Master X-HEC 2016-17 École Polytechnique

Probability Refresher

Lucas Gerin *(Département de Mathématiques Appliquées, École Polytechnique)*
lucas.gerin@polytechnique.edu

Contents

Merci à Mathieu Richard et Manon Costa pour les nombreuses remarques et corrections sur des versions préliminaires de ces notes.

1 Sets and Measures

The goal of this chapter is to define *probability spaces*, which are basically the sets of outcomes of a random experiment. In particular, we wonder what are the sets that we can compute the probability, such sets will be called *measurable sets*. This approach of probability is due to Kolmogorov (Russia, XXth).

1.1 σ-algebras and measures

Let Ω be a set (later it will be the *sample space* of outcomes of a random experiment) and let $\mathcal{A}$ be a collection of subsets of Ω, $\mathcal{A}$ is to be understood as the collection of events for which one can compute the probability. Here is a more formal definition.

Definition 1.1 (*σ-algebra and measurable sets*)

$\mathcal{A}$ is a σ-algebra *if the following conditions hold:*

i) The empty set $\varnothing$ and the entire set Ω are in $\mathcal{A}$.

ii) If A is in $\mathcal{A}$, then so is A^c (the complement *of A).*

iii) If A_1, A_2 are in $\mathcal{A}$, then so are $A_1 \cup A_2$ and $A_1 \cap A_2$.

iv) More generally, if $A_1, A_2, A_3, \cdots \in \mathcal{A}$, then

$$\bigcup_{n\geq 1} A_n = A_1 \cup A_2 \cup A_3 \cup \cdots \in \mathcal{A}$$
$$\bigcap_{n\geq 1} A_n = A_1 \cap A_2 \cap A_3 \cap \cdots \in \mathcal{A}.$$

If $A \in \mathcal{A}$ we say that A is measurable with respect to $\mathcal{A}$. It is also simply called an event.

Example : For instance if we take for Ω the set of outcomes of a dice

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

here are two examples of σ-algebras:

1. The collection of **all subsets** of Ω

$$\varnothing, \Omega, \{1\}, \{2\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{1, 2, 3\}, \ldots$$

is a σ-algebra. It is denoted by $\mathcal{P}(\Omega)$, and called the *power set*.

2. The collection

$$\mathcal{A} = \varnothing, \Omega, \{1, 3, 5\}$$

is not a σ-algebra since $\{1, 3, 5\}$ is in the collection but not its complement $\{2, 4, 6\}$. However

$$\mathcal{A} = \varnothing, \Omega, \{1, 3, 5\}, \{2, 4, 6\}$$

is a σ-algebra.

Remark : In order to understand item iv), let us imagine what would be the probability space defined by throwing a coin infinitely many times. Let $X_n \in \{$ Heads/Tails $\}$ be the n-th result. Then each event $\{X_n = \text{Heads}\}$ is in the σ-algebra and we would like to compute the probability of, for instance,

$$\bigcap_{n\geq 1} \{X_n = \text{ Heads }\} = \text{"The coin always turns Heads"}.$$

▷ Measures

A pair $(\Omega, \mathcal{A})$ is called a *measurable space*, we can endow it with a *measure*. Basically, a *measure* gives a *mass* to every set.

Definition 1.2

Let $(\Omega, \mathcal{A})$ be measurable space, a measure *on Ω is an application*

$$\mathcal{A} \to [0, +\infty]$$

(note that $+\infty$ is allowed) such that

- $\mu(\varnothing) = 0$
- **(Countable additivity)** *For every disjoint events $A_1, A_2, \ldots$ in $\mathcal{A}$*

$$\mu\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mu(A_n).$$

(This property should remind you of your undergraduate probability courses.)

If furthermore $\mu(\Omega) = 1$ then μ is a probability measure*, and is usually denoted by $\mathbb{P}$.*

A triple $(\Omega, \mathcal{A}, \mu)$ is called a *measured* space. Three important examples:

- The *uniform measure* on a finite set Ω is defined by

$$\mu(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}.$$

- A useful notation is that of the *Dirac measure* (or *Dirac mass*) at some point a, denoted by δ_a. It puts a *mass* one on a:

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

▷ Properties of measures

Plainly from the definition, we get that

Proposition 1.3

Let $(\Omega, \mathcal{A}, \mu)$ be a measured space.

1. *If $A \subset B$, then $\mu(A) \leq \mu(B)$.*

2. *If $\mu(\Omega) < +\infty$,*
$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B),$$
in particular $\mu(A \cup B) \leq \mu(A) + \mu(B)$.

3. **(Union bound)** *More generally, let $(A_n)_{n\geq 1}$ be any sequence of sets (not necessarily disjoint),*
$$\mu\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mu(A_n).$$

4. **(Law of total probability)** *Let A be an event and $B_1, B_2, \ldots$ be a sequence of disjoint sets such that $\cup_n B_n = \Omega$,*
$$\mu(A) = \sum_n \mu(A \cap B_n).$$

Exercise : Prove item 3. (Hint: Put $B_n = A_n \setminus (A_1 \cup A_2 \cup \ldots A_{n-1})$.)

1.2 Borel sets and the Lebesgue measure

Many interesting measures in this course are defined on the real line $\mathbb{R}$, so we need to equip it with a σ-algebra.

Definition 1.4

The Borel σ-algebra *on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra that contains all open intervals (a, b) for every $a, b \in \mathbb{R}$. A set in $\mathcal{B}(\mathbb{R})$ is said to be* a Borel set.

What's there in $\mathcal{B}(\mathbb{R})$? Open intervals of course but also

- closed intervals $[a, b]$, since
$$[a, b] = \left(\underbrace{(-\infty, a)}_{\text{Borel set}} \cup \underbrace{(b, +\infty)}_{\text{Borel set}} \right)^{\mathrm{c}}.$$

- more complicated sets such as $\mathbb{N}$ since it is a countable union of Borel sets:
$$\mathbb{N} = \bigcup_{k\geq 1} \underbrace{\{k\}}_{\text{Borel set}}.$$

In fact, every set you might think of is a Borel set! It is quite difficult (and not very useful for you) to build a set which is not Borel-measurable.

Definition 1.5

A Borel function f *is a function* $\mathbb{R} \to \mathbb{R}$ *such that for all Borel set* A, $f^{-1}(A)$ *is also a Borel set.*
Every continuous function, or piecewise continuous function is a Borel function.

▷ The Lebesgue measure

Many interesting measures in this course are defined on the real line $\mathbb{R}$, so we need to equip it with a measure.

Definition/Theorem 1.6

The Lebesgue measure, *usually denoted by* λ, *is the* **only** *measure on* $\mathbb{R}$ *such that for any real numbers* a, b,

$$\lambda\left((a,b)\right) = b - a.$$

For a set A, $\lambda(A)$ is interpreted as the *length* of A. Let us compute $\lambda(A)$ for some A's:

- For a real a we have, for any integer $n \geq 1$,

  $$\lambda(\{a\}) \leq \lambda\left((a - 1/n, a + 1/n)\right) = 2/n$$

  and thus $\lambda(\{a\}) = 0$. This is consistent since $\{a\}$ has no length.

- For any arbitrary $x > 0$, we have $\lambda(\mathbb{R}) \geq \lambda\left((-x, x)\right) = 2x$, thus $\lambda(\mathbb{R}) = +\infty$.

▷ Random variables

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, a random variable X is a function

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega). \end{array}$$

We say that X is measurable (w.r.t. $\mathcal{A}$) if for every interval I, $X^{-1}(I)$ is in $\mathcal{A}$. Recall that the notation X^{-1} stands for

$$X^{-1}(I) = \left\{\omega \text{ such that } X(\omega) \in I\right\}.$$

Let's see some examples:

- For a finite Ω with its power set $\mathcal{P}(\Omega)$, every function $X : \Omega \to \mathbb{R}$ is measurable.

- A very useful example of measurable function is that of *indicator function*. For $A \in \mathcal{A}$, the indicator function of A is denoted by $\mathbb{1}_A$ and defined by

  $$\begin{array}{rccc} \mathbb{1}_A: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases} \end{array}$$

  Note that indicator functions are very close to Dirac masses: $\mathbb{1}_A(\omega) = \delta_\omega(A)$.

1.3 Limits of events

We now are given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where $\mathbb{P}$ is a probability measure.

Let $(A_n)_{n\geq 1}$ be a sequence of events. The question we address here is, "when do we have $\mathbb{P}(\lim_n A_n) = \lim_n \mathbb{P}(A_n)$?" And, by the way, does $\lim_n A_n$ make sense? A first case is when $(A_n)_{n\geq 1}$ is a sequence of *monotone* events.

Theorem 1.7

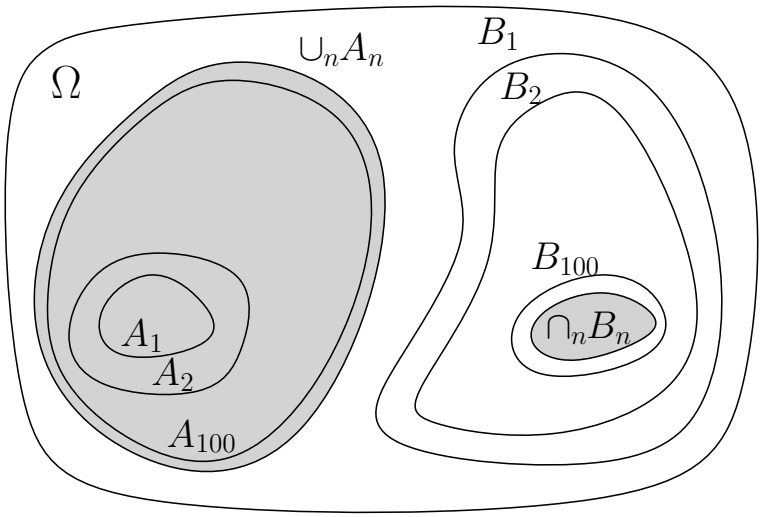
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

1. *Let $(A_n)_{n\geq 1}$ be an increasing sequence of events,* i.e. $A_1 \subset A_2 \subset A_3 \subset \dots$ *Then*

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \lim_{n\to\infty} \nearrow \mathbb{P}(A_n).$$

2. *Let $(B_n)_{n\geq 1}$ be a decreasing sequence of events: $B_1 \supset B_2 \supset B_3 \supset \dots$, then*

$$\mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \lim_{n\to\infty} \searrow \mathbb{P}(B_n).$$



Exercise : Prove item 1. of Theorem 1.7. (Hint: Set $E_n = A_n \setminus A_{n-1}$.)

Example : (A fair coin eventually turns Tails).

We turn back to the example of a fair coin flipped infinitely many times, we want to prove rigorously that the coin turns Tails at least once.

Let $X_n \in \{\text{Heads,Tails}\}$ be the n-th result. We have

$$\{ \text{ the coin never turns Tails } \} = \bigcap_{n\geq 1} \{X_1 = X_2 = \dots = X_n = \text{"Heads"}\} = \bigcap_{n\geq 1} B_n$$

where we set $B_n = \{X_1 = X_2 = \cdots = X_n = \text{"Heads"}\}$. The sequence (B_n) is clearly decreasing:

$$\{X_1 = X_2 = \cdots = X_n = X_{n+1} = \text{"Heads"}\} \subset \{X_1 = X_2 = \cdots = X_n = \text{"Heads"}\}$$

Clearly

$$\mathbb{P}\,(X_1 = X_2 = \cdots = X_n = \text{ "Heads" }) = \frac{1}{2^n}.$$

Thus item 2. in Theorem 1.7 says that

$$\begin{aligned}\mathbb{P}\,(\text{ the coin never turns Tails }) = \mathbb{P}(\cap_{n\geq 1} B_n) &= \lim_{n\to+\infty} \searrow \mathbb{P}\,(B_n) \\ &= \lim_{n\to+\infty} \searrow \mathbb{P}\,(X_1 = X_2 = \cdots = X_n = \text{ "Heads" }) \\ &= \lim_{n\to+\infty} \frac{1}{2^n} = 0.\end{aligned}$$

Then

$$\mathbb{P}\,(\text{ the coin turns Tails at least once }) = 1 - \mathbb{P}\,(\text{ the coin never turns Tails }) = 1.$$

▷ limsup of events

Let (A_n) be a sequence of events, we are often interested in "how many of the A_n's occur?". There is a useful notation for that. Consider the event

$$\begin{aligned}\text{"}A_n \text{ occurs infinitely often"} &= \text{"For any } p\text{, there is } n \geq p \text{ such that } A_n \text{ occurs"} \\ &= \text{"For any } p\text{, } A_p \cup A_{p+1} \cup A_{p+2} \cup \ldots\text{"} \\ &= \textstyle\bigcap_{p\geq 1} \bigcup_{n\geq p} A_n.\end{aligned}$$

This event is denoted by $\limsup_{n\to+\infty} A_n$.

