Then,*

$$Cov(X,Y) = Cov(Y,X), \quad Cov(aX,bY) = abCov(x,Y),$$

$$Cov(X_1, Y_1 + Y_2) = Cov(X,Y_1) + Cov(X,Y_2).$$

In words : Cov is symmetric, and linear im each of its arguments.

1.4.2 Pearson correlation coefficient

In statistics, a coefficient obtained form the covariance and standard deviation is frequently used : the Pearson correlation coefficient.

Définition 26

For two random variables X, Y define their **Pearson correlation coefficient** :

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_Y \sigma_x} \in [-1,1].$$

$|\rho_{X,Y}| = 1$ if and only if X and Y are related by an affine transformation (i.e : there are $a, b \in \mathbb{R}$ such that $Y = aX + b$).

Remarque

We normalize by the product of variance because " the height of Bob influences the height of Alice" should not depend on unit we chose to measure height, but the covariance does, so it's a way to correct this

1.5 Classical example of random variables

Ω will be an abstract space of realisation.
P will be an abstract probability measure.

1.5.1 Discrete random variables

Constant random variable

$X : \Omega \to \mathbb{R},\ \omega \to c$. The law of X is a **Dirac measure.**

$$\delta_c(A) = \begin{cases} 1 \text{ if } c \in A \\ 0 \text{ else} \end{cases}$$

Bernoulli random variable

$X : \Omega \to \mathbb{R}$ is a random variables of Bernoulli of parameter p if

$$P(X = 1) = p = 1 - P(X = 0)$$

Example

$A \in P(\Omega)$ event,

$$\mathbb{1}_A : \Omega \to \mathbb{R},\ \mathbb{1}_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ else} \end{cases}$$

is a random variable of Bernoulli parameter $P(A)$.

Binomiale random variable

A random variable $X : \Omega \to \mathbb{R}$ is a binomial random variable with parameter $n \in \mathbb{N},\ p \in [0, 1]$, denoted $X \sim \text{Bin}(n, p)$ if

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

In particular $P(X \in \{0, \dots, n\}) = 1$.

Théorème 15

Let $n \in \mathbb{N},\ p \in [0, 1]$. Let $X_1, \dots, X_n$ be an independent family of Bernoulli random variables of parameter p. Define

$$Y = \sum_{k=1}^{n} X_k.$$

Then, $Y \sim\ Bin(n, p)$.

Geometric random variable

A random variable $X : \Omega \to \mathbb{R}$ is a geometric random variable with parameter $p \in [0, 1]$, denoted $X \sim \text{Geo}(p)$, if

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

Théorème 16

Let $X_1, X_2, \dots$ be an i.i.d sequence of Bernoulli random variables with parameter p. Define

$$Y = 1 + \sum_{n \geq 1} \prod_{i=1}^{n} (1 - X_i),$$

the number of trials before getting a 1 in the sequence. Then, $Y \sim$ Geo (p).

Théorème 17

Let $X \sim$ Geo (p) be a geometric random variable. Then, for any $n > k \in \mathbb{N}$,

$$P(X = n | X > k) = P(X = n - k).$$

In particular, under the law $P(\,|X > k)$, $X - k$ follows a geometric law of parameter p.

Remarque

We can see this "loss of memory" porperty as follows : a gemotric random variable is the number of independent coin toss needed to make a 1. If we pause after k tosses and that the first k coins all gave 0, the following coins being independent of the first k, we end up with simply a sequence of independent coins tosses, exactly as we started.

Poisson random variable

Let $\lambda \geq 0$. A random variable $X : \Omega \to \mathbb{R}$ is a **Poisson random variable of parameter λ**, denoted $X \sim \text{Poi}(\lambda)$, if

$$P(X = k) = \mathbb{1}_{k \in \mathbb{N}} e^{-\lambda} \frac{\lambda^k}{k!}.$$

Théorème 18

Let X be a random variable. Then the two following points are equivalent :

- *$X \sim Poi(\lambda)$;*
- *$P(X = 0) = e^{-\lambda}$ and for all $k \in \mathbb{N}$,*

$$\frac{P(X = k + 1)}{P(X = k)} = \frac{\lambda}{k + 1}.$$

Uniform random variable (finite case)

Let $J \subset \mathbb{R}$ be finite. A random variable $X : \Omega \to \mathbb{R}$ is a **uniform random variable on J** denoted $X \sim \text{Uni}(J)$ if

$$P(X = x) = \frac{1}{|J|} \forall x \in J.$$

In particular, $P(X \in J) = 1$. We will often look at $J = \{0, 1, \ldots, n\}$ or $J = \{1, \ldots, n\}$ for some $n \geq 1$.

Théorème 19

Let $J \subset \mathbb{R}$ be finite, and let $X \sim Uni(J)$. Let $I \subset J$. Then, for any $A \subset I$,

$$P(X \in A | X \in I) = \frac{|A|}{|I|}.$$

1.5.2 Continuous random variables

Uniform random variable on an interval

Let $a < b \in \mathbb{R}$. A random variable $X : \Omega \to \mathbb{R}$ is a **uniform random variable on $[a, b]$**, denoted $X \sim \text{Uni}([a, b])$, if it is a continuous random variable with probability density given by

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

Théorème 20

Let $a < b < c < d \in \mathbb{R}$. Then, if $X \in Uni([a,d])$,

$$P(t_1 \leq X \leq t_2 | b \leq X \leq c) = \frac{t_1 - t_2}{c - b},\ \forall b \leq t_1 \leq t_2 \leq c,$$

which is equivalent to say that under the conditioning $\{X \in [b,c]\}$, X is a uniform random variable on $[b,c]$.

Gaussian random variables

Let $\mu \in \mathbb{R}, \sigma \geq 0$. A random variable $X : \Omega \to \mathbb{R}$ is a **Gaussian random variable with mean μ and variance μ^2**, denoted $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, if it is a continuous random variable with density

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

Théorème 21

Let $X, Y : \Omega \to \mathbb{R}$ be two independent Gaussian random variables. Suppose that $x \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Y \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$. Then,

- *the random variable $\widetilde{X} = (X - \mu_1)/\sigma_1$ is a centred and reduced Gaussian random variable : $\widetilde{X} \sim \mathcal{N}(0,1)$;*
- *the random variable $Z = X + Y$ is a Gaussian random variable with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$: $Z \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$.*

Expononential random variable

let $\lambda > 0$. A random variable $X : \Omega \to \mathbb{R}$ is an **exponential random variable of parameter λ**, denoted $X \sim \text{Exp}(\lambda)$, if X is a continuous random variable with density

$$f_X(x) = \mathbb{1}_{[0,\infty)}(x) \lambda e^{-\lambda x}$$

Remarque The exponential random variable is the continuous version of the geometric random variable, it is therefore not a surprise that they share the "memory loss"property.

Théorème 22

Let $\lambda > 0$, and $X \sim Exp(\lambda)$. Then for any $0 < a < b$,

$$P(X \geq b | Y \geq a) = P(X \geq b - a).$$

In particular, under the conditioning $\{X \geq <\}$, the variable $X - a$ is an exponential random variable with parameter λ.

