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*
$$E(g(X)) = \sum_{x \in \text{ Image }(X)} g(x) P(X = x)$$
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 $E[X_1] = \int_{\mathbb{R}} x f_X(x)\, dx = \int_{\mathbb{R}} \int_{\mathbb{R}} x f_{XY}(x, y)\, dx dy$.

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 $f, g : \mathbb{R} \to \mathbb{R}$,

$$E(f(X)g(Y)) = E(f(X))E(g(Y)).$$

- 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 $(X_{i i \in I}$ 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*

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

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 $x \in \mathbb{R}^d, A \subset \mathbb{R}^{d'}$,*

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

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 $P(A), P(B) > 0$. Then

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

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.**

$$\delta_c(A) = \left\{ \begin{array}{l} 1 \text{ if } c \in A \\ 0 \text{ else} \end{array} \right.$$

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,

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

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

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

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

Théorème 15

Let $n \in \mathbb{N},\ p \in [0,1]$. Let $X_1,\ldots,X_n$ 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 $X_1, X_2, \ldots$ 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 $X \sim \mathrm{Poi}(\lambda)$, 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 $X \sim \mathrm{Uni}(J)$ 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 $X \sim \mathrm{Uni}([a, b])$, 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 $\widetilde{X} = \left(X - \mu_1\right)/\sigma_1$ 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 $X \sim \operatorname{Exp}(\lambda)$, if X is a continuous random variable with density

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

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 $X \sim \operatorname{Cauchy}(x_0, \alpha)$, 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}
$\text{Bern}(p)$	p	$p(1-p)$	$1+p(e^t-1)$
$\text{Bin}(n,p)$	np	$np(1-p)$	$\left(1+p(e^t-1)\right)^n$
$\text{Geo}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
$\text{Poi}(\lambda)$	λ	λ	$\exp(\lambda(e^t-1))$
$\text{Uni}(\{0,1,\ldots,n\})$	$\frac{n}{2}$	$\frac{n^2+2}{12}$	$\frac{e^{(n+1)t}-1}{(n+1)(e^t-1)}$
$\text{Uni}(\{1,\ldots,n\})$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{e^{nt}-1}{n(1-e^{-t})}$
$\text{Uni}([a,b])$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{t(b-a)}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	$\exp\left(\mu t+\frac{\sigma^2t^2}{2}\right)$
$\text{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
$\text{Cauchy}(x_0,\alpha)$	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 $a>0$,

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

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, $P\left(X\in\mathbb{N}\right)=1$, we can upper bound the probability that X is non-zero by using its mean :

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

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,

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

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

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