Example : Take $\Omega = \mathbb{R}$ and

$$A_n = \begin{cases} [0,1] \text{ if } n \text{ is odd,} \\ [0,2] \text{ if } n \text{ is even.}\end{cases}$$

Then it is clear that $\bigcup_{n\geq p} A_n = [0,2]$ for all p, and then

$$\limsup_n A_n = \bigcap_{p\geq 1} \bigcup_{n\geq p} A_n = \bigcap_{p\geq 1} [0,2] = [0,2].$$

2 Random variables and expectation

From now on, we work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where $\mathbb{P}$ is a probability measure. We say that an event A is $\mathbb{P}$-*almost sure*, or just *almost sure* if no ambiguity, if $\mathbb{P}(A) = 1$.

Elements of Ω are often denoted by ω. Recall that a random variable X is just a measurable function $X : \Omega \to \mathbb{R}$.

2.1 Random variables and their laws

Definition 2.1

The law *(or* distribution*) of X, denoted by $\mathbb{P}_X$ is the measure on $\mathbb{R}$ such that for any Borel set A*

$$\mathbb{P}_X(A) = \mathbb{P}\left(\{\omega \text{ such that } X(\omega) \in A\}\right) = \mathbb{P}(X \in A).$$

We write $X \sim \mathbb{P}_X$ which reads "X has distribution $\mathbb{P}_X$".

Example : Fair coin. Take $\Omega = \{H, T\}$, $\mathbb{P}$ the uniform measure on Ω and

$$\begin{array}{rccc} X: & H & \mapsto & 2, \\ & T & \mapsto & 0. \end{array}$$

Then

$$\mathbb{P}_X(\{2\}) = 1/2 = \mathbb{P}_X(\{0\}),$$

and we have $\mathbb{P}_X = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$ (while $\mathbb{P} = \frac{1}{2}\delta_H + \frac{1}{2}\delta_T$).

Probability laws are measures and, as such, are complicated objects. This is why we prefer to deal with simpler objects: cumulative distribution functions.

Definition/Theorem 2.2

The cumulative distribution function *(or just distribution function) of X is the function F_X defined by*

$$\begin{array}{rccc} F_X: & \mathbb{R} & \to & [0,1] \\ & t & \mapsto & \mathbb{P}(X \le t). \end{array}$$

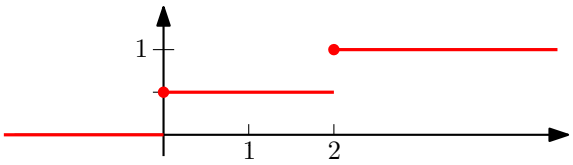
If $F_X(t) = F_Y(t)$ for every t, then X and Y have the same law.

Some properties of F_X:

- If $s \le t$, then $\{X \le s\} \subset \{X \le t\}$, and so $F_X(s) \le F_X(t)$. So F_X is non-decreasing.
- $\lim_{t \to -\infty} F_X(t) = \mathbb{P}(\varnothing) = 0, \quad \lim_{t \to +\infty} F_X(t) = \mathbb{P}(\Omega) = 1.$
- F_X is right-continuous:

$$\begin{aligned} \lim_{n \to +\infty} F_X(t + 1/n) &= \lim_{n \to +\infty} \mathbb{P}(X \le t + 1/n) \\ &= \mathbb{P}(\cap_{n \ge 1} \{X \le t + 1/n\}) \qquad \text{(by Theorem 1.7)} \\ &= \mathbb{P}(X \le t) = F_X(t). \end{aligned}$$

Example : Fair coin cont'd



▷ Examples of laws: discrete random variables

We say that X is *discrete* if X takes its values in a finite or countable set.

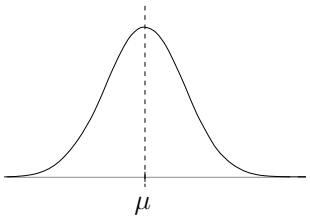
- Bernoulli distribution with parameter $p \in [0,1]$:
$$\mathbb{P}(X=1) = p, \qquad \mathbb{P}(X=0) = 1-p.$$
- Binomial distribution with parameters $n \geq 1, p \in [0,1]$ = number of successes in n Bernoulli trials:
$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n.$$
- Geometric distribution with parameter $p \in [0,1]$ = first success in Bernoulli trials:
$$\mathbb{P}(X=k) = (1-p)^{k-1} p \text{ for } k = 1, 2, \ldots$$
- Poisson distribution with parameter $\lambda > 0$:
$$\mathbb{P}(X=k) = e^{-\lambda} \frac{\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

▷ Examples of laws: continuous random variables

We say that X is *continuous* if there is a non-negative function f such that
$$\mathbb{P}(X \in A) = \int_A f(x)dx.$$
The function f is the *density* of X. Of course $\int_{\mathbb{R}} f(x)dx = \mathbb{P}(X \in \mathbb{R}) = 1$.

- Uniform distribution on $[a,b]$:
$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$
- Exponential distribution $\mathcal{E}(\lambda)$ with parameter $\lambda > 0$:
$$f(x) = \lambda \exp(-\lambda x) \mathbb{1}_{x \geq 0}.$$
- Normal distribution (or gaussian distribution) with parameters $\mu \in \mathbb{R}$, $\sigma^2 > 0$:

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



Remark : There exist random variables which are neither discrete nor continuous!
For instance $X = \max\{1, Y\}$ where $Y \sim \mathcal{E}(1)$.

▷ cumulative distribution functions and densities

If X is continuous, then $F_X(t) = \int_{-\infty}^{t} f(x)dx$ and thus

$$F'_X(t) = f(t).$$

Example : Let X follow the uniform distribution in $[0,1]$. What is the density (if any) of X^2 ? To answer this we compute the cdf of X^2. The map $x \mapsto \sqrt{x}$ is increasing and maps $[0,1]$ onto itself so for each $t \in (0,1)$,

$$F(t) = \mathbb{P}(X^2 \leq t) = \mathbb{P}(X \leq \sqrt{t}) = \int_0^{\sqrt{t}} dx = \sqrt{t}.$$

We now differentiate in order to get the density of X^2:

$$F'(t) = \frac{1}{2\sqrt{t}} \mathbb{1}_{[0,1]}(t).$$

▷ Simulations of continuous random variables

In many programming languages `rand()` returns a uniform number in the interval $(0,1)$. Imagine that we want to sample a random variable X with cdf F, it is very easy if F is one-to-one:

Algorithm
U=`rand()`
`return` $F^{-1}(U)$.

We easily check that the cdf of $F^{-1}(U)$ is indeed F: F is increasing so

$$\mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t)) = \int_0^{F(t)} dx \text{ (since } U \text{ is uniform)} = F(t).$$

If F is not injective, we must use the *pseudo-inverse* of F defined by

$$F^{-1}(u) = \inf \{x \text{ such that } F(x) \geq u\}.$$

Example : (Simulation of an exponential random variable) For the case of $X \sim \mathcal{E}(1)$, we have $F(t) = \int_0^t e^{-u}du = 1 - e^{-t}$. In order to find F^{-1} we have to solve

$$F(t) = x \Leftrightarrow 1 - e^{-t} = x \Leftrightarrow 1 - x = e^{-t} \Leftrightarrow t = -\log(1-x).$$

This proves that $F^{-1}(x) = -\log(1-x)$. Thus the command `-log(1-`U`)` returns an exponential random variable.

2.2 Abstract expectation

Let X be a **non-negative** random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition/Theorem 2.3 (*Expectation of a non-negative random variable*)

The expectation $\mathbb{E}[X]$ *of* X *is a real number in* $[0, +\infty) \cup \{+\infty\}$ *with properties*

- *If* $X = \mathbb{1}_A$ *then*
$$\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A).$$
- **(Linearity)** *For any real numbers* a, b *and any random variables* X, Y
$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$
- **(Monotonicity)** *If* $X(\omega) \leq Y(\omega)$ *for every* ω*, then* $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
- **(Monotone convergence)** *If for each* $\omega \in \Omega$, $X_n(\omega) \nearrow X(\omega)$ *then* X *is measurable and*
$$\mathbb{E}[X] = \mathbb{E}[\lim X_n] = \lim \mathbb{E}[X_n].$$

Remark : From a more theoretical point of view, the expectation $\mathbb{E}[X]$ is constructed as the integral of function $\omega \mapsto X(\omega)$ with respect to the measure $\mathbb{P}$. This is the theory of *Lebesgue integration* (or *abstract integration*) and goes much beyond the scope of this course. Yet, it explains why you also can find the notation
$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega),$$
where $\int_\Omega$ is an abstract integral (recall that Ω is not an interval of $\mathbb{R}$!) and "$d\mathbb{P}$" reads "with respect to $\mathbb{P}$".

▷ Expectation of a discrete random variable

Let Ω be some finite or countable space $\{\omega_1, \omega_2, \dots\}$ and X a non-negative random variable defined on Ω. Then we can write
$$X(\omega) = X(\omega_1) \times \mathbb{1}_{\omega=\omega_1} + X(\omega_1) \times \mathbb{1}_{\omega=\omega_2} + \dots$$
and thus
$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X(\omega_1) \times \mathbb{1}_{\omega_1}] + \mathbb{E}[X(\omega_2) \times \mathbb{1}_{\omega_2}] + \dots \\
&= X(\omega_1) \times \mathbb{E}[\mathbb{1}_{\omega_1}] + X(\omega_2) \times \mathbb{E}[\mathbb{1}_{\omega_2}] + \dots \\
&= X(\omega_1) \times \mathbb{P}(\omega_1) + X(\omega_2) \times \mathbb{P}(\omega_2) + \dots \\
&= \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(X = \omega),
\end{aligned}$$
and this coincides with the usual notion of expectation (or average).

▷ Expectation of a continuous random variable

Assume that X has density f on $\mathbb{R}_+$. We won't get into mathematical details but we always can approach such an X by below with a sequence of discrete random variables: for all n

$$X \approx \sum_{i\geq 0} \frac{i}{n} \mathbb{1}_{\frac{i}{n}\leq X\leq \frac{i+1}{n}}$$
$$\mathbb{E}[X] \approx \sum_{i\geq 0} \frac{i}{n} \mathbb{P}\left(\tfrac{i}{n} \leq X \leq \tfrac{i+1}{n}\right)$$

By letting n go to infinity the right-hand side gets closer and closer to the area below the curve of $xf(x)$. If f is integrable in the Riemann sense (the usual integral that you already know), we get

$$\mathbb{E}[X] = \int_{\mathbb{R}_+} xf(x)dx.$$

▷ Expectations of random variables of any sign

Now, how to define $\mathbb{E}[X]$ when X has arbitrary sign? Set

$$X^+ = \begin{cases} X & \text{if } X \geq 0, \\ 0 & \text{otherwise.} \end{cases} \qquad X^- = \begin{cases} 0 & \text{if } X \geq 0, \\ -X & \text{otherwise.} \end{cases}$$

Then we can write $X = X^+ - X^-$ (note also that $|X| = X^+ + X^-$).

Definition/Theorem 2.4 (*Expectation of a random variable of any sign*)

Let $X : \Omega \to \mathbb{R}$ be a random variable such that $\mathbb{E}[X^+] < +\infty$ and $\mathbb{E}[X^-] < +\infty$ (those expectations are well-defined since X^+, X^- are non-negative random variables).
We say that X is integrable *and we set*

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

Remark : Recall that $|X| = X$ or $-X$, and note that by monotonicity

$$\begin{cases} X \leq |X| \\ -X \leq |X| \end{cases} \quad \Rightarrow \quad \begin{cases} \mathbb{E}[X] \leq \mathbb{E}[|X|] \\ -\mathbb{E}[X] \leq \mathbb{E}[|X|] \end{cases}.$$

But now $|\mathbb{E}[X]| = \mathbb{E}[X]$ or $-\mathbb{E}[X]$ so

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

Here is the practical formula that tells how to compute expectations.

Proposition 2.5 (*Expectation of a function of X*)

Let X be a random variable and ϕ a Borel function such that $\mathbb{E}[|\phi(X)|] < +\infty$.

- *If X has density f, then $\mathbb{E}[\phi(X)] = \int_{\mathbb{R}} \phi(x) f(x) dx$.*
- *If X is discrete, then $\mathbb{E}[\phi(X)] = \sum_k \phi(k) \mathbb{P}(X = k)$, where the sums runs over all the possible values for X: $\mathbb{N}, \mathbb{Z}, \ldots$*

Example : Let X follow the uniform distribution in $[0, 1]$,

$$\mathbb{E}[X^n] = \int x^n \times \mathbb{1}_{[0,1]}(x) dx = \int_{[0,1]} x^n dx = \frac{x^{n+1}}{n+1}\Big|_{x=0}^{x=1} = \frac{1}{n+1}.$$

Exercise : Let $X \sim \mathcal{E}(1)$. Prove by induction that $\mathbb{E}[X^n] = n!$.