Cauchy random variable

Let $x_0 \in \mathbb{R}$ and $\alpha > 0$. A random variable $X : \Omega \to \mathbb{R}$ is a **Cauchy random variable**, denoted $X \sim \text{Cauchy}(x_0, \alpha)$, if it is a continuous random variable with density

$$f_X(x) = \frac{\alpha}{\pi((x - x_0)^2 + \alpha^2)}.$$

Summary of usual random variables

Variable	Expectation	Variance	Mom. Gen. Fct.
δ_c	c	0	e^{tc}
$\mathrm{Bern}(p)$	p	$p(1-p)$	$1+p(e^t-1)$
$\mathrm{Bin}(n,p)$	np	$np(1-p)$	$\left(1+p(e^t-1)\right)^n$
$\mathrm{Geo}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
$\mathrm{Poi}(\lambda)$	λ	λ	$\exp(\lambda(e^t-1))$
$\mathrm{Uni}(\{0,1,\ldots,n\})$	$\frac{n}{2}$	$\frac{n^2+2}{12}$	$\frac{e^{(n+1)t}-1}{(n+1)(e^t-1)}$
$\mathrm{Uni}(\{1,\ldots,n\})$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{e^{nt}-1}{n(1-e^{-t})}$
$\mathrm{Uni}([a,b])$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{t(b-a)}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	$\exp\left(\mu t+\frac{\sigma^2t^2}{2}\right)$
$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
$\mathrm{Cauchy}(x_0,\alpha)$	N.D.	N.D.	N.D.

1.6 Probabilistic inequalities and applications

1.6.1 Markov's inequality

Théorème 23

Let X be a non-negative random variable ($X:\Omega\to[0,+\infty)$). Then, for any $a>0$,

$$P(X\geq a)\leq\frac{E(X)}{a}.$$

1.6.2 First moment method

The first moment method is a simple observation : if we have a random variable X taking values in the non-negative integers, $P(X\in\mathbb{N})=1$, we can upper bound the probability that X is non-zero by using its mean :

$$P(X\neq 0)=P(X>0)=E(\mathbb{1}_{X>0}X)\leq E(X).$$

Remarque This means that if you have a small expectation it implies that you have a large probability to be 0.

Example

- We define M_n = the maximum length of a consecutive run of 1's in the n bits.
- To study M_n, we look at $Y_{=}k$ the number of runs of 1's of length k.
- We compute the expected value of Y_k :

$$E(Y_k) = (n - k + 1) \cdot 2^{-k}.$$

- Applying the first moment method

$$P(M_n \geq k) \leq E(Y_k) \leq n \cdot 2^{-k}$$

- if $k > \log_2(n)$, then $n \cdot 2^{-k}$ becomes very small, so the probability of having such a long run of 1's is close to 0. In particular, we obtain that the longest run of 1's is at most of oder $\log_2(n)$

1.6.3 Chebychev's inequality

Chebychev's inequality can be seen as a re-phrasing of Markov's inequality.

Théorème 24

Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $g : \mathbb{R} \to \mathbb{R}_+$ be an increasing function such that $g(X)$ is a random variable. Then,

$$P(X \geq a) \leq \frac{E(g(X))}{g(a)}, \forall a \in \mathbb{R}.$$

The following cases are of particular interest.

- *p-th moment version : for all $p \in (0, +\infty)$,*

$$P(X \geq a) \leq \frac{E(|X|^p)}{a^p}, \forall a > 0.$$

- *Exponential version : for all $\delta \in \mathbb{R}_+$,*

$$P(X \geq a) \leq e^{\delta a} E\left(e^{\delta X}\right), \forall a > 0.$$

Remarque

Taking the function $g(x) = x^2$ applied to the random variable $|X - E(X)|$, we obtain the useful particular case

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}, \forall a > 0.$$

1.6.4 A weak Law of Large Numbers

This is an application of Chebychev's inequality. This law states that with high probability, the empirical average of independent indentically distributed (i.i.d) random variables is close to its expectation.

Théorème 25

*Let $X_1, X_2, \ldots$ be a sequence of **identically ditributed** random variables. Suppose that*

1. *thy admit a second moment : $E\left(X_1^2\right) < \infty$,*
2. *they are **uncorrelated** : $Cov(X_i, X_j) = 0$ if $i \neq j$.*

Denote

$$E(X_1) = \mu,\ Var(X_1) = \sigma^2,\ \bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for all $\epsilon > 0$,

$$P\left(\left|\bar{S}_n - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 n}.$$

In particular, $\bar{S}_n$ *converges in probability towards* μ.

1.6.5 Cauchy-Schwartz and Hölder's inequalities

Théorème 26

For any random variables X, Y,

$$E(XY)^2 \leq E\left(X^2\right) E\left(Y^2\right).$$

A direct application is that Pearson correlation coefficient defined in 1.4.2

Théorème 27

Let X, Y *be two random variables. Then,*

$$|Cov(X,Y)| \leq \sqrt{Var(X)}\sqrt{Var(Y)}.$$

In particular, $\rho_{XY} \in [-1, 1]$.

The generalisation of this theoreme is the Hölder's inequality.

Théorème 28

Let $p, q \in (1, +\infty)$ *be such that* $\frac{1}{p} + \frac{1}{q} = 1$. *Then, for all random variables* X, Y,

$$E(|XY|) \left(|X|^p\right)^{\frac{1}{p}} E\left(|Y|^q\right)^{\frac{1}{q}}.$$

1.6.6 Second moment method

The second moment method is useful to show that a random variable is often positiv. This is the complement of the first moment method. It relies on the following :

for X a $\mathbb{N}$-valued random variable,

$$P(X > 0) \geq \frac{E(X)^2}{E(X^2)}.$$

Example

Let defin M_n = the lenght of the longest run of 1's, Y_k = number of consecutive blocks of lenght k consisting only of 1's

- $Y_k > 0 \leftrightarrow$ there exist a run of length $\geq k$
- $P(M_n \geq k) = P(Y_k > 0)$

Step 1 : First Moment
We compute the expectation :

$$E(Y_k) = (n-k+1)\,2^{-k}.$$

This gives an upper bound, if this expectation is small, it's unlikely to have a run of 1's
Step 2 : Second Moment
To show that run actually exists, we use the second moment method :

$$Var(Y_k) = \sum_{i=1}^{n-k+1} \sum_{j=1}^{n-k+1} Cov(B_i, B_j),$$

Where $B_i = \prod_{l=0}^{k-1} X_{i+l}$ is the indicator that the block starting at i is all 1's.

- If the block do not overlap ($|i-j| \geq k$) they are independent $\rightarrow$ covariance = 0.
- If they overlap ($|i-j| < k$), the covariance is compute explicitly : $Cov(B_i, B_j) = d^{-j+i-k} - 2^{-2k}$.

It gives : $Var(Y_k) \leq 3(n-k+1)\,2^{-k}$
Step 3 : Apply the second moment formula

$$P(Y_k > 0) \geq \left(1 + \frac{Var(Y_k)}{(E[Y_k])^2}\right)^{-1} \geq \left(1 + \frac{3}{(n-k+1)\,2^{-k}}\right)^{-1}.$$

Step 4 : Asymptotic consequence
Choose $k = (1-\epsilon)\log_2 n$. Then $2^{-k} = n^{-(1-\epsilon)}$, so :

$$(n-k+1)\,2^{-k} n^{\epsilon}$$

and thus

$$P(M_n \geq (1-\epsilon)\log_2 n) = P(Y_k > 0) \geq \left(1 + \frac{6}{n^{\epsilon}}\right)^{-1} \lim_{n\to\infty} 1.$$

1.6.7 Jensen's inequality

Théorème 29

Let $I \subset \mathbb{R}$ be an interval. Let $g :\rightarrow \mathbb{R}$ be a convex functions. Then, for any random variable X taking values in I ($P(X \in I) = 1$),

$$E(g(X)) \geq g(E(X)).$$

Moreover, for any $h : I \rightarrow \mathbb{R}$ concave,

$$E(h(X)) \leq h(E(X)).$$

Chapitre 2

Convergence of random variables

2.1 What does converge mean ?

2.1.1 Modes of converges

The weakest notion of convergence is the convergence in law. In words, convergence in law means that the statistical properties of the sequence approach the statistcal properties of the limiting random variable as we go further and further in the sequence.