2.3 How to find the distribution of X?

To find the distribution of X, we have already seen that it is enough to prove that they have the same cdf. Here are two other useful criteria.

▷ 1st method: change of variable

The strategy is to use the following Theorem:

Theorem 2.6 (*Characterization with bounded and continuous ϕ*)

- *If $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(Y)]$ for every bounded and continuous function ϕ, then X and Y have the same distribution.*
- *In particular, if one can find a function f such that, for every bounded and continuous ϕ,*

$$\mathbb{E}[\phi(X)] = \int \phi(x) f(x) dx$$

then X has density f.

Example : Let us use this criterion to prove once again that if X is uniform in $[0, 1]$ then X^2 has density $1/2\sqrt{x}$. Let us compute

$$\mathbb{E}[\phi(X^2)] = \int \phi(x^2) \underbrace{\mathbb{1}_{[0,1]}(x)}_{\text{density of } X} dx = \int_0^1 \phi(x^2) dx.$$

We make the change of variables $t = x^2$, $x = \sqrt{t}$, $\frac{dx}{dt} = \frac{1}{2\sqrt{t}}$, this gives

$$\mathbb{E}[\phi(X^2)] = \int_{x=0}^{x=1} \phi(x^2) dx = \int_{t=0}^{t=1} \phi(t) \frac{1}{2\sqrt{t}} dt.$$

This holds for any bounded and continuous ϕ, thus X^2 has density $\frac{1}{2\sqrt{t}}$ on interval $[0, 1]$.

▷ 2d method: Characteristic functions

We introduce an important tool: the *characteristic function* (also called *Fourier transform*).

Definition/Theorem 2.7

The characteristic function *of a random variable X is the function $\Phi_X(t)$:*

$$\begin{array}{rccc} \Phi_X(t): & \mathbb{R} & \to & \mathbb{C} \\ & t & \mapsto & \mathbb{E}[e^{itX}]. \end{array}$$

If $\Phi_X(t) = \Phi_Y(t)$ for all t, then X and Y have the same law.

(Here i is the complex number $i^2 = -1$. All you need to know about the exponential of a complex number is that the power rule $e^{z+z'} = e^z e^{z'}$ also holds for complex numbers and that if r is a real number then $|e^{ir}| = 1$.)

Example : Let X have the exponential distribution with parameter 1. Then

$$\begin{aligned} \mathbb{E}[e^{itX}] &= \int_0^{+\infty} e^{itx} e^{-x} dx \\ &= \int_0^{+\infty} e^{x(it-1)} dx = \frac{e^{x(it-1)}}{it-1}\Big|_{x=0}^{x=+\infty} \\ &= \frac{1}{it-1}\left(\lim_{x\to+\infty} e^{x(it-1)} - 1\right) \\ &= \frac{1}{it-1}\left(\lim_{x\to+\infty} \underbrace{e^{xit}}_{\text{of modulus 1}} \times \underbrace{e^{-x}}_{\to 0} - 1\right) = \frac{1}{1-it}. \end{aligned}$$

Plainly from the definition, we have some interesting properties of the characteristic function.

Proposition 2.8

- $\Phi_X(0) = \mathbb{E}[e^0] = \mathbb{E}[1] = 1$.
- *For every t,* $|\Phi_X(t)| = |\mathbb{E}[e^{itX}]| \leq \mathbb{E}[|e^{itX}|] = 1$.
- *Assume that $\mathbb{E}[|X|] < +\infty$, then (using Theorem 2.20)*

$$\Phi'_X(t) = \frac{\partial}{\partial t}\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\frac{\partial}{\partial t} e^{itX}\right] = \mathbb{E}[iXe^{itX}].$$

In particular, $\Phi'_X(0) = i\mathbb{E}[X]$.

For a discrete random variable X, it is more usual to deal with the *generating function* $G_X(z)$ defined by

$$G_X(z) = \mathbb{E}[z^X] = \sum_{k\geq 0} \mathbb{P}(X = k) z^k.$$

As for characteristic functions, X, Y have the same law if $G_X(z) = G_Y(z)$ for any z.

2.4 Inequalities

We first recall the definition of variance. First note that if X such that $\mathbb{E}[X^2] < +\infty$ then, since $|X| \leq 1 + X^2$, we have $\mathbb{E}[|X|] < +\infty$, and then $\mathbb{E}[X]$ is well-defined.

Definition 2.9 (*Variance*)

Let X be a random variable such that $\mathbb{E}[X^2] < +\infty$. The variance *of X is defined by*

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

It tells how much X deviates from its mean.

Let us expand what's inside the expectation:

$$\begin{aligned}
\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\right] \\
&= \mathbb{E}[X^2] + \mathbb{E}\left[\mathbb{E}[X]^2\right] - 2\mathbb{E}\left[X\mathbb{E}[X]\right] \text{ (by linearity)} \\
&= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

Note also that, plainly from the definition, we have for all constants a, b

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

▷ Some useful inequalities

You already know that $\mathbb{E}[|X|] \geq |\mathbb{E}[X]|$. This is in fact a particular case of:

Theorem 2.10 (*Jensen's inequality*)

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function and X be an integrable r.v. Then

$$\mathbb{E}\left[\phi(X)\right] \geq \phi\left(\mathbb{E}[X]\right).$$

For instance we have $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2, \mathbb{E}[e^X] \geq e^{\mathbb{E}[X]}, \ldots$

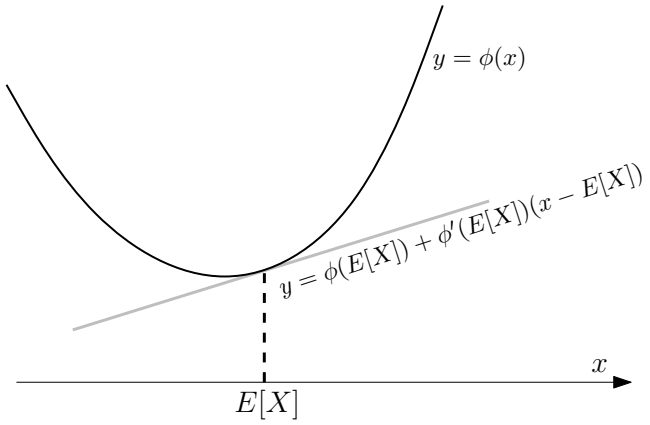
Proof:

For simplicity we assume that ϕ is differentiable. Consider its curve, by convexity the tangent line at some point a is below the curve: for all real x we have

$$\phi(x) \geq \phi(a) + \phi'(a)(x-a).$$

In particular, for $a = \mathbb{E}[X]$ this gives (see the picture on the right)

$$\phi(x) \geq \phi(\mathbb{E}[X]) + \phi'(\mathbb{E}[X])(x - \mathbb{E}[X]).$$



In particular, this inequality is true for the real number $x = X$. By taking expectation,

$$\begin{aligned}\mathbb{E}[\phi(X)] &\geq \mathbb{E}\Big[\phi(\mathbb{E}[X]) + \phi'(\mathbb{E}[X])(X - \mathbb{E}[X])\Big] \\ &\geq \mathbb{E}[\phi(\mathbb{E}[X])] + \phi'(\mathbb{E}[X])\mathbb{E}\left[X - \mathbb{E}[X]\right] = \phi(\mathbb{E}[X]) + 0,\end{aligned}$$

since $\phi(\mathbb{E}[X])$ is a constant, and $\mathbb{E}\left[X - \mathbb{E}[X]\right] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. ■

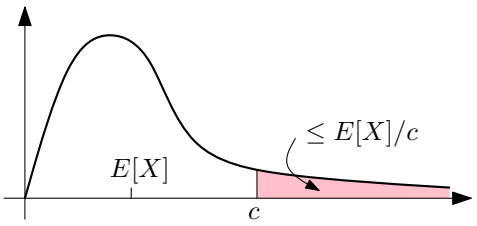
We now can state two important inequalities that estimate the probability that X deviates from its mean:

Theorem 2.11

Let X be an integrable random variable.

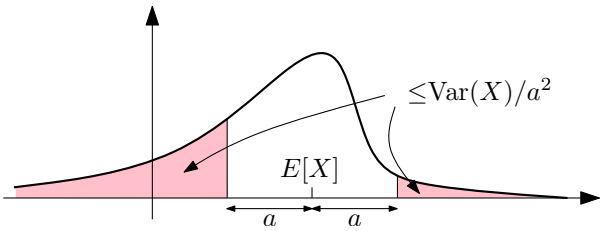
- **Markov's inequality.** *If X is non-negative and $c > 0$ is a constant,*

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



- **Chebyshev's inequality.** *If $\mathrm{Var}(X) < +\infty$ and $a > 0$ is a constant,*

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



Proof: Markov's inequality.

$$\begin{aligned} 1 &\geq \mathbb{1}_{X\geq c} \\ X &\geq X\mathbb{1}_{X\geq c} \qquad (\text{since } X \geq 0) \\ \mathbb{E}[X] &\geq \mathbb{E}\left[X\mathbb{1}_{X\geq c}\right] \geq c\mathbb{E}\left[\mathbb{1}_{X\geq c}\right], \end{aligned}$$

and now remember that $\mathbb{E}[\mathbb{1}_{X\geq c}]$ is just $\mathbb{P}(X \geq c)$.

Chebyshev's inequality. First note that $\{|X - \mathbb{E}[X]| \geq a\}$ and $\left\{(X - \mathbb{E}[X])^2 \geq a^2\right\}$ both denote the same event, so that

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq a\right) = \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq a^2\right).$$

Now, $(X - \mathbb{E}[X])^2$ is a non-negative random variable. So by Markov's inequality

$$\mathbb{P}\left((X - \mathbb{E}[X])^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

■

2.5 L^p spaces

Definition 2.12

Let $p \geq 1$ be a real number, we denote by $L^p(\Omega, \mathcal{A}, \mathbb{P})$ (or just L^p if there is no ambiguity) the set of random variables X such that $\mathbb{E}[|X|^p] < +\infty$. In this case, we define the L^p norm of X as

$$\|X\|_p = \mathbb{E}\left[|X|^p\right]^{1/p}.$$

Note that in the definition, p is any real number in $[1, +\infty)$ but in practice we often consider integer values of p: $L^1, L^2, \ldots$

Example:

- If $X \leq c$, then $\mathbb{E}[|X|^p] \leq c^p < +\infty$. Then bounded r.v. are in all L^p's.
- We saw that if X follows the exponential distribution then $\mathbb{E}[X^p] = p! < +\infty$. Then X also belongs to all L^p's.

L^p is a *vector space*, meaning that if $X, Y \in L^p$ and $a \in \mathbb{R}$, then

- aX is in L^p,
- $X + Y$ is in L^p.

It is not so easy to check that $\mathbb{E}[|X + Y|^p]$ is finite. To prove so, let us observe that for any real numbers x, y,

$$|x + y|^p \leq |2\max\{x, y\}|^p \leq 2^p(|x|^p + |y|^p),$$

and then, applying this to X, Y and taking expectations gives

$$\mathbb{E}[|X + Y|^p] \leq 2^p(\mathbb{E}[|X|^p] + \mathbb{E}[|Y|^p]),$$

which is finite. We admit in these notes that $X \mapsto \|X\|_p$ is indeed a norm, that is to say:

- $\|aX\|_p = |a|\ \|X\|_p$,
- $\|X\|_p = 0$ if and only if $X = 0$ almost surely,
- **Triangle inequality:** $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$.

A very important property of L^p spaces is that there are included into each others.

Theorem 2.13

For $q > p$ we have $L^q \subset L^p$:

$$\dots L^p \subset L^{p-1} \subset L^{p-2} \subset \cdots \subset L^2 \subset L^1.$$

Proof: We will prove that for $p < q$ we have $\|X\|_p \leq \|X\|_q$. Then, if $\|X\|_q$ is finite, so is $\|X\|_p$. The trick is to write $|X|^q = (|X|^p)^{q/p}$. Since $q/p > 1$ the map $x \mapsto x^{q/p}$ is convex on $\mathbb{R}_+$ and thus, if we apply Jensen's inequality to the r.v. $|X|^p$ we obtain

$$\begin{aligned} & & \mathbb{E}\left[(|X|^p)^{q/p}\right] &\geq (\mathbb{E}\left[|X|^p\right])^{q/p} \\ \text{to the power } 1/q: & & \mathbb{E}\left[|X|^q\right]^{1/q} &\geq (\mathbb{E}\left[|X|^p\right])^{1/p} \\ & & \|X\|_q &\geq \|X\|_p . \end{aligned}$$

■

Definition 2.14 (*L^p convergence*)

Let $(X_n)_{n\geq 0}$ be a sequence of random variables, one says that X_n converges to X in L^p *if $\mathbb{E}\left[\ |X_n - X|^p\ \right] \overset{n\to\infty}{\to} 0$. One writes $X_n \overset{L^p}{\to} X$.*

Of course this amounts to say that $\|X_n - X\|_p$ goes to zero. Let us note that if $X_n \overset{L^q}{\to} X$ for some q, then $X_n \overset{L^p}{\to} X$ for every $p < q$, since we have

$$\|X_n - X\|_p \leq \underbrace{\|X_n - X\|_q}_{\to 0}.$$

2.6 The special case of L^2

Let X, Y be in some $L^2(\Omega, \mathcal{A}, \mathbb{P})$, it turns out that XY is integrable, due to the following:

Theorem 2.15 (*Cauchy-Schwarz's inequality*)

If X, Y are in L^2, then

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2}\mathbb{E}[Y^2]^{1/2}.$$

In particular, since the right-hand side is finite, $X \times Y \in L^1$.

The first inequality is not new, the second one has a nice proof:

Proof (sketch of): Take some real number t, we obviously have

$$\begin{aligned} 0 &\leq \mathbb{E}[(t|X| + |Y|)^2] \\ &= t^2\mathbb{E}[|X|^2] + 2t\mathbb{E}[|XY|] + \mathbb{E}[|Y|^2] =: P(t). \end{aligned}$$

If we see t as the variable, then $P(t)$ is a polynomial of order two whose sign does not change. Thus its discriminant is non-positive:

$$4\mathbb{E}[|XY|]^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0,$$

which can be rewritten as $\mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2}\mathbb{E}[Y^2]^{1/2}$. ■

Since $\mathbb{E}[XY]$ is finite , we can define the *covariance* of X and Y by

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

(note that $\mathrm{Cov}(X,X)$ is just $\mathrm{Var}(X)$).

▷ Scalar product in L^2

Set $\langle X, Y\rangle = \mathbb{E}[XY]$, it is often convenient to see the application

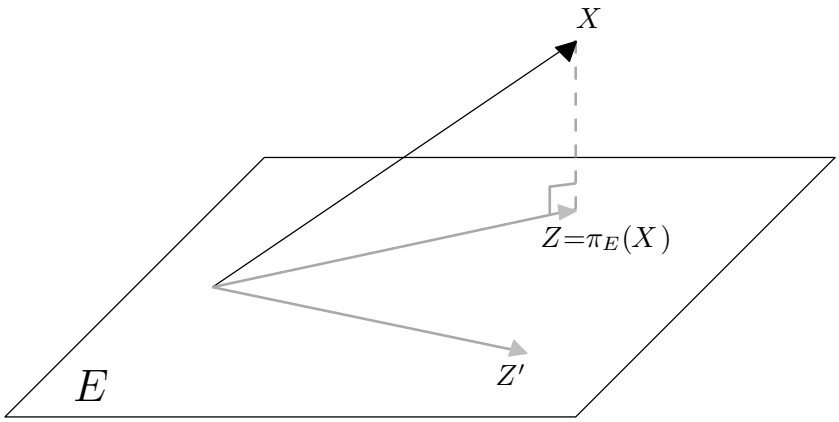
$$\begin{array}{ccc} L^2 \times L^2 & \to & \mathbb{R} \\ (X,Y) & \mapsto & \langle X, Y\rangle \end{array}$$

as a *scalar product* (one also says *inner product*) which means that the following properties hold:

- **Symmetry:** $\langle X, Y\rangle = \langle Y, X\rangle$.
- **Linearity:** $\langle aX + bX', Y\rangle = a\langle X, Y\rangle + b\langle X', Y\rangle$.
- **Positive-definiteness:** For every X, we have $\langle X, X\rangle \geq 0$. Moreover, $\langle X, X\rangle = 0$ if and only if $X = 0$ a.s.

By analogy with the usual scalar product in geometry, we say that X, Y are **orthogonal** if $\langle X, Y\rangle = 0$. This analogy is very useful, especially when we will define conditional expectation.

Let $X \in L^2(\Omega)$ and let E be a subspace of $L^2(\Omega)$.



Legend: By definition, the orthogonal projection $\pi_E(X)$ of X is the only random variable $\pi_E(X) \in E$ such that $(X - \pi_E(X)) \perp Z'$ for every element $Z' \in E$.

It turns out that the orthogonal projection $\pi_E(X)$ of X onto E can be characterized as a minimization problem:

Theorem 2.16 (*Projection as a minimization problem*)

Let $X \in L^2$, E be a vector subspace of L^2. Then

$$Z = \pi_E(X) \Leftrightarrow \begin{cases} Z \in E \\ \mathbb{E}[(X-Z)^2] = \min_{Z' \in E} \mathbb{E}[(X-Z')^2] \end{cases}$$

Example : Let $X \in L^2$. What is the orthogonal projection $\pi_E(X)$ of X onto the subspace $E =$ { constant random variables $a, a \in \mathbb{R}$}?

1st method: using orthogonality. First, we know that $\pi_E(X) \in E$, *i.e.* can be written as a constant a. Then

$$X - \pi_E(X) = X - a \perp \mathbf{1}$$

(where $\mathbf{1} \in E$ is the constant r.v. equal to one). This means that

$$0 = \mathbb{E}[(X-a) \times 1] = \mathbb{E}[X] - a,$$

i.e. $a = \mathbb{E}[X]$. Finally,

$$\pi_E(X) = \mathbb{E}[X]$$

2d method: using minimization. Using Theorem 2.16, we know that $a = \pi_E(X)$ is the solution of

$$\mathbb{E}[(X-a)^2] = \min_{b \in \mathbb{R}} \mathbb{E}[(X-b)^2].$$

Let us minimize $f(b) := \mathbb{E}[(X-b)^2]$. We have

$$\begin{aligned} f(b) &= \mathbb{E}[X^2] + \mathbb{E}[b^2] + \mathbb{E}[-2bX] = \mathbb{E}[X^2] + b^2 - 2b\mathbb{E}[X] \\ f'(b) &= 0 + 2b - 2\mathbb{E}[X]. \end{aligned}$$

This proves that f is minimal for $b = \mathbb{E}[X]$. Therefore

$$\mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{b \in \mathbb{R}} \mathbb{E}[(X-b)^2].$$

ans this proves (again) that $\pi_E(X) = \mathbb{E}[X]$.

2.7 Bonus: Swapping $\mathbb{E}$ and limit

Our main concern here is:

When can we swap expectation and limit: when is $\mathbb{E}[\lim_n X_n]$ equal to $\lim_n \mathbb{E}[X_n]$?

In the definition of the expectation we already saw the

Theorem 2.17 (*Monotone convergence*)

Let (X_n) be a sequence of non-negative measurable random variables. Assume that (X_n) is $\nearrow$: for all ω, $(X_n(\omega))_{n \geq 1}$ is non-decreasing. Then $X = \lim_n X_n$ is measurable and

$$\mathbb{E}[X] = \mathbb{E}[\lim_{n \to +\infty} X_n] = \lim_{n \to +\infty} \mathbb{E}[X_n].$$

Here is an important application of monotone convergence:

Proposition 2.18 (*Swapping $\sum$ and $\mathbb{E}$*)

Let (X_k) be a sequence of non-negative random variables. Then $\sum_{k=0}^{n} X_k \nearrow \sum_{k=0}^{\infty} X_k$ and then

$$\mathbb{E}\left[\sum_{k=0}^{\infty} X_k\right] = \sum_{k=0}^{\infty} \mathbb{E}[X_k]$$

(both sides can be infinite).

If we want to deal with arbitrary sequences, we need another assumption: *domination*. Before going into details, let us begin by a definition. We say that $(X_n)_{n\geq 1}$ converges to X *almost surely* if

$$\mathbb{P}\left(X_n \overset{n\to\infty}{\longrightarrow} X\right) = \mathbb{P}\left(\omega \text{ such that } X_n(\omega) \overset{n\to\infty}{\longrightarrow} X(\omega)\right) = 1.$$

Theorem 2.19 (*Dominated-convergence Theorem*)

Assume that (X_n) converges to X almost surely. Assume also that all X_n's are dominated *by Y: for all $n \geq 1$ and $\omega \in \Omega$,*

$$|X_n(\omega)| \leq |Y(\omega)|$$

where Y is integrable: $\mathbb{E}[|Y|] < +\infty$. Then

$$\mathbb{E}[\lim_{n\to+\infty} X_n] = \lim_{n\to+\infty} \mathbb{E}[X_n].$$

Remark : If fact, with these assumptions, we even have a stronger result:

$$\mathbb{E}[|X_n - X|] \to 0.$$

We now give an useful result which turns to be an application of the dominated-convergence Theorem.

Theorem 2.20 (*Differentiating inside expectations*)

Let I be an interval, and

$$\begin{array}{rccc} f: & I \times \mathbb{R} & \to & \mathbb{R} \\ & (t, X) & \mapsto & f(t, X). \end{array}$$

Assume that for every $t \in I$, the random variable $x \mapsto f(t, X)$ is integrable and that $\left|\frac{\partial}{\partial t} f(t, X)\right| \leq g(X)$ with $\mathbb{E}[g(X)] < +\infty$, then

$$\frac{\partial}{\partial t}\mathbb{E}[f(t, X)] = \mathbb{E}\left[\frac{\partial}{\partial t} f(t, X)\right].$$

3 Random vectors

The objective of the present Chapter is to introduce some tools to study the distribution of a random vector $(X_1, \ldots, X_d) \in \mathbb{R}^d$. The **joint** distribution of $(X_1, \ldots, X_d)$, denoted by $\mathbb{P}_{(X_1,\ldots,X_d)}$ is the measure on $\mathbb{R}^d$ defined by

$$\mathbb{P}_{(X_1,\ldots,X_d)}(A) = \mathbb{P}\left((X_1, \ldots, X_d) \in A\right).$$

In what follows, in order to lighten notations, we only write formulas for the case $d = 2$ but all the results are valid in the general d-dimensional case.

3.1 Joint densities and the Fubini Theorems

We first need a couple of Theorems in order to properly define and handle multiple integrals.

Theorem 3.1 (*The first Fubini Theorem*)

Let f be a **non-negative** *measurable function*

$$\begin{array}{rccc} f: & \mathbb{R}\times\mathbb{R} & \to & \mathbb{R}_+ \\ & (x,y) & \mapsto & f(x,y). \end{array}$$

Then

$$\int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} f(x,y)dx \right) dy = \int_{x\in\mathbb{R}} \left(\int_{y\in\mathbb{R}} f(x,y)dy \right) dx$$

(Note that both sides might be equal to $+\infty$.)
Thus we can without ambiguity denote this quantity by $\iint_{\mathbb{R}^2} f(x,y)dxdy$.

Theorem 3.2 (*The second Fubini Theorem*)

Let f be a measurable function **of any sign**

$$\begin{array}{rccc} f: & \mathbb{R}\times\mathbb{R} & \to & \mathbb{R} \\ & (x,y) & \mapsto & f(x,y). \end{array}$$

Assume that $\iint |f|dxdy$ is finite (you check this with the first Fubini Theorem), then

$$\int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} f(x,y)dx \right) dy = \int_{x\in\mathbb{R}} \left(\int_{y\in\mathbb{R}} f(x,y)dy \right) dx$$

(In this case both sides are finite.)
We denote this quantity by $\iint_{\mathbb{R}^2} f(x,y)dxdy$.

Definition 3.3 (*Joint density*)

We say that (X, Y) has joint density $f : \mathbb{R}^2 \to \mathbb{R}_+$ if for any set $A \subset \mathbb{R}^2$

$$\mathbb{P}\left((X,Y) \in A\right) = \iint_A f(x,y)dxdy.$$

In this case, for every measurable function ϕ,

$$\mathbb{E}\left[\phi(X,Y)\right] = \iint_{\mathbb{R}^2} \phi(x,y) f(x,y) dxdy. \qquad (*)$$

(when this quantity is well-defined: either if $\phi \geq 0$ or if $\iint |\phi| f < +\infty$).