Définition 27

Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. Say that $(X_i)_{i\geq 1}$ **converges in law** towards X, denoted $X_n \to X$, if one of the two following equivalent conditions is fulfilled.

1. $F_{X_n}(t) \to F_X(t)$ for all $t \in \mathbb{R}$ such that F_X is continuous at t.
2. $E_P(\phi(X_n)) \to E_P(\phi(X))$ for all $\phi : \mathbb{R} \to \mathbb{R}$ continuous and bounded.

Remarque

Note thate in the particular case of X being a continuous random variable, F_X is continuous (primitive of the density function). Moreover, still for X continuous, as $P(X = x) = 0$ for all x, we have that $P(X \in (a,b)) = P(X \in [a,b])$ for all $a < b \in \mathbb{R}$. In particular, if X is a continuous random variable, convergence in law of X_n towards X is equivalent to : for any interval $I \subset \mathbb{R}$,

$$P(X_n \in I) \xrightarrow{n\to\infty} P(X \in I).$$

Théorème 30

Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. Suppose that there is $\delta > 0$ such that

$$sup_{|t|<\delta} E\left(e^{tX}\right) < +\infty,\ sup_{|t|<\delta} E\left(e^{tX_i}\right) < +\infty\ \forall i \geq 1,$$

$$E\left(e^{tX_n}\right) \xrightarrow{n\to\infty} E\left(e^{tX}\right)\ \forall |t| < \delta.$$

Then,

$$X_n \xrightarrow{Law} X \text{ as } n \to \infty,$$

and for any $p \geq 1$,

$$E(X_n^p) \xrightarrow{n\to\infty} E(X^p).$$

Examples

1. Let $X_n \sim \text{Bern}\left(p + n^{-1}\right)$, and $X \sim \text{Bern}(p)$. Then, $X_n \xrightarrow{\text{Law}} X$. Indeed, we have $F_X(t) = (1-p)\,\mathbb{1}_{t\geq 0} + p\mathbb{1}_{t\geq 1}$ is continuous everywhere except at 0 and 1.
   For $t \in \mathbb{R}\backslash\{0,1\}$,

$$P(X_n \leq t) = \left(1 - p + n^{-1}\right)\mathbb{1}_{t\geq 0} + \left(p + n^{-1}\right)\mathbb{1}_{t\geq 1} \xrightarrow{n\to\infty} F_X(t).$$

2. Let $X_n \sim \mathcal{N}\left(0, n^{-1}\right)$, and $X \sim \delta_0$. Then, $X_n \xrightarrow{\text{Law}} X$. Indeed, we have $E\left(e^{tX}\right) = 1$, and for any $t \in \mathbb{R}$,

$$E\left(e^{tX_n}\right) = e^{t^2\backslash(2n)} \xrightarrow{n\to\infty} 1.$$

   So Theorem 30 implies $X_n \xrightarrow{\text{Law}} X$. This is a particular instance of : Gaussian random variables with 0 variance are Dirac random variables.

Théorème 31

Let $X : \Omega \to \mathbb{R}$ be a random variable with Image$(X) \subset \mathbb{Z}$. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. Then, the two following conditions are equivalent.

1. $X_n \xrightarrow{Law} X$.
2. *For every $k \in \mathbb{Z}$,*

$$\lim_{\epsilon\to 0^+} \lim_{n\to\infty} P(X_n \in (k-\epsilon, k+\epsilon)) = P(X = k).$$

Now, let's see the second mode of convergence, it is saying : if I allow a small probability of failure ϵ, and a small error of approximation ϵ', I can fing a rank in the sequence which approximate the limit object up to an error ϵ' with probability of success at least $1 - \epsilon$.

Définition 28

Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables. Say that $(X_n)_{n\geq 1}$ **converges in probability** towards X, denoted $X_n \xrightarrow{\text{Proba}} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X - X_n| \geq \epsilon) = 0.$$

The third mode of convergence states that the sequence of random variables converge with probability one.

Définition 29

Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variable. We say that the sequence $(X_n)_{n\geq 1}$ **converges almost surely** towards X, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Théorème 32

$P(\lim_{n\to\infty} X_n = X) = 1$ *is equivalent to*

$$\forall \epsilon > 0,\ \lim_{n\to\infty} P(sup_{m\geq n} |X_m - X| \geq \epsilon) = 0.$$

Définition 30

Let $p \in (0, +\infty)$. Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables admitting a moment of order p. We say that the sequence $(X_n)_{n\geq 1}$ **converges in L^p** towards X, denoted $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} E\left(|X_n - X|^p\right) = 0.$$

And there is a $p = \infty$ version of this definition :

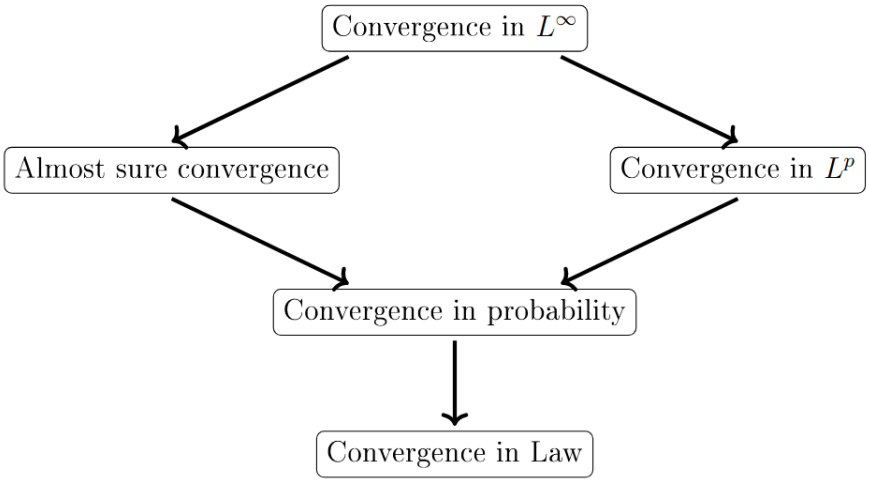
Définition 31

Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be essentially bounded random variables. We say that $(X_n)_{n\geq 1}$ **converges enssentially uniformly** towards X, denoted $X_n \xrightarrow{L^\infty} X$, if

$$\lim_{n\to\infty} \|X - X_n\|_\infty = 0,$$

where $\|Y\|_\infty = inf\{M \geq 0 : \ P\left(|Y| \leq M\right) = 1\}$.

2.1.2 Relations between convergence modes



2.2 Limit theorems

We will now study the two fundamentals results on which most of statistics are based : the **law of large number** which formalizes the fact that the probability of an event represents its occurence frequency ; and the **central limit theorem** which quantifies the typical deviations of frequency in a long sequence of experiments. Both will be about independent sequences of identically random variables : the setup will be as follows.