Example : Function $(x,y) \mapsto (x+y)\mathbb{1}_{[0,1]\times[0,1]}(x,y)$ is a density. It is clearly non-negative and by first Fubini's Theorem

$$\begin{aligned}
\iint_{[0,1]\times[0,1]} (x+y)dxdy &= \int_{x\in[0,1]} \left(\int_{y\in[0,1]} (x+y)dy \right) dx \\
&= \int_{x\in[0,1]} \left(\int_{y\in[0,1]} xdy + \int_{y\in[0,1]} ydy \right) dx \\
&= \int_{x\in[0,1]} (x+1/2)dx \\
&= 1/2 + 1/2 = 1.
\end{aligned}$$

▷ Marginal densities

Let (X,Y) have density f and let us compute the density of X: for any bounded and continuous function ϕ of a single variable x, we can apply formula $(*)$ just above:

$$\mathbb{E}\left[\phi(X)\right] = \iint \phi(x) f(x,y) dxdy.$$

Function $|\phi|$ is bounded by some $c > 0$,

$$\iint |\phi(x) f(x,y)| dxdy \leq \iint c f(x,y) dxdy = c \iint f(x,y) dxdy = c \times 1 < +\infty$$

so we can apply the second Fubini Theorem and integrate first with respect to y:

$$\mathbb{E}\left[\phi(X)\right] = \int_x \phi(x) \underbrace{\left(\int_y f(x,y) dy \right)}_{\text{density of } X} dx.$$

Proposition 3.4 (*Marginal densities*)

If (X,Y) has density $(x,y) \mapsto f(x,y)$ then X has density $x \mapsto \displaystyle\int_{y\in\mathbb{R}} f(x,y)dy$. It is called the marginal density *of X. Similarly, Y has density $y \mapsto \displaystyle\int_{x\in\mathbb{R}} f(x,y)dx$.*

Example (continuation) : Let (X,Y) have density $(x+y)\mathbb{1}_{[0,1]\times[0,1]}(x,y)$. By the proposition the density of X is

$$x \mapsto \begin{cases} \int_{y\in[0,1]} (x+y)dy = x + 1/2 \text{ if } x \in [0,1], \\ 0 \text{ otherwise.} \end{cases}$$

Remark : It might happen that X, Y both have densities while pair (X, Y) has not. Consider the case where $X \sim \mathcal{E}(1)$, and $Y = 2X$. Then the pair $(X, Y) = (X, 2X)$ lies on the line $D = \{y = 2x\}$ with probability one:

$$\mathbb{P}((X, Y) \in D) = 1.$$

But if (X, Y) had a density f then this probability would be zero:

$$\begin{aligned}\mathbb{P}((X, Y) \in D) = \mathbb{E}[\mathbb{1}_{Y=2X}] &= \iint \mathbb{1}_{y=2x} f(x, y) dxdy \\ &= \int_x \left(\int_y \mathbb{1}_{y=2x} f(x, y) dy \right) dx \\ &= \int_x \left(\int_{y=2x}^{2x} f(x, y) dy \right) dx \\ &= \int_x 0 \times dy = 0.\end{aligned}$$

To conclude this section, note that there is a criterion analogous to Theorem 2.6:

Theorem 3.5 (*Characterization with bounded and continuous ϕ*)

If one can find a function $f : \mathbb{R}^2 \to \mathbb{R}_+$ such that, for every bounded and continuous function $\phi : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}\left[\phi(X, Y)\right] = \iint \phi(x, y) f(x, y) dxdy.$$

then (X, Y) has joint density f.

3.2 Bivariate change of variables

▷ A simple example

Let (X, Y) have joint density $f(x, y)$, imagine that we want to compute the joint density of $(u(X, Y), v(X, Y))$ for some simple functions u, v. To do this we have to compute

$$\mathbb{E}\left[\phi\left(u(X, Y), v(X, Y)\right)\right] = \iint_{\mathbb{R}^2} \phi\left(u(x, y), v(x, y)\right) f(x, y) dxdy$$

for any bounded and continuous $\phi : \mathbb{R}^2 \to \mathbb{R}$.
The general method to do that is to use a bivariate change of variables, we won't make the detailed theory but rather work out an example.

Example : Let (X, Y) have joint density

$$f(x, y) = \tfrac{3}{4} \exp\left(-|x + 2y| - |x - y|\right)$$

what is the joint density of $(X + 2Y, X - Y)$? We set $U = X + 2Y$ and $V = X - Y$,

$$\mathbb{E}\left[\phi(U, V)\right] = \iint_{\mathbb{R}^2} \phi\left(x + 2y, x - y\right) \tfrac{3}{4} \exp\left(-|x + 2y| - |x - y|\right) dxdy.$$

In the latter integral we make the change of variables

$$\begin{cases} u = x + 2y, \\ v = x - y. \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = (u + 2v)/3, \\ y = (u - v)/3. \end{cases}$$

If (x, y) runs in all $\mathbb{R}^2$ then so does (u, v) so the integration domain is still $\mathbb{R}^2$. We have to make the change $dxdy \leftrightarrow dudv$, we need to compute the so-called *Jacobian matrix*

$$\mathrm{Jac}(x, y) = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial u}\left(\frac{u+2v}{3}\right) & \frac{\partial}{\partial v}\left(\frac{u+2v}{3}\right) \\ \frac{\partial}{\partial u}\left(\frac{u-v}{3}\right) & \frac{\partial}{\partial v}\left(\frac{u-v}{3}\right) \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & -1/3 \end{pmatrix}.$$

Now the formula is (don't forget absolute values)

$$\frac{dxdy}{dudv} = |\det(\mathrm{Jac}(x, y))| = \left|\frac{1}{3} \times (-\frac{1}{3}) - \frac{1}{3} \times \frac{2}{3}\right| = |-3/9| = 1/3.$$

We get

$$\mathbb{E}\left[\phi(U, V)\right] = \iint_{\mathbb{R}^2} \phi\left(u, v\right) \tfrac{3}{4} e^{-|u|-|v|} \frac{dudv}{3} = \iint_{\mathbb{R}^2} \phi\left(u, v\right) \tfrac{1}{4} e^{-|u|-|v|} dudv,$$

which proves, using Theorem 3.5, that $(U, V) = (X + 2Y, X - Y)$ has density $(u, v) \mapsto \frac{1}{4} e^{-|u|-|v|}$.

Remark : 1. It is not necessary to check that $\frac{1}{4} e^{-|u|-|v|}$ is indeed a density on $\mathbb{R}^2$, this is always the case if there is no miscomputation in the change of variables.

2. One can also compute $\frac{dudv}{dxdy}$ in the reverse way. Just interchange $(u, v) \leftrightarrow (x, y)$:

$$\frac{dudv}{dxdy} = |\det(\mathrm{Jac}(u, v))| = \left|\det \begin{pmatrix} \frac{\partial}{\partial x}(x + 2y) & \frac{\partial}{\partial y}(x + 2y) \\ \frac{\partial}{\partial x}(x - y) & \frac{\partial}{\partial y}(x - y) \end{pmatrix}\right| = \left|\det \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}.\right| = |-3| = 3,$$

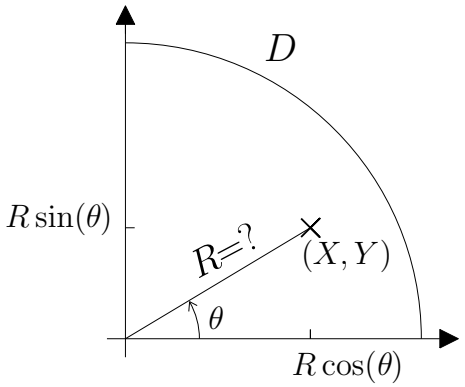
which is consistent with the previous computation.

▷ Another example: polar coordinates

Let (X, Y) be uniform in the quarter disc $D = \left\{(x, y) \in \mathbb{R}^2_+, \sqrt{x^2 + y^2} \leq 1\right\}$. In other words, (X, Y) has density

$$\frac{1}{\pi/4} \mathbb{1}_D(x, y),$$

(recall $\mathrm{Vol}(D) = \pi/4$). Let $R = \sqrt{X^2 + Y^2}$ be the distance from (X, Y) to origin, what is the distribution of R?



We have to compute

$$\mathbb{E}\left[\phi\left(R\right)\right)] = \iint_{\mathbb{R}^2} \phi\left(\sqrt{x^2+y^2}\right) \frac{1}{\pi/4} \mathbb{1}_D(x,y) dxdy.$$

for every bounded and continuous $\phi : \mathbb{R} \to \mathbb{R}$. We have to put $r = \sqrt{x^2+y^2}$, therefore we make the *polar* change of variables:

$$\begin{cases} r\cos(\theta) = x, \\ r\sin(\theta) = y. \end{cases}$$

This way we get

$$\sqrt{x^2+y^2} = \sqrt{r^2\cos^2(\theta) + r^2\sin^2(\theta)} = \sqrt{r^2 \times 1} = r.$$

Now we have (check the picture!)

$$(x,y) \in D \Leftrightarrow \begin{cases} r \leq 1, \\ 0 \leq \theta \leq \pi/2. \end{cases}$$

We have to make the change $dxdy \leftrightarrow drd\theta$, here is the Jacobian matrix:

$$\mathrm{Jac}(x,y) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial r} r\cos(\theta) & \frac{\partial}{\partial \theta} r\cos(\theta) \\ \frac{\partial}{\partial r} r\sin(\theta) & \frac{\partial}{\partial \theta} r\sin(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}.$$

We get

$$\frac{dxdy}{drd\theta} = \left|\det(\mathrm{Jac}(x,y))\right| = \left|r\cos^2(\theta) + r\sin^2(\theta)\right| = r.$$

Finally

$$\begin{aligned} \mathbb{E}\left[\phi(R)\right] &= \int_{r=0}^{1} \left(\int_{\theta=0}^{\pi/2} \phi(r) \frac{r}{\pi/4} d\theta \right) dr \\ &= \int_{r=0}^{1} \phi(r) r \left(\int_{\theta=0}^{\pi/2} \frac{1}{\pi/4} d\theta \right) dr \\ &= \int_{r=0}^{1} \phi(r) 2r dr. \end{aligned}$$

This proves, using Theorem 2.6, that R has density $2r$ on interval $[0,1]$.

3.3 Independence

Definition 3.6 (*Independence of random variables*)

Let $X_1, X_2, \ldots, X_n$ be random variables on the same space $(\Omega, \mathcal{A}, \mathbb{P})$. We say that $X_1, \ldots, X_n$ are independent *if for any Borel sets $B_1, \ldots, B_n$ we have*

$$\mathbb{P}\left(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i).$$

We say that a sequence of random variables $X_1, X_2, \ldots$ is *i.i.d.* (independent and identically distributed) if, for every n, $X_1, \ldots, X_n$ are independent and if X_i's all have the same law.

Remark : If X, Y are discrete random variables, then X and Y are independent if and only if for every i, j

$$\mathbb{P}(X = i \cap Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j).$$

Independence of events is more subtle. For $A_1, A_2, \ldots, A_n$ to be independent, we have to check independence of every sub-family of A_i's:

Definition 3.7 (*Independence of events*)

Events $A_1, \ldots, A_n$ are independent *if for every $k \leq n$ and every $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ we have*

$$\mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) = \mathbb{P}(A_{i_1}) \times \mathbb{P}(A_{i_2}) \times \cdots \times \mathbb{P}(A_{i_k}).$$

For instance, for $n = 3$ we have

$$A_1, A_2, A_3 \text{ independent } \Leftrightarrow \begin{cases} \mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) \\ \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2) \\ \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_3) \\ \mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2)\mathbb{P}(A_3). \end{cases}$$

If X, Y are independent then by definition we have

$$\mathbb{P}(X \leq s, Y \leq t) = \mathbb{P}(X \leq s)\mathbb{P}(Y \leq t) = F_X(s)F_Y(t).$$

More generally,

Proposition 3.8 (*Product of independent r.v.*)

If $X_1, X_2, \ldots, X_n$ are independent and $\phi_1, \ldots, \phi_n$ are Borel functions, then

- *$\phi_1(X_1), \phi_2(X_2), \ldots, \phi_n(X_n)$ are independent.*
- *If all $\phi_k(X_k)$'s are integrable we have*

$$\mathbb{E}\left[\phi_1(X_1)\phi_2(X_2)\ldots\phi_n(X_n)\right] = \mathbb{E}[\phi_1(X_1)]\mathbb{E}[\phi_2(X_2)]\ldots\mathbb{E}[\phi_n(X_n)].$$

An interesting consequence is that if X and Y are integrable and independent then

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

But the converse is not true: there exist random variables X, Y that are **not** independent for which $\mathrm{Cov}(X,Y) = 0$.

▷ Independence and densities

Theorem 3.9

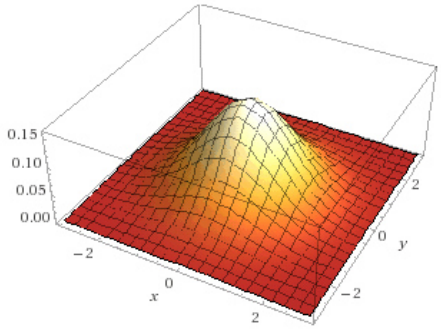
Let X, Y be two random variables.

- *If X, Y have densities f_X and f_Y, and if X, Y are independent, then*

$$(X,Y) \text{ has density } f_{(X,Y)}(x,y) = f_X(x) f_Y(y).$$

- *Conversely, if (X,Y) has a density which can be written as a product $g_1(x) \times g_2(y)$ then X, Y are independent.*

the density of (X,Y) where X, Y are independent $\mathcal{N}(0,1)$:



Example : Assume (X,Y) has density $6x^2y\mathbb{1}_{(x,y)\in[0,1]^2}$ (exercise: check that this is a density). Then the theorem says that X, Y are independent since one can write

$$6x^2y\mathbb{1}_{[0,1]\times[0,1]}(x,y) = 6x^2\mathbb{1}_{x\in[0,1]} \times y\mathbb{1}_{y\in[0,1]}.$$

Though, we have to care about constants if we want marginal densities: the density of X is

$$\int_{y=0}^{1} 6x^2 y dy = 6x^2 \int_{y=0}^{1} y dy = 3x^2 \text{ for } x \in [0,1].$$

Finally,

$$6x^2y\mathbb{1}_{x,y\in[0,1]} = \underbrace{3x^2\mathbb{1}_{x\in[0,1]}}_{\text{density of } X} \times \underbrace{2y\mathbb{1}_{y\in[0,1]}}_{\text{density of } Y}.$$

We conclude by a simple but yet very useful proposition.

Proposition 3.10 (*i.i.d. random variables are distinct*)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with density f.

Then almost surely they all are pairwise distinct:

$$\mathbb{P}\,(\ \textit{for all } i \neq j,\ X_i \neq X_j) = 1.$$

Proof: We prove that the complement event {there are i, j such that $X_i = X_j$} has probability zero. First we notice that

$$\begin{aligned}\mathbb{P}(\text{ there are } i,j \text{ such that } X_i = X_j) &= \mathbb{P}\left(\cup_{i,j\in\mathbb{N}} \{\text{ there are } i,j \text{ such that } X_i = X_j\}\right) \\ &\leq \sum_{i,j\in\mathbb{N}} \mathbb{P}\left(X_i = X_j\right),\end{aligned}$$

by the union bound (Proposition 1.3). Now,

$$\begin{aligned}\mathbb{P}\left(X_i = X_j\right) &= \int_{\mathbb{R}^2} \mathbb{1}_{x=y} f(x)f(y)dxdy \\ &= \int_{y\in\mathbb{R}} \left(\int_{x\in\mathbb{R}} \mathbb{1}_{x=y} f(x)dx\right) f(y)dy \\ &= \int_{y\in\mathbb{R}} \left(\int_{x=y}^{y} f(x)dx\right) f(y)dy \\ &= \int_{y\in\mathbb{R}} 0 \times f(y)dy = 0.\end{aligned}$$

■

3.4 Sums of independent random variables

Let X, Y be random variables, what can we say about $X + Y$? First, by linearity of expectation

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

Now, assume that X, Y are independent. Then

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2 \\ &= \mathbb{E}[X^2 + Y^2 + 2XY] - \mathbb{E}[X]^2 - \mathbb{E}[Y]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] = \mathrm{Var}(X) + \mathrm{Var}(Y).\end{aligned}$$

More generally, if $X_1, \ldots, X_n$ are independent, then expectations and variances add up:

$$\begin{aligned}\mathbb{E}[X_1 + \cdots + X_n] &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n], \\ \mathrm{Var}(X_1 + \cdots + X_n) &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \quad \text{(here } X_k\text{'s need to be } \textbf{independent}\text{!)}\end{aligned}$$

More precisely, what can we say about the distribution of the sum of independent random variables?

Assume first that X, Y have densities f, g. We want to compute the density (if any) of $X+Y$. Take a bounded and continuous function ϕ and compute

$$\begin{aligned}
\mathbb{E}[\phi(X+Y)] &= \iint \phi(x+y)f(x)g(y)dxdy \\
&= \int_x f(x)\left(\int_y \phi(x+y)g(y)dy\right)dx \quad \text{(by Fubini n.2)} \\
&= \int_x f(x)\left(\int_u \phi(u)g(x-u)du\right)dx \quad \text{(by ch. of variables } u = x+y,\ \tfrac{du}{dy} = 1) \\
&= \int_u \phi(u)\underbrace{\left(\int_x f(x)g(u-x)dx\right)}_{\text{density of } \phi(X+Y)}du \quad \text{(again by Fubini n.2) .}
\end{aligned}$$

Then we have a formula for the density of $X+Y$. It is called the *convolution* of f and g:

Definition/Theorem 3.11 (*Convolution of two densities*)

Let X, Y be two independent random variables with densities f and g. Then $X+Y$ has a density, it is denoted $f * g$ and given by the formula

$$f * g\ (u) = \int_{x\in\mathbb{R}} f(x)g(u-x)dx.$$

The function $u \mapsto f * g\ (u)$ is called the convolution *of f and g.*

Remark : Obviously $X+Y$ has the same density as $Y+X$ so we should have $f * g = g * f$ *i.e.*

$$\int_{x\in\mathbb{R}} f(x)g(u-x)dx = \int_{x\in\mathbb{R}} g(x)f(u-x)dx$$

for every u. You can check this with the change of variables $v = u - x$.

Example : Let X, Y be i.i.d., with the exponential distribution, *i.e.* $f(x) = e^{-x}\mathbb{1}_{x\geq 0}$ and $g(y) = e^{-y}\mathbb{1}_{y\geq 0}$. The random variable $X+Y$ also takes its values in $[0, +\infty)$ and by the previous computation, the density of $X+Y$ is given by

$$\begin{aligned}
\text{(for all } u \geq 0), \qquad \int_{x=0}^{\infty} f(x)g(u-x)dx &= \int_{x=0}^{\infty} e^{-x}e^{-(u-x)}\mathbb{1}_{u-x\geq 0}\ dx \\
&= e^{-u}\int_{x=0}^{\infty} \mathbb{1}_{u-x\geq 0}\ dx = e^{-u}\int_{x=0}^{\infty} \mathbb{1}_{x\leq u}\ dx \\
&= e^{-u}\int_{x=0}^{u} dx = ue^{-u}.
\end{aligned}$$

Then $X+Y$ has density $ue^{-u}\mathbb{1}_{u\geq 0}$ (you can check that it is a density).

▷ Sums of random variables and characteristic functions

Another very efficient tool for the sum of r.v. is the use of characteristic functions (in fact this is the very reason for which characteristic functions are introduced in this course). Indeed, we have

$$\Phi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] \overset{\text{by ind.}}{=} \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}] = \Phi_X(t)\Phi_Y(t),$$

the CF of a sum of independent random variables is the product of CFs!

As an important application we have:

Proposition 3.12 (*Sum of two independent gaussian r.v.*)

If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and if X, Y are independent, then $X + Y$ is also a gaussian random variable. More precisely,

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Proof: We admit for this proof that we have the following formula for the characteristic function of a $\mathcal{N}(\mu, \sigma^2)$:

$$\Phi_X(t) = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right).$$

This will be proved later in Proposition 4.1. In our case it gives that

$$\Phi_{X_1}(t) = \exp\left(it\mu_1 - \frac{t^2\sigma_1^2}{2}\right), \qquad \Phi_{X_2}(t) = \exp\left(it\mu_2 - \frac{t^2\sigma_2^2}{2}\right)$$

Let us compute the CF of $X_1 + X_2$:

$$\begin{aligned}\Phi_{X_1+X_2}(t) = \Phi_{X_1}(t)\Phi_{X_2}(t) &= \exp\left(it\mu_1 - \frac{t^2\sigma_1^2}{2}\right)\exp\left(it\mu_2 - \frac{t^2\sigma_2^2}{2}\right) \\ &= \exp\left(it(\mu_1+\mu_2) - \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right),\end{aligned}$$

and we recognize the CF of a $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. ■

3.5 Bonus: The Borel-Cantelli lemmas

Given events $A_1, A_2, \ldots$, a main concern is often "how many of the A_n's occur?". The first Borel-Cantelli Lemma says that if $\mathbb{P}(A_n)$ is too small, then A_n cannot occur infinitely often. Recall that

$$\limsup_{n\to\infty} A_n = \text{ "}A_n \text{ occurs infinitely often" } = \bigcap_{p\geq 1}\bigcup_{n\geq p} A_n.$$

Theorem 3.13 (*First Borel-Cantelli Lemma*)

If $\sum_{n\geq 1}\mathbb{P}(A_n) < +\infty$ then, almost surely, A_n occurs finitely many times:

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

On the contrary, if $\mathbb{P}(A_n)$ is large enough, then we can prove that A_n occurs infinitely often. This is the second Borel-Cantelli Lemma, but we need the extra assumption that A_n's are independent:

Theorem 3.14 (*Second Borel-Cantelli Lemma*)

If $\sum_{n\geq 1}\mathbb{P}(A_n) = +\infty$ and if furthermore $A_1, A_2, \dots$ are independent then, almost surely, A_n occurs infinitely often:

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1.$$

You can skip the proofs of Borel-Cantelli's lemmas, more important is the application just below, which helps to understand how B-C's lemmas work in practice.

Proof of the first Borel-Cantelli lemma : We consider the random variable $\sum_{n\geq 1}\mathbb{1}_{A_n}$ that counts the number of A_n's that occur. By positivity (see Proposition 2.18) one can swap $\sum$ and $\mathbb{E}$ and then

$$\mathbb{E}\Big[\sum_{n\geq 1}\mathbb{1}_{A_n}\Big] = \sum_{n\geq 1}\mathbb{E}\left[\mathbb{1}_{A_n}\right] = \sum_{n\geq 1}\mathbb{P}(A_n) < +\infty.$$

Then $\sum_{n\geq 1}\mathbb{1}_{A_n}$ has finite expectation, therefore it is finite with probability one: A_n occurs finitely many times. ■

Proof of the second Borel-Cantelli lemma : Fix $p \geq 1$,

$$\begin{aligned}\mathbb{P}\left(A_n \text{ doesn't occur for } n\geq p\right) &= \mathbb{P}\left(\overline{A_p}\cap\overline{A_{p+1}}\cap\overline{A_{p+2}}\cap\dots\right)\\ &\overset{\text{by indep.}}{=} \prod_{n\geq p}\mathbb{P}(\overline{A_n}) = \prod_{n\geq p}\left(1-\mathbb{P}(A_n)\right).\end{aligned}$$

Now, we use the fact that $1-x\leq e^{-x}$ for any real x:

$$\mathbb{P}\left(A_n \text{ doesn't occur for } n\geq p\right) \leq \prod_{n\geq p}\exp\left(-\mathbb{P}(A_n)\right) = \exp\left(-\sum_{n\geq p}\mathbb{P}(A_n)\right) = \exp(-\infty) = 0.$$

$$\begin{aligned}\text{Then}\qquad \mathbb{P}\left(A_n \text{ occurs finitely many times }\right) &= \mathbb{P}\left(\bigcup_{p\geq 1}\left\{A_n \text{ doesn't occur for } n\geq p\right\}\right)\\ &\leq \sum_{p\geq 1}\mathbb{P}\left(A_n \text{ doesn't occur for } n\geq p\right) = \sum_{p\geq 1} 0 = 0\end{aligned}$$

(we used item 3. in Proposition 1.3). This proves that $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$. ■

An application of Borel-Cantelli's lemmas: extreme values

Let $X_1, X_2, \dots$ be i.i.d. exponential random variables with parameter 1. We have for each n

$$\mathbb{P}(X_n \geq t) = \int_t^{+\infty} e^{-u}du = e^{-t}.$$

What can we say about very large values of the sequence (X_n) ? First of all, it should be obvious that this infinite sequence is not bounded. To prove so, take a huge number, 10^{100} say. We have, for each n, $\mathbb{P}(X_n \geq 10^{100}) = e^{-10^{100}} > 0$. Then

$$\sum_{n\geq 1} \mathbb{P}(X_n \geq 10^{100}) = \sum_{n\geq 1} e^{-10^{100}} = +\infty.$$

Then, applying Borel-Cantelli n.2 with events $A_n = \{X_n \geq 10^{100}\}$ (which are independent) shows that A_n occurs infinitely often : $\{X_n \geq 10^{100}\}$ for infinitely many n's.

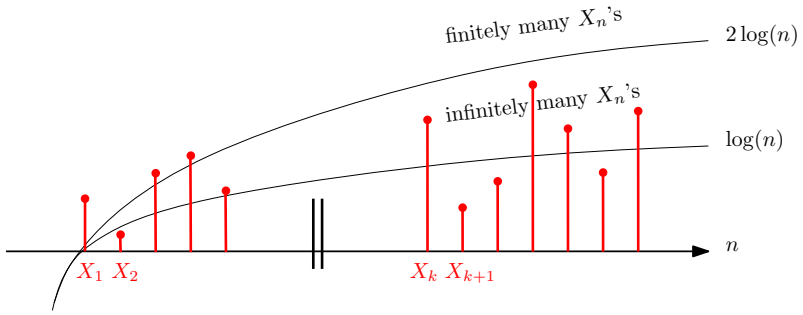
Now we would like to be more precise: for a fixed n, how large are extreme values among $X_1, \ldots, X_n$? We can prove that they are of order $\log(n)$. First,

$$\sum_{n\geq 1} \mathbb{P}(X_n \geq \log(n)) = \sum_{n\geq 1} e^{-\log(n)} = \sum_{n\geq 1} \frac{1}{n} = +\infty.$$

Then, again by Borel-Cantelli n.2, we have $\{X_n \geq \log(n)\}$ for infinitely many n's. On the other hand,

$$\sum_{n\geq 1} \mathbb{P}(X_n \geq 2\log(n)) = \sum_{n\geq 1} e^{-2\log(n)} = \sum_{n\geq 1} \frac{1}{n^2} < +\infty.$$

Hence, by Borel-Cantelli n.1, $\{X_n \geq 2\log(n)\}$ occurs only finitely many times. Here is a picture that sums up the situation:



We have precisely proved that, almost surely,

$$1 \leq \limsup_{n\to+\infty} \frac{X_n}{\log(n)} \leq 2.$$

4 Gaussian random variables and gaussian vectors

We first gather some useful properties of gaussian random variables.