- Take a random variable X. For example, some measurement in a random experiment.
- Take a sequence $X_1, X_2, \ldots$ of random variables such that
  1. $(X_i)_{i=1,2,\ldots}$ is an independent family,
  2. X_i has the same law as X for every $i \geq 1$.

  This will be the repeated measurement in sequence of repetitions of our random experiment.
- The goal : for n large, study
  1. whether $\frac{1}{n}\sum_{i=1}^n X_i$ gets typically close or not to $E\left(X\right)$ as $n \to \infty$,
  2. quantify how far from $E\left(X\right)$ $\frac{1}{n}\sum_{i=1}^n X_i$ typically is.

2.2.1 Weak Law of Large Numbers

Théorème 33

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if $E(|X|) < \infty$, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X)\right| \geq \epsilon\right) = 0.$$

In other words, $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{Proba} E(X)$ as $n \to \infty$.

Another version of this theorem :

Théorème 34

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if $E\left(X^2\right) < \infty$, for any $\epsilon > 0$, and any $n \geq 1$,

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X)\right| \geq \epsilon\right) \leq \frac{Var(X)}{\epsilon^2 n}.$$

2.2.2 Strong Law of Large numbers

Théorème 35

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if $E(|X|) < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} E(X)$$

as $n \to \infty$.

2.2.3 Central limit theorem

Théorème 36

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if $E\left(X^2\right) < \infty$,

$$\frac{1}{\sqrt{\sigma_X^2 n}}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{Law} \mathcal{N}(0,1)$$

as $n \to \infty$, where

$$\mu = E(X), \sigma_X^2 = Var(X).$$

Another version of this theorem :

Théorème 37

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if $E\left(|X|^3\right) < \infty$, for any $n \geq 1$,

$$sup_{t\in\mathbb{R}}\left|P\left(\frac{1}{\sqrt{\sigma_X^2 n}}\sum_{i=1}^{n}(X_i - \mu) \leq t\right) - P(Z \leq t)\right| \leq \frac{0.5E\left(|X|^3\right)}{\sqrt{n}},$$

where $Z \sim \mathcal{N}(0,1)$, and

$$\mu = E(X),\ \sigma_X^2 = Var(X).$$

Now, let's see a more restrictive version of CLT.

Théorème 38

Let $X, X_1, X_2, \ldots$ be an independent family of identically distributed random variables. Then, if for some $\delta > 0, E\left(e^{\delta|X|}\right) < \infty$,

$$\frac{1}{\sqrt{\sigma_X^2 n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{Law} \mathcal{N}(0,1)$$

as $n \to \infty$, where

$$\mu = E(X), \sigma_X^2 = Var(X).$$

Chapitre 3

Statistics : starter guide

3.1 Basic objects

3.1.1 Statistics

Définition 32

Let P be a probability measure on $\mathbb{R}^d$. Let $n \geq 1$. A **size n sample** of law P is an i.i.d. sequence $X_1, \ldots, X_n$ of random variables/vectors having law P. A **realisation** of a size n sample is a sequence of values of the random variables : that is a sequence $(x_1, \ldots, x_n) \in \left(\mathbb{R}^d\right)^n$.

Définition 33

A **statistic** on a size n sample $X_1, \ldots, X_n$ with values in $\mathbb{R}^d$ is a random variable $Y : \left(\mathbb{R}^d\right)^n \to \mathbb{R}$.

3.1.2 Estimators

Définition 34

We are interested in **parametric estimation**. We will denote $\mathbb{P}_\theta$ the law that we are trying to approximate, $\mathbb{P}_\theta$ is thus a probability measure on $\mathbb{R}^d$. We will suppose that it is entirely determined by a parameter $\theta \in \Theta$, where $\Theta \subset \mathbb{R}^m$ for some $m \geq 1$ is the **parameter space**.

Examples

We could for example look at

1. a squence of height measurements in the Swiss population,
2. a sequence of life duration for computers in a company ,
3. the number of blue cars exiting the highway at Ecublens during a day.

We can now chose an apprriate family of probability measures as candidates for approximating each of these situations.

1. One can use a Gaussian law $\mathbb{P}_\theta = \mathcal{N}\left(\mu, \sigma^2\right)$. The parameter defining the law is $\theta = (\mu, \sigma)$. It leads to the parameter space $\Theta = \mathbb{R}x\,[0, +\infty)$.
2. One can use an exponential law $\mathbb{P}_\theta = \mathrm{Exp}\,(\lambda)$. The parameter is then $\theta = \lambda$, and the parameter space is $\Theta = (0, +\infty)$.
3. One can use a Poisson law $\mathbb{P}_\theta = \mathrm{Poi}\,(\lambda)$. The parameter is then $\theta = \lambda$, and the parameter space is $\Theta = [0, +\infty)$.

Définition 35

Let $X_1, \ldots, X_n$ be a n-sample of law $\mathbb{P}_\theta$. For $f : \Theta \to \mathbb{R}$, an **estimator** of $f(\theta)$ is a statistic $\hat{f} : \left(\mathbb{R}^d\right)^n \to f(\Theta)$, which does not depend on the parameter θ. In words, an estimator is a way to probe the parameter space without prior knowledge of the parameter value to approximate the parameter defining our probability law $\mathbb{P}_\theta$.

The first notion makes precise the idea that as we increase a sample size, we can find better and better approximations of our parameter.

Définition 36

For $n \geq 1$, let $X_1, \ldots, X_n$ be a n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}_n$ be an estimator of $f(\theta)$. We say that the sequence of estimators $\left(\hat{f}_n\right)_{n>=1}$ is **convergent** if $\hat{f}_n$ converges in probability towards $f(\theta)$: for any $\epsilon > 0$ and for any $\theta \in \Theta$,

$$\lim_{n\to\infty} P\left(\left|\hat{f}_n(X_1, \ldots, X_n) - f(\theta)\right| \geq \epsilon\right) = 0$$

The next notions ar formalizations of the estimator gives the correct value on average.

Définition 37

Let $X_1, \ldots, X_n$ be an n-sample of law $\mathbb{P}_\theta$. Let $\hat{f}$ be an estimator of $f(\theta)$. The **bias of** $\hat{f}$ is the difference between the expected value of $\hat{f}$ and $f(\theta)$:

$$\text{Bias}_\theta\left(\hat{f}\right) = E\left(\hat{f}(X_1, \ldots, X_n)\right) - f(\theta).$$

$\hat{f}$ is called **unbiased** if for any $\theta \in \Theta$, $\text{Bias}_\theta\left(\hat{f}\right) = 0$. Otherwise, it is called **biased**.

Définition 38

For $n \geq 1$, let $X_1, \ldots, X_n$ be a n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}_n$ be an estimator of $f(\theta)$. We say that the sequence of estimators $\left(\hat{f}_n\right)_{n\geq 1}$ is **asymptotically unbiased** if for any $\theta \in \Theta$,

$$\lim_{n\to\infty} \text{Bias}_\theta\left(\hat{f}_n\right) = 0.$$

Définition 39

For $n \geq 1$, let $X_1, \ldots, X_n$ be a n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}_n$ be an estimator of $f(\theta)$. The **variance** of $\hat{f}_n$, denoted $\text{Var}_\theta\left(\hat{f}_n\right)$, is simply the variance of the random variable $\hat{f}_n(X_1, \ldots, X_n)$:

$$\text{Var}_\theta\left(\hat{f}_n\right) = \text{Var}\left(\hat{f}_n(X_1, \ldots, X_n)\right).$$

We can give a first example of criterion guaranteeing that a sequence of estimators converges.

Théorème 39

For $n \geq 1$, let $X_1, \ldots, X_n$ be a n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}_n$ be an estimator of $f(\theta)$. Suppose that

1. *$\hat{f}_n$ is asymptotically unbiased,*
2. *$Var\left(\hat{f}_n\right) \xrightarrow{n\to\infty} 0$.*

Then, $\hat{f}_n$ is a convergent sequence of estimators.

3.1.3 Some examples of estimators

Empirical mean

The empirical mean is the estimator we already encountered several times. It is an estimator for the expected value and it is given by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is so widelyused that is got its own standard notation. It is diret to see that this estimator is unbiased :

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = E\left(X_1\right),$$

as the X_i's are identically distributed. Moreover, the weak Law of Large Numbers implies that it is a convergent estimator as soon as $E\left(X_1\right)$ is well defined.

Empirical median

A median for a random variable X is a number M such that

$$P\left(X\right) = P\left(X \geq M\right) = \frac{1}{2}.$$

The empirical median is obtained as follows.

1. Order the sample $X_1, \ldots, X_n$ in a non-decreasing fashion : let $\left(\widetilde{X}_1, \ldots, \widetilde{X}_n\right)$ be a permutation of $X_1, \ldots, X_n$ such that $\widetilde{X}_{i+1} \geq \widetilde{X}_i$ for all i's.
2. Define the empirical median of $X_1, \ldots, X_n$ to be

$$= \begin{cases} \widetilde{X}_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(\widetilde{X}_{\frac{n}{2}} + \widetilde{X}_{1+\frac{n}{2}}\right) & \text{if } n \text{ is even.} \end{cases}$$

One can show that when the law $\mathbb{P}_\theta$ is symmetric around its mean (if it is a continuous law with density f_θ, this means $f_\theta\left(\mu + x\right) = f_\theta\left(\mu - x\right)$ with $\mu = \int_{-\infty}^{+\infty} x f_\theta\left(x\right) dx$, the median is an unbiased estimator of the mean.

Empirical variance

We already have an estimator for the expectation : $\bar{X}_n$. A natural thing to do is to define the empirical variance as

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \bar{X^2}_n - \bar{X}_n^2.$$

I.e : take the difference between the empirical mean of $X_1^2, \ldots, X_n^2$ and the square of the empirical mean of $X_1, \ldots, X_n$.

Empirical covariance

Consider an n-sample $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)$ of random vectors of law $\mathbb{P}_\theta$. The goal is to estimate the covariance $\mathrm{Cov}\left(X, Y\right)$ where $\left(X, Y\right) \sim \mathbb{P}_\theta$. Define the estimator

$$\hat{\tau}_n\left(\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)\right) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right) = \bar{XY}_n - \bar{X}_n\bar{Y}_n.$$

3.2 Constructing estimators

3.2.1 Moments method

Suppose one wants to estimate a parameter θ from a sample $X_1, \ldots, X_n$ of law $\mathbb{P}_\theta$. The general idea is

1. find functions h,g such that we have the relation
$$\theta = h\left(E\left(g\left(X\right)\right)\right)$$
with $X \sim \mathbb{P}_\theta$;
2. use the empirical mean $\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right)$ to estimate $E\left(g\left(X\right)\right)$;
3. use the estimator of θ given by
$$\hat{\theta}_n = h\left(\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right)\right).$$

Example

We can look at the case of $\mathbb{P}_\theta = \text{Geo}\left(p\right)$ (so that $\theta = p, \Theta = [0,1]$). We know that if $X \sim \text{Geo}\left(p\right)$,
$$E\left(X\right) = \frac{1}{p},\ E\left(X^2\right) = \frac{2-p}{p^2}.$$
Each of these give rise to estimator via the moment method.

- First consider the first moment. We can use $h\left(x\right) = x^{-1}$ and $g\left(x\right) = x$. This gives the estimator of p
$$\hat{p}_n\left(X_1, \ldots, X_n\right) = \frac{n}{\sum_{i=1}^{n} X_i}.$$
- Then consider the second moment. We can use $h\left(x\right) = \frac{\sqrt{1+8x}-1}{2x}$ and $g\left(x\right) = x^2$.
This gives the estimator of p
$$\hat{p}_n\left(X_1, \ldots, X_n\right) = \frac{n}{2\sum_{i=1}^{n} X_i^2}\left(\left(1 + \frac{8}{n}\sum_{i=1}^{n} X_i^2\right)^{\frac{1}{2}} - 1\right).$$

3.2.2 Maximum likelihood estimator

This estimator is deeply linked to the Bayes formula : imagine that you have a realisation $x_1, \ldots, x_n$ of a sample $X_1, \ldots, X_n$ of law $\mathbb{P}_\theta$. The idea is to look for the value of θ which maximises the probability to observe the realisation $x_1, \ldots, x_n$.

Définition 40

Let $n >= 1$, and a sample $X_1, \ldots, X_n$ of law $\mathbb{P}_\theta$. For a realisation $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$, the **likelihood** of $x_1, \ldots, x_n$ given θ, $\mathcal{L}\left(x_1, \ldots, x_n|\theta\right)$, is defined by

- if $\mathbb{P}_\theta$ is a discrete probability measure (i.e. : the X_i's are discrete random variables),
$$\mathcal{L}\left(x_1, \ldots, x_n|\theta\right) = P\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} P\left(X_i = x_i\right) = \prod_{i=1}^{n} \mathbb{P}_\theta\left(\{x_i\}\right)$$
as the X_i's are independent and of law $\mathbb{P}_\theta$;
- if $\mathbb{P}_\theta$ is a continuous probability measure (i.e. : the X_i's are continuous random variables),
$$\mathcal{L}\left(x_1, \ldots, x_n|\theta\right) = \prod_{i=1}^{n} f_{X_i}\left(x_i\right) = \prod_{i=1}^{n} f_\theta\left(x_i\right)$$

where f_θ is the density function associated to $\mathbb{P}_\theta$.

Définition 41

The **maximum likelihood estimator** of θ is given by

$$MLE\left(X_1,\ldots,X_n\right) = argmax_\theta \mathcal{L}\left(X_1,\ldots,X_n|\theta\right),$$

the point θ at which $\mathcal{L}\left(X_1,\ldots,X_n|\theta\right)$ is maximized.

Example

Consider the case $X_1,\ldots,X_n$ is a n-sample of law $\mathcal{N}\left(\mu,\sigma^2\right)$ (and take $\theta = \left(\mu,\sigma^2\right)$). The likelihood function of a realisation $x_1,\ldots,x_n \in \mathbb{N}^*$ is given by

$$\mathcal{L}\left(x_1,\ldots,x_n|\mu,\sigma^2\right) = \prod_{i=1}^{n} \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

Maximizing $\mathcal{L}$ is equivalent to minimizing $-\ln\left(\mathcal{L}\right)$.

$$-\mathcal{L}\left(x_1,\ldots,x_n|\mu,\sigma^2\right) = \frac{n}{2}\ln\left(2\pi\sigma^2\right) + \sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}.$$

For any fixed σ^2, the minima of this expression is reached at the empirical mean : $\mu_* = \frac{1}{n}\sum_{i=1}^{n} x_i$ (which is found by finding the critical ponts as a function of μ). Then, the minima is reached at the empirical variance : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_*\right)^2$ (which is again found by finding the critical points as a function of). The maximum likelihood estimator of $\left(\mu,\sigma^2\right)$ is thus given by

$$MLE\left(X_1,\ldots,X_n\right) = \left(\bar{X}_n,\hat{\sigma}_n^2\right).$$

3.3 Quantile tables

In the topic of the next section, we will often be confronted to problem of this form :

$$P\left(X \leq t\right) = 0.95,$$

To avoid doing computations every time, people have computed thes values and made quantile tables.