Proposition 4.1 (*Properties of gaussian random variables*)

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ i.e.

$$\mathbb{P}(X \in A) = \int_A \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.$$

then $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.

- *The characteristic function of* X *is given by*

$$\Phi_X(t) = \mathbb{E}[e^{itX}] = \exp(it\mu - t^2\sigma^2/2). \qquad (\Diamond)$$

- *If* $X \sim \mathcal{N}(0,1)$ *then for two constants* $a, b \in \mathbb{R}$, $aX + b \sim \mathcal{N}(b, a^2)$.
- *If* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \ldots, X_d \sim \mathcal{N}(\mu_d, \sigma_d^2)$ *are independent then*

$$X_1 + \cdots + X_d \sim \mathcal{N}\left(\mu_1 + \cdots + \mu_d, \sigma_1^2 + \cdots + \sigma_d^2\right).$$

The two last items can be proved using characteric functions.

Proof of formula $(\Diamond)$**:** (You can skip the proof.)

First of all, let us prove the claimed result if $X \sim \mathcal{N}(0,1)$, in this case we need to prove that $\Phi_X(t) = \exp(-t^2/2)$. Let us differentiate Φ_X:

$$\begin{aligned}
\frac{\partial}{\partial t}\Phi_X(t) &= \mathbb{E}\left[\frac{\partial}{\partial t} e^{itX}\right] \qquad \text{(using Theorem 2.20)} \\
\Phi'_X(t) &= \int_{\mathbb{R}} \frac{\partial}{\partial t} e^{itx} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \\
&= \int_{\mathbb{R}} ixe^{itx} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \\
&= \frac{i}{\sqrt{2\pi}} \int_{\mathbb{R}} \underbrace{e^{itx}}_{v} \underbrace{x\exp(-x^2/2)}_{u'} dx = \frac{i}{\sqrt{2\pi}} \left(\cancel{[uv]_{-\infty}^{+\infty}} - \int_{-\infty}^{+\infty} uv' \right) \\
&= \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-x^2/2) ite^{itx} = -t\Phi_X(t).
\end{aligned}$$

Thus we have to solve the differential equation $\frac{\Phi'_X(t)}{\Phi_X(t)} = -t$, which is equivalent to

$$(\log(\Phi_X(t)))' = -t.$$

Hence $\log(\Phi_X(t)) = -t^2/2 + c$ for some constant c, *i.e.* $\Phi_X(t) = \exp(-t^2/2)e^c$. But recall that $\Phi_X(0) = 1$, so $e^c = 1$.

We turn to the general case where $X \sim \mathcal{N}(\mu, \sigma^2)$. We can write $X = \mu + \sigma \times Z$, where $Z \sim \mathcal{N}(0,1)$. Now,

$$\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[e^{it\mu + it\sigma Z}\right] = e^{it\mu}\mathbb{E}\left[e^{it\sigma Z}\right] = e^{it\mu}\mathbb{E}\left[e^{i(t\sigma)Z}\right] = e^{it\mu}\Phi_Z(t\sigma) = \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right). \quad \blacksquare$$

We are now interested in *multivariate* gaussian random variables.

Definition 4.2 (*Gaussian vectors*)

A random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ *is a* gaussian vector *(one also says that* $\mathbf{X}$ *follows the* multivariate gaussian distribution*) if for every real numbers* $t_1, t_2, \ldots, t_d$ *the linear combination*

$$t_1X_1 + t_2X_2 + \cdots + t_dX_d$$

is a gaussian random variable (here, by convention, we say that the constant random variable c *follows the gaussian distribution* $\mathcal{N}(c, 0)$*).*

Remark : 1. If $X_1, \ldots, X_d$ are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then Proposition 4.1 tells that the linear combination

$$t_1X_1 + t_2X_2 + \cdots + t_dX_d \sim \mathcal{N}\left(\ t_1\mu_1 + \cdots + t_d\mu_d\ ,\ t_1^2\sigma_1^2 + \cdots + t_d^2\sigma_d^2\ \right)$$

and therefore is gaussian: **a vector made of independent gaussian r.v. is a gaussian vector**.

2. If $\mathbf{X} = (X_1, \ldots, X_d)$ is a gaussian vector, then every sub-vector of $\mathbf{X}$ is a gaussian vector as well: for instance, (X_2, X_5, X_{11}) is a gaussian vector. To show this it suffices to take in the definition above $t_2 = t_5 = t_{11} = 1$ and every other t_i equal to zero.
3. In particular, by taking $t_i = 1$ and $t_j = 0$ for $j \neq i$, the definition tells that X_i is gaussian: **every component of a gaussian vector is gaussian**.

Definition 4.3 (*Parameters of a gaussian vector*)

Let $\mathbf{X}$ *be a gaussian vector. The* mean vector μ *of* $\mathbf{X}$ *is the column vector of expectations:*

$$\mu = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

and the covariance matrix C *of* $\mathbf{X}$ *is the* $d \times d$ *matrix where entry* $C_{i,j} = \mathrm{Cov}(X_i, X_j)$.

Example : If $\mathbf{X}$ is made of i.i.d. $X_1, \dots, X_d$ with $X_i \sim \mathcal{N}(0,1)$, then $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = 1$, and $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ (by independence). Then C is the identity matrix:

$$C = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

Since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$, the matrix C is always symmetric. It is less obvious, but true, that C is also positive-semi-definite, which means that for all vector $\mathbf{t} = (t_1, \dots, t_d) \in \mathbb{R}^d$ one has the inequality $\mathbf{t}' C \mathbf{t} \geq 0$, where $\mathbf{t}'$ is the *transpose* of $\mathbf{t}$.

Remark : About matrix operations:

Recall that

$$\mathbf{t}' C \mathbf{t} = \mathbf{t}' \times (C\mathbf{t}) = \mathbf{t}' \times \begin{pmatrix} \sum_j C_{1,j} t_j \\ \sum_j C_{2,j} t_j \\ \vdots \\ \sum_j C_{d,j} t_j \end{pmatrix} = \sum_{i,j} t_i C_{i,j} t_j.$$

Note also that

$$(\mathbf{X}-\mu)(\mathbf{X}-\mu)' = \begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_d-\mu_d \end{pmatrix} \begin{pmatrix} X_1-\mu_1 & X_2-\mu_2 & \dots & X_d-\mu_d \end{pmatrix} = \begin{matrix} \\ i \end{matrix} \begin{pmatrix} & j & \\ & \vdots & \\ \dots & (X_i-\mu_i)(X_j-\mu_j) & \dots \\ & \vdots & \end{pmatrix}.$$

Thus by taking expectation we get

$$C = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)']. \qquad (\star)$$

Definition/Theorem 4.4

Let $\mathbf{X}$ be a gaussian vector with mean vector μ and covariance matrix C. Then the multivariate characteristic function *of $\mathbf{X}$ is defined by*

$$\begin{array}{rccc} \Phi_{\mathbf{X}} : & \mathbb{R}^d & \to & \mathbb{C} \\ & \mathbf{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_d \end{pmatrix} & \mapsto & \mathbb{E}\left[\exp(it_1X_1 + \cdots + it_dX_d)\right]. \end{array}$$

and for all $\mathbf{t}$ we have the formula

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \exp\left(i\mathbf{t}'\mu - \frac{\mathbf{t}' C \mathbf{t}}{2}\right). \qquad (\#)$$

Proof of formula (#): We fix a vector $\mathbf{t} = (t_1, \ldots, t_d)$ in $\mathbb{R}^d$. Since $\mathbf{X}$ is a gaussian vector, the random variable Y defined by the linear combination $Y = t_1X_1 + \cdots + t_dX_d$ is a gaussian random variable. We have

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[\exp\left(i\left(t_1X_1 + \cdots + it_dX_d\right)\right)\right] = \mathbb{E}\left[\exp(i \times 1 \times Y)\right] = \Phi_Y(1)$$

so it suffices to compute the characteristic function of Y. To do so, it is enough (since Y is gaussian) to compute $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$. First, by linearity of expectation,

$$\mathbb{E}[Y] = \mathbb{E}\left[t_1X_1 + \cdots + t_dX_d\right] = t_1\mathbb{E}[X_1] + \cdots + t_d\mathbb{E}[X_d] = \mathbf{t}'\mu.$$

Let us now compute $\mathbb{E}[Y^2]$:

$$\mathbb{E}[Y^2] = \mathbb{E}\left[(t_1X_1 + \cdots + t_dX_d)^2\right] = \sum_{i,j} t_it_j\mathbb{E}[X_iX_j]$$

$$\mathbb{E}[Y]^2 = \left(\sum_i t_i\mathbb{E}[X_i]\right)^2 = \sum_{i,j} t_it_j\mathbb{E}[X_i]\mathbb{E}[X_j].$$

Hence

$$\mathrm{Var}(Y) = \sum_{i,j} t_it_j\left(\mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\right) = \sum_{i,j} t_it_jC_{i,j} = \mathbf{t}'C\mathbf{t}.$$

The proof is finished since by Proposition 4.1

$$\Phi_Y(1) = \exp\left(i \times 1 \times \mathbb{E}[Y] - 1^2 \times \frac{\mathrm{Var}(Y)}{2}\right) = \exp\left(i\mathbf{t}'\mu - \frac{\mathbf{t}'C\mathbf{t}}{2}\right). \qquad \blacksquare$$

A consequence of Theorem 4.4 is that if $\mathbf{X}$ and $\mathbf{Y}$ are two gaussian vectors with the same mean vector and the same covariance matrix, then $\Phi_{\mathbf{X}}(\mathbf{t}) = \Phi_{\mathbf{Y}}(\mathbf{t})$ for all $\mathbf{t}$ and then $\mathbf{X}$ and $\mathbf{Y}$ have the same law. We summarize this in a proposition.

Proposition 4.5 (*The covariance is enough*)

The distribution of a gaussian vector $(X_1, \ldots, X_d)$ is fully characterized by its mean vector μ and its covariance matrix C.
In particular, if for some i, j we have $\mathrm{Cov}(X_i, X_j) = 0$ then X_i and X_j are independent.

Example: Let X, Y be independent $\mathcal{N}(0, 1)$. Let us use the proposition to show that $X + Y$ and $X - Y$ are independent.

First, $(X + Y, X - Y)$ is a gaussian vector: for all t_1, t_2, $t_1(X + Y) + t_2(X - Y) = (t_1 + t_2)X + (t_1 - t_2)Y$ is indeed a gaussian random variable. Now it is enough to compute the covariance. Using bilinearity repeatedly we get:

$$\begin{aligned}\mathrm{Cov}(X + Y, X - Y) &= \mathrm{Cov}(X, X - Y) + \mathrm{Cov}(Y, X - Y)\\ &= \mathrm{Cov}(X, X) - \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X) - \mathrm{Cov}(Y, Y)\\ &= \mathrm{Cov}(X, X) - \mathrm{Cov}(Y, Y)\\ &= \mathrm{Var}(X) - \mathrm{Var}(Y) = 0.\end{aligned}$$

A useful property is that if $\mathbf{X}$ is a gaussian vector then so is its image by a linear transformation.

Proposition 4.6 (*Linear transformation of a gaussian vector*)

Let $\mathbf{X} \in \mathbb{R}^d$ be a gaussian vector with mean vector μ and covariance matrix C and let M be a $p \times d$ matrix. Then $M\mathbf{X} \in \mathbb{R}^p$ is a gaussian vector with mean $M\mu$ and covariance matrix MCM'.