Définition 42

Let X be a random variables, $q > 0$ be an integer. For integer $k \geq 1, t \in \mathbb{R}$ is a **kth q-quantile** of X if

$$P\left(X < t\right) \leq \frac{k}{q} \text{ AND } P\left(X\right) \geq \frac{k}{q}.$$

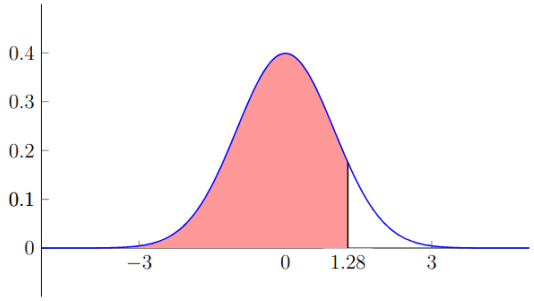
Of X is a continuous random variabl with strictly positive density, there is only one kth q-quantile of X. Sometimes, one speaks about v-quantiles with v $\in (0,1)$. In this case, a kth v-quantile of X is a number $t \in \mathbb{R}$ such that

$$P\left(X < t\right) \leq kv \text{ AND } P\left(X \leq t\right) \geq kv.$$

Example

The 10-quantiles of the $\mathcal{N}(0,1)$ distribution are given in the next table. See the graph after for an illustration of the concept. In the table, t_k is the number such that $P(X \leq t) = \frac{k}{10}$ where $X \sim \mathcal{N}(0,1)$.

k	1	2	3	4	5	6	7	8	9	10
t_k	-1.282	-0.841	-0.524	-0.253	0	0.253	0.524	0.842	1.282	$+\infty$



3.4 Confidence intervals

Confidence intervals will allow to state things such as "given the measurements $X_1, \ldots, X_n$, the parameter θ belongs to the interval $I(X_1, \ldots, X_n)$ with probability $p(X_1, \ldots, X_n)$", where I, p, are to be sepcified. One is typically interested in finding intervals with very short length (precise estimation) with p close to 1 (small probability of error).

Définition 43

Let $X_1, \ldots, X_n$ be a sample of law $\mathbb{P}_\theta$. Let $\alpha \in (0,1)$. A ranndom interval $I = I(X_1, \ldots, X_n)$ note depending on θ is called a **level $1-\alpha$ confidence interval fo $f(\theta)$** if for every $\theta \in \Theta$,

$$P(f(\theta) \in I(X_1, \ldots, X_n)) = 1 - \alpha.$$

$1-\alpha$ is called the **confidence level** of the estimation.

Définition 44

A confidence interval $I = I(X_1, \ldots, X_n)$ is an **excess confidence interval for $f(\theta)$ at level $1-\alpha$** if

$$P(f(\theta) \in I(X_1, \ldots, X_n)) \geq 1 - \alpha.$$

3.5 Hypothese testing

3.5.1 General principle

Let's start with an example.

Example

Let

- p_1 be the probability that a pipe frome company 1 breaks,
- p_2 be the probability that a pipe frome company 2 breaks.

The parameter is $\theta = (p_1, p_2)$ with parameter space $\Theta = [0,1]^2$. Our goal is not to estimate θ precisely but to determine which region of the parameter space it belongs to.

Hypotheses :

- Null hypothesis $H_0 : \theta \in \Theta_0 = \{(p_1, p_2) : p_1 > p_2\}$, (company 1 produces less safes pipes).
- Alternative hypothesis $H_1 : \theta \in \Theta = \{(p_1, p_2) : p_1 \leq p_2\}$. (company 1 is at least as safe as company 2).

Based on the data, we decide whether to reject H_0. Two types of error may occur :

- **Type I error :** Rejecting H_0 when it is true →serious consequence.
- **Type II error :** not rejecting H_0 when it is false →minor consequence.

General framework : we are trying to decide whether a parameter θ belongs to a region $\Theta_0 \subset \Theta$ or not, based on a sample $X_1, \ldots, X_n$ of law $\mathbb{P}_\theta$.

Définition 45

The hypotheses "$\theta \in \Theta_0$", usually denotes H_0, is called the **null hypotheses**, Its complement, the hypotheses "$\theta \in \Theta \backslash \Theta_0$", usually denoted H_1, is called the **alternative hypotheses**.

Définition 46

A **rejection region D** is an event for the random variables $X_1, \ldots, X_n$. I.e : if the X_i take values in $\mathbb{R}^d$, $D \subset \left(\mathbb{R}^d\right)^n$. In practice, one usually takes

$$D = \{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) \in [a, b]\}$$

for some statistic T and real numbers $a \leq b$.

Définition 47

Given D a rejection region, and H_0, H_1 two hypotheses that are tested one against the other, a **test procedure** corresponds to

1. reject H_0 if $(X_1, \ldots, X_n) \in D$;
2. do not reject H_0 if $(X_1, \ldots, X_n) \in D$.

Failure of prediciton are divided into two classes :

- **Type-I error** : we reject H_0 whereas it was correct.
- **Type-II error** : we do not reject H_0 whereas it was false.

Définition 48

Let $\alpha \in [0,1]$. We say that the test procedure has a **risk level** α, or a **confidence level** $1-\alpha$ if

$$sup_{\theta \in \Theta_0} P((X_1, \ldots, X_n)) = \alpha.$$

Définition 49

The **power** of a test is given by

$$inf_{\theta\in\Theta_1} P\left((X_1,\ldots,X_n)\in D\right) = 1-\beta.$$

In particular,

$$\beta = sup_{\theta\in\Theta_1} P\left((X_1,\ldots,X_n)\notin D\right).$$

In words : the risk level α is the "worst case scenario" of the probability to fall in the rejection region whilst having a value of the parameter satisfying H_0 (type-I error), and the β in the power of a test is the "worst case scenario" of the probability to fall outside of the rejection region whilst having a value of the parameter not satisfying H_0 (type-II error).

3.5.2 Chi-square distribution

Définition 50

Let $k\in\mathbb{N}^*$ be a positive integer. A random variable X follows the $\mathcal{X}^2$ **distribution with k degrees of freedom**, denoted $X\sim\mathcal{X}^2_k$, if it is a continuous random variable with density given by

$$f_X(x) = \mathbb{1}_{[0,+\infty)}(x)\,\frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}.$$

Recall the Gamma functiong $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}dx$. For $n\in\mathbb{N}^*$, we have $\Gamma(n)=(n-1)!$

This law have links with the Gaussian random variables,

- if $k\geq 1$ is an integer, and $X_1,\ldots,X_k$ are independent $\mathcal{N}(0,1)$ random variables, then

$$\sum_{i=k}^{k} X_i^2 \sim X_k^2.$$

- if $k\geq 1$ is an integer, $\mu\in\mathbb{R},\sigma^2>0$, and $X_1,\ldots,X_k$ are independent $\mathcal{N}\left(\mu,\sigma^2\right)$ random variables, then the following rescaling of the emprical variance follows a X^2_{k-1} law :

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}\left(X_i=\frac{1}{k}\sum_{i=1}^{k}X_i\right)^2\sim X^2_{k-1}.$$

Définition 51

Let $k\geq 2$ be a positive integer. Let $n\geq 1$. Let $p_1,\ldots,p_k\in[0,1]$ be such that $\sum_{i=1}^k p_i=1$. A random vector $X=(X_1,\ldots,X_k)$ follows the **multinomial distribution with parameters** $(k;n;p_1,\ldots,p_k)$ if it is a discrete random vector with mass function given by

$$P\left(X=(x_1,\ldots,x_k)\right)=\begin{cases}\frac{n!}{x_1!\ldots x_k!}p_1^{x_1}\ldots p_k^{x_k} & \text{if } x_1,\ldots,x_k\in\mathbb{N},\sum_{i=1}^k x_i=n,\\ 0 & \text{else.}\end{cases}$$

When $k=2$, its reduces to the binomial distribution.

The convergence result we will use is the following.

Théorème 40

Let $k\geq 2$, $p_1,\ldots,p_k\in(0,1)$ such that $p_1+\ldots+p_k=1$. For $n\geq 1$, let $(N_{n,1},\ldots,N_{n,k})$ be a multinomial random vector with parameters $(n,k;\ p_1,\ldots,p_k)$. Then the random variable

$$\sum_{i=1}^{k}\frac{(N_{n,i}-np_i)^2}{p_i}$$

converges in law to a $\mathcal{X}^2$ with $k-1$ degrees of freedom :

$$\sum_{i-1}^{k} \frac{(N_{n,i} - np_i)^2}{p_i} \xrightarrow{Law} \mathcal{X}^2_{k-1}, \text{ as } n \to \infty.$$

This is used as follows. Let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be some independent sequence of i.i.d. Let $A_1, \ldots, A_k \subset \mathbb{R}$ be a partition of $\mathbb{R}$:

$$A_i \cap A_j = \emptyset \text{ if } i \neq j, \ \cup_{i=1}^{k} A_i = \mathbb{R}.$$

Then, the random vector

$$\left(\sum_{i=1}^{n} \mathbb{1}_{A_1}(X_i), \sum_{i=1}^{n} \mathbb{1}_{A_2}(X_i), \ldots, \sum_{i=1}^{n} \mathbb{1}_{A_k}(X_i) \right),$$

which simply counts the number of measurements which fell into each class, follows a multinomila distribution with parameters $(n, k; P(X_1), \ldots, P(X_1 \in A_k))$.

3.5.3 $\mathcal{X}^2$ tests

Adequacy tests

Adequacy tests are a way to check whether the observed realisation of the sample was obtain from a given law or not. Consider an experiment with a certain realisation space Ω. Suppose that you can partition Ω into disjoint classes $\Omega_1, \ldots, \Omega_k$. If we let $\mathbb{P}$ be the lae of the sample (which we do not know), we then have the probability of falling in each of the classes is a number $p_i = \mathbb{P}(\Omega_i) \in [0, 1]$, and that $\sum_{i=1}^{k} p_i = 1$. Observing a realisation of a n-sample $X_1, \ldots, X_n$, we can associate the numbers of times weobserved a realisation in the class Ω_i :

$$N_{n,i}(x_1, \ldots, x_n) = \sum_{j=1}^{n} \mathbb{1}_{\Omega_i}(x_j).$$

If we assume that the law of the sample is $\mathbb{P}$, we have thtat the vector $(N_{n,1}, \ldots, N_{n,k})$ folllows a **multinomial distribution** of parameters $(n, k; p_1, \ldots, p_k)$: recall that it is a discrete random vector with, for $m_1, \ldots, m_k \in \mathbb{N}$,

$$P((N_{n,1}, \ldots, N_{n,k}) = (m_1, \ldots, m_k)) = \mathbb{1}_{m_1 + \ldots + m_k = n} \frac{n!}{m_1! \ldots m_k!} p_1^{m_1} \ldots p_k^{m_k}.$$

We can now describe the idea of the test. Our goal is to test wether $\mathbb{P}$, the law of the sample, is a certain law $\mathbb{Q}$ or not. We thus introduce

$$q_i = \mathbb{Q}(\Omega_i), \ i = 1, \ldots, k.$$

We then have that if $\mathbb{P} = \mathbb{Q}$, $p_i = q_i$ for every $i = 1, \ldots, k$ ...but if $p_i \neq q_i$ for some i, then we know for sure that $\mathbb{P} \neq \mathbb{Q}$! We will therefore test the null hypotheses H_0 : "$p_i = q_i$ for $i = 1, \ldots, k$" against the alternative hypotheses H_1 " there is $i \in \{1, \ldots, k\}$ such that $p_i \neq q_i$".
Under H_0, the q_i's should be well approximated by the empirical frequencues : $q_i \approx N_{n,i} \backslash n$. We just transformed a non-parametric question "is $\mathbb{P}$ equal to $\mathbb{Q}$? into a parametric one about the parameters of a multinomial law !

We now need to construct our rejection region. Introduce the statistic

$$Z_n(X_1, \ldots, X_n) = \sum_{i=1}^{k} \frac{(N_{n,i}(X_1, \ldots, X_n) - nq_i)^2}{nq_i} = n \sum_{i=1}^{k} \frac{\left(\frac{N_{n,i}(X_1, \ldots, X_n)}{n} - q_i \right)^2}{q_i}.$$

It is some measure of the difference between the observed frequencies $\frac{N_{n,i}(X_1, \ldots, X_n)}{n}$ and the ones we should see under the null hypotheses. The reason to chose this particular statistic is the lemma thats, in our setup, says that, under the null Hypotheses H_0,

$$Z_n(X_1, \ldots, X_n) \xrightarrow{\text{Law}} \mathcal{X}^2_{k-1}.$$

From this Lemma, we have on the one hand under H_0, Z_n follows asymptotically a $\mathcal{X}^2$ law with $k-1$ degrees of freedom. On the onther hand, under H_1, there is $i_* \in \{1, \dots, k\}$ such that

$$\left(\frac{N_{n,i_*}}{n} - q_{i_*}\right)^2 \xrightarrow{\text{a.s.}} (p_{i_*} - q_{i_*})^2 > 0,$$

as $n \to \infty$, which implies that $Z_n \xrightarrow{\text{a,s,}} +\infty$. If we want a risk level α for our test, we can thus take a rejection region of the form $D = \{Z_n > C\}$ with C such that, with $Z \sim \mathcal{X}^2_{k-1}$,

$$P(Z > C) = \alpha.$$

This gives a region with risk level α asymptotically as, under H_0,

$$P(Z_n > C) \xrightarrow{n\to\infty} P(Z > C).$$

Independence test

We will only consider an example of this test and not describe the general theory. The goal of the test is to test the hypotheses "are property A and property B independent ?". Here are the measures of hairs and eyes colour in a group of people.

eyes\hairs	**Blond**	**Brown**	**Black**	**Ginger**	**Total**	**Freq.**
Blue	25	9	3	7	44	44/124
Grey	13	17	10	7	47	47/124
Brown	7	13	8	5	33	33/124
Total	45	39	21	19	124	
Freq	45/124	39/124	21/124	19/124		

We then want to test wether the colour of the eyes is independent of the colour of the hairs. Let the null hypotheses be "the eyes and hairs colours are independent" against the alternative hypotheses.

Denote the eyes colours by

$$1 \equiv \text{ blue}, \ 2 \equiv \text{ grey}, \ 3 \equiv \text{ brown}.$$

and the hair colours by

$$1 \equiv \text{ blond}, \ 2 \equiv \text{ brown}, \ 3 \equiv \text{ black}, \ 4 \equiv \text{ ginger}.$$

Denote then p_{ij} the probability that an individual has eyes colour i and hair colour j and set

$$p_{i*} = \sum_{i=1}^{4} P_{ik}, \ p_{*j} = \sum_{j=1}^{3} p_{kj},$$

the probability that an individual has eye colour i, hair colour j. Under the null hypotheses, we have that hair and eyes colours are independent, which translates into

$$p_{ij} = p_{i*}p_{*j}, \ i = 1, 2, 3, \ j = 1, 2, 3, 4.$$

We therefore have that the null hypotheses can be re-phrased "$p_{ij} = p_{i*}p_{*j}$ for every $i \in \{1, 2, 3\}$, and every $j \in \{1, 2, 3, 4\}$", which turns out to be a parametric hypotheses.

Under the null hypotheses, we have that the number of individual with eye colour i and hair colour j in a sample of n people is on average

$$np_{ij} = np_{i*}p_{*j}.$$

Also, we know from the LLN that p_{i*} is well approximated by the empirical frequency. We can then test this hypotheses as before by introducing the statistic

$$Z_n = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{\left(N_{ij} - n\frac{N_{i*}N_{*j}}{n^2}\right)^2}{n\frac{N_{i*}N_{*j}}{n^2}} = n\sum_{i=1}^{3}\sum_{j=1}^{4} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i*}N_{*j}}{n^2}\right)^2}{\frac{N_{i*}N_{*j}}{n^2}}$$

Where N_{ij} counts the number of individual eye colour i and hair colour j, N_{i*} counts the number of individual with eye colour i, and N_{*j} with haire colour j. Under the null hypotheses, Z_n converges to a $\mathcal{X}^2_6$ law.

Using our data in Z_{124} with a rejection region at risk 5%, we obtain that we should reject the null hypotheses (so no independence).

3.6 t-test

Définition 52

Let $v \in (0, +\infty)$ be a positive real. A random variable X follows the **Student's t-distribution with parameter ν**, denoted $X \sim \text{Student}_t(\nu)$, if it is a continuous random variable with density given by

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Often, one takes $\nu \in \mathbb{N}^*$. In this case, we call $\text{Student}_t(\nu)$ the **Student's t-distribution with ν degrees of freedom**.

Théorème 41

Let Z, V be two independent random variables with

$$Z \sim \mathcal{N}(0,1),\ V \sim \mathcal{X}^2_k,\ k \in \{2,3,4,\ldots\}.$$

Then, the random variable $Z\sqrt{\frac{k}{V}}$ follows a Student's t-distribution with parameter k :

$$Z\sqrt{\frac{k}{V}} \sim \textit{Student}_t(k).$$

One-sample t-test

It is used to test the null Hypotheses thatt the mean of a sample is given by a fixed value.

Setup : we have a sample $X_1, \ldots, X_n$ of a law $\mathbb{P}$. We want to test the Hypotheses H_0"$E(X_1) = \mu_0$ for a fixed $\mu_0 \in \mathbb{R}$.

Assumptions : to be an exact test, we assume that there are $\mu \in \mathbb{R}$, $\sigma^2 > 0$ such that $\mathbb{P} = \mathcal{N}(\mu, \sigma^2)$. One can make approximate versions of this in applications.

Running the test : we want to test the null Hypotheses "$E(X_1) = \mu_0$" using the test statistic

$$\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}}.$$

Under the null Hypotheses, our assumptions and (3.2) imply that $\frac{\sqrt{n}}{\sigma^2}(\bar{X}_n - \mu_0) \sim \mathcal{N}(0,1)$, and thus that

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}} \sim \text{Student}_t(n-1).$$

The rejection region is then taken to be of the form $\{|T| \geq C\}$ with $C > 0$ depending on the wanted confidence level.

Two-samples t-test

It is used to check whether two independent samples have the same mean or not.

Setup : we have two samples of the same size $X_1, \ldots, X_n$ of law $\mathbb{P}$, $Y_1, \ldots, Y_n$ of a law $\mathbb{Q}$. We want to test the Hypotheses H_0 "$E(X_1) = E(Y_1)$".

Assumptions : to be an exact test, we assum that $\mathbb{P} = \mathcal{N}(\mu_X, \sigma^2)$, and that $\mathbb{Q} = \mathcal{N}(\mu_Y, \sigma^2)$ for some $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma^2 > 0$. Note that we in particular assume that the variance is the same in both sample (σ^2).

Running the test : we will use the test statistic

$$T_n(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \frac{\sqrt{n}(\bar{X}_n - \bar{Y}_n)}{\sqrt{s_X^2 + s_Y^2}}$$

where

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2,\quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2,$$

are the un-biased versions of the empirical variances. Under our assumptions, if H_0 holds (i.e : if $\mu_X = \mu_Y$),

$$\frac{\sqrt{n}}{\sigma^2}\left(\bar{X}_n - \bar{Y}_n\right) \sim \mathcal{N}(0,1), \quad \frac{n-1}{\sigma^2}\left(s_X^2 + s_Y^2\right) \sim \mathcal{X}^2_{2n-2},$$

and $T_n \sim$ Student$_t(2n-2)$. On the other hand, if H_0 does not hold, $\bar{X}_n - \bar{Y}_n \xrightarrow{\text{a.s.}} \delta \neq 0$ as $n \to \infty$, which implies that T_n takes extremely large values for large n. We thus take a rejection region of the form $\{|T_n| \geq C\}$ with C to be chosen depending on the wanted confidence level.

3.7 Comparing estimators

3.7.1 Mean square error

Définition 53

Let $n \geq 1, X_1, \ldots, X_n$ be an n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}$ be an estimator of $f(\theta)$. The **mean square error** of $\hat{f}$ is

$$MSE_\theta\left(\hat{f}\right) = E\left(\left(\hat{f}(X_1, \ldots, X_n) - f(\theta)\right)^2\right).$$

Remarque

$$MSE_\theta\left(\hat{f}\right) = Var_\theta\left(\hat{f}\right) + \left(Bias_\theta\left(\hat{f}\right)\right)^2.$$

Minimizing the mean square error is yet another way to construct an estimator.

Définition 54

Let $n \geq 1, X_1, \ldots, X_n$ be an n-sample of law $\mathbb{P}_\theta$, and $f(\theta)$ be a quantity to estimate. The **minimal mean square error estimator** of $f(\theta)$ is

$$MMSE_{f(\theta)}(X_1, \ldots, X_n) = argmin_{\hat{f}} E\left(\left(\hat{f}(X_1, \ldots, X_n) - f(\theta)\right)^2\right).$$

where the minimum is over all estimators $\hat{f}(X_1, \ldots, X_n)$ of $f(\theta)$. The existence of this estimator is a non-trivial fact of probability theory. It is in general not so easy to compute.

3.7.2 Asymptotic normality

Définition 55

For $n \geq 1$, let $X_1, \ldots, X_n$ be an n-sample of law $\mathbb{P}_\theta$, and let $\hat{f}_n$ be an estimator of $f(\theta)$. We say that $\hat{f}_n$ is an **asymptotically normal** sequence of estimators if

1. $\hat{f}_n$ is convergent ;
2. there is $C \geq 0$ such that $\sqrt{n}\left(\hat{f}_n - f(\theta)\right) \xrightarrow{\text{Law}} \mathcal{N}(0, C)$ as $n \to \infty$.