Proof: First,

$$M\mathbf{X} = \begin{pmatrix} \sum_j M_{1,j} X_j \\ \sum_j M_{2,j} X_j \\ \vdots \\ \sum_j M_{p,j} X_j \end{pmatrix},$$

so each linear combination of the components of $M\mathbf{X}$ is in fact a linear combination of the components of $\mathbf{X}$ and thus is gaussian. Therefore $M\mathbf{X}$ is a gaussian vector. It suffices to compute its mean and its covariance matrix. First, by linearity,

$$\mathbb{E}[M\mathbf{X}] = M\mathbb{E}[\mathbf{X}] = M\mu.$$

For the covariance we apply formula $(\star)$ just above (recall $(AB)' = B'A'$):

$$\begin{aligned} \text{Cov. matrix of } M\mathbf{X} &= \mathbb{E}[(M\mathbf{X} - M\mu)(M\mathbf{X} - M\mu)'] \\ &= \mathbb{E}[M(\mathbf{X} - \mu)\,(M(\mathbf{X} - \mu))'] \\ &= \mathbb{E}[M(\mathbf{X} - \mu)(\mathbf{X} - \mu)'M'] = M\mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu))']M' = MCM'. \end{aligned}$$

■

▷ How to simulate a gaussian vector?

Consider the following example: how to simulate a gaussian vector (X, Y) with mean and covariance

$$\mu = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad C = \begin{pmatrix} 5 & -1 \\ -1 & 10 \end{pmatrix},$$

By linearity, it suffices to find (X', Y') with mean zero and covariance C, and put $(X, Y) = (1, 3) + (X', Y')$.

Now, consider Z_1, Z_2 two independent $\mathcal{N}(0, 1)$, *i.e.* (Z_1, Z_2) is a gaussian vector with mean zero and covariance matrix Id_2. According to Theorem 4.6, if M is a 2×2 matrix then $M \times (Z_1, Z_2)$ is a gaussian vector with mean zero and covariance matrix

$$M\mathrm{Id}_2 M' = MM'.$$

Thus, to simulate (X, Y) it suffices to simulate two independent $\mathcal{N}(0, 1)$, find M such that $MM' = C$, and put

$$(X, Y) = (1, 3) + M \times (Z_1, Z_2).$$

You can check that $M = \begin{pmatrix} -1 & 2 \\ 3 & 1 \end{pmatrix}$ is a solution of our problem:

$$\begin{pmatrix} -1 & 2 \\ 3 & 1 \end{pmatrix} \times \begin{pmatrix} -1 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -1 \\ -1 & 10 \end{pmatrix}.$$

To conclude this chapter we will admit the following formula:

Proposition 4.7 (*Density of a gaussian vector*)

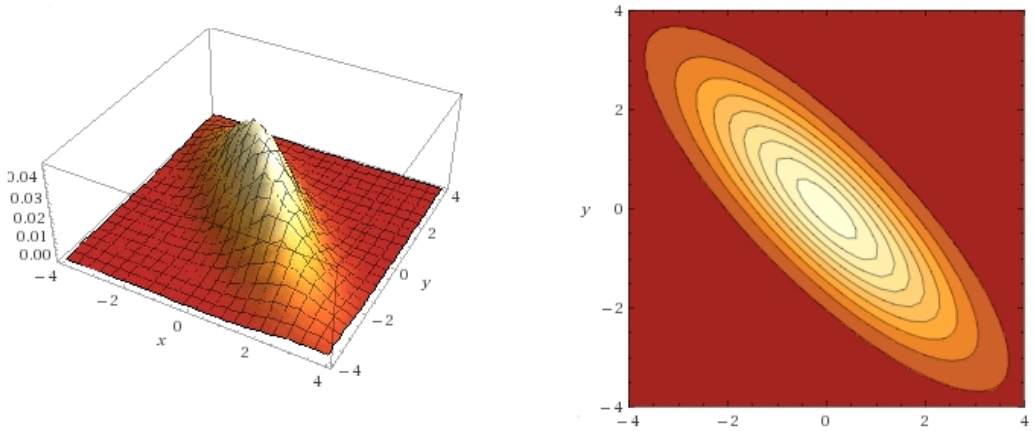
Let $\mathbf{X}$ *be a gaussian vector with mean vector* μ *and covariance matrix* C.
If C *is invertible* (i.e. *there exists* C^{-1} *such that* $C^{-1} \times C = \mathrm{Id}$, *this implies* $\det(C) \neq 0$) *then* $\mathbf{X}$ *has a density on* $\mathbb{R}^d$: *for all Borel set* A *of* $\mathbb{R}^d$,

$$\mathbb{P}(\mathbf{X} \in A) = \iint \dots \int_A \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\mu)' C^{-1} (\mathbf{x}-\mu)\right) dx_1 dx_2 \dots dx_d,$$

where $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$.

Remark :

- When $d = 1$ the matrix C is just $\big(\mathrm{Var}(X)\big)$ and then $\det(C) = \mathrm{Var}(X)$. We recover the density of the (one-dimensional) gaussian distribution.
- When C is not invertible then one can prove that $\mathbf{X}$ lies in a strict sub-space of $\mathbb{R}^d$ (for which Lebesgue measure is zero) and then $\mathbf{X}$ has no density.



The joint density of a centered gaussian vector with covariance matrix $\begin{pmatrix} 4 & -3 \\ -3 & 4 \end{pmatrix}$ *(compare with page 29). We have* $\mathbb{E}[XY] = -3$ *and consistently one sees that with high probability* X, Y *are of opposite signs.*

5 Conditioning

▷ Conditional probabilities

Recall that if A, B are events of $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{P}(B) > 0$, then we define

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

this is a prediction made on A, given that B occurs. Note that one can iterate this formula:

Proposition 5.1 (*Multiplicative formula for events*)

Let $A_1, \ldots, A_n$ be some events. Then

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \times \ldots \times \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

We consider now a different but related problem: what is the best prediction that can be made on X, given the value of another random variable Y. This should depend on Y, and therefore be a random variable.

5.1 Conditional expectation

Definition 5.2 (*Conditional expectation*)

- **(The discrete case)** *Let (X, Y) be a pair of discrete random variables, set*

$$p(x, y) = \mathbb{P}(X = x, Y = y), \qquad p_Y(y) = \mathbb{P}(Y = y) = \sum_x p(x, y).$$

The conditional expectation *of X given Y is defined by*

$$\mathbb{E}[X|Y] = \frac{\sum_x xp(x, Y)}{p_Y(Y)}.$$

- **(The continuous case)** *Let (X, Y) with joint density $f_{(X,Y)}(x, y)$, denote by $f_Y(y)$ the marginal density of Y. The* conditional expectation *of X given Y is defined by*

$$\mathbb{E}[X|Y] = \frac{\int_x x f_{(X,Y)}(x, Y)dx}{f_Y(Y)}.$$

More generally, let ϕ be any Borel function,

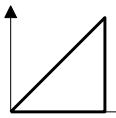
$$\mathbb{E}[\phi(X, Y)|Y] = \frac{\int_x \phi(x, Y) f_{(X,Y)}(x, Y)dx}{f_Y(Y)}.$$

Note that in both cases, by construction $\mathbb{E}[X|Y]$ is a function of Y, and therefore is a random variable.

We also introduce a useful notation in the continuous case:

$$\mathbb{E}[X|Y=y] = \frac{\int x f_{(X,Y)}(x,y)dx}{f_Y(y)}.$$

Note that $\mathbb{E}[X|Y]$ is a random variable, while $\mathbb{E}[X|Y=y]$ is a simple function of y.

Example : Let (X,Y) be uniform in the unit triangle  of area $1/2$, *i.e.* (X,Y) has density $f_{(X,Y)}(x,y) = 2 \times \mathbb{1}_{0\le y\le x\le 1}$. Then

$$f_Y(y) = \int_x 2\mathbb{1}_{0\le y\le x\le 1}dx = \int_{x=y}^{1} 2dx = 2(1-y).$$

Let us apply the proposition to compute $\mathbb{E}[X|Y]$. First,

$$\int_x x f(x,Y)dx = \int_x 2x\mathbb{1}_{0\le Y\le x\le 1}dx = 2\int_{x=Y}^{1} xdx = 2\left[x^2/2\right]_{x=Y}^{x=1} = 1-Y^2.$$

Then the formula reads

$$\mathbb{E}[X|Y] = \frac{\int_x x f_{(X,Y)}(x,Y)dx}{f_Y(Y)} = \frac{1-Y^2}{2(1-Y)} = \frac{1+Y}{2}.$$

(in particular you see that $\mathbb{E}[X|Y]$ is a function of Y).

Proposition 5.3 (*Properties of conditional expectations*)

Let X, X', Y be random variables. In either discrete or continuous case:

(i) **(Linearity)** $\mathbb{E}[aX+X'|Y] = a\mathbb{E}[X|Y] + \mathbb{E}[X'|Y]$ *for all constant a.*

(ii) **(Averaging)**

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}[X|Y]\right] = \int_y \mathbb{E}[X|Y=y] f_Y(y)dy.$$

(iii) **('Taking out what is known')**
For any Borel function g, we have $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$.
In particular, $\mathbb{E}[g(Y)|Y] = g(Y)$.

(iv) **(Independence)** *If X is independent of Y then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.*

Let us briefly justify these properties, at least in the continuous case (we skip the proof of (i), which is intuitive but not obvious in the settings of this course):

(ii) is obtained as follows. By definition of $\mathbb{E}[X|Y]$,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[X|Y]\right] &= \mathbb{E}\left[\frac{\int_x xf(x,Y)dx}{f_Y(Y)}\right] \\
&= \int_y \left(\frac{\int_x xf(x,y)dx}{f_Y(y)}\right) f_Y(y)dy \qquad \text{(note that this is equal to } \int_y \mathbb{E}[X|Y=y]f_Y(y)dy.\text{)} \\
&= \int_y \left(\frac{\int_x xf(x,y)dx}{\cancel{f_Y(y)}}\right) \cancel{f_Y(y)}dy \\
&= \int_y \int_x xf(x,y)dxdy = \mathbb{E}[X].
\end{aligned}$$

(iii) is also easy:

$$\begin{aligned}
\mathbb{E}[Xg(Y)|Y] &= \frac{\int_x xg(Y)f(x,Y)dx}{f_Y(Y)} \\
&= g(Y)\frac{\int_x xf(x,Y)dx}{f_Y(Y)} \\
&= g(Y)\mathbb{E}[X|Y].
\end{aligned}$$

(iv) is intuitive: Y doesn't bring information about X.

$$\begin{aligned}
\mathbb{E}[X|Y] &= \frac{\int_x xf(x,Y)dx}{f_Y(Y)} \\
&= \frac{\int_x xf_X(x)f_Y(Y)dx}{f_Y(Y)} \qquad \text{(by independence of } X,Y\text{)} \\
&= \frac{\int_x xf_X(x)\cancel{f_Y(Y)}dx}{\cancel{f_Y(Y)}} \\
&= \mathbb{E}[X].
\end{aligned}$$

Example : We briefly give an example that shows how we can use $\mathbb{E}[Y|X]$ in order to evaluate $\mathbb{E}[Y]$.

Pick X uniformly in $[0,1]$, and then pick Y uniformly in $[0,X]$. How to compute $\mathbb{E}[Y]$? Thanks to item (ii) (Averaging) in Proposition 5.3 we have:

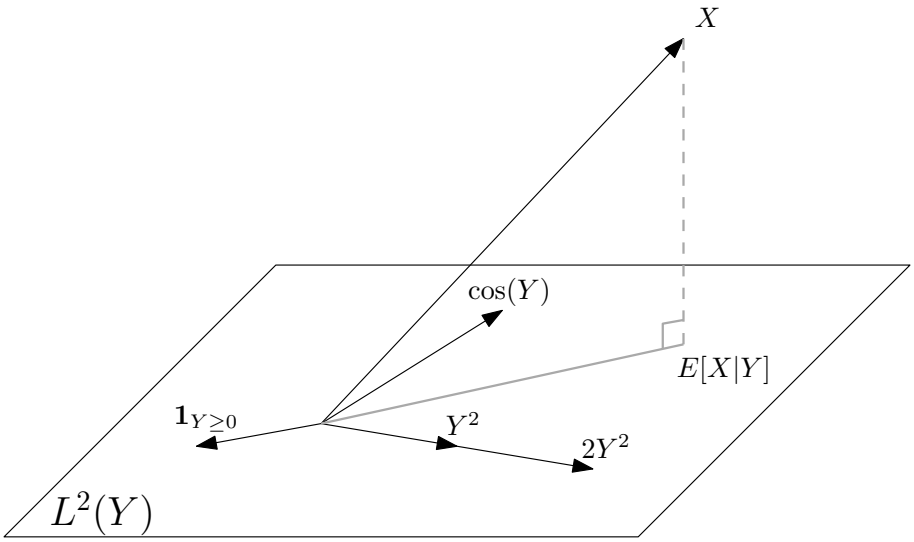
$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[\mathbb{E}[Y|X]\right] \\
&= \mathbb{E}\left[X/2\right] \qquad (Y \text{ is uniform in } [0,X]) \\
&= \int_{x=0}^{1} \tfrac{1}{2}xdx = \left[x^2/2\right]_{x=0}^{x=1} = 1/4.
\end{aligned}$$

▷ Conditional expectation as the best predictor

Let $X, Y \in L^2$. It turns out that in this case $\mathbb{E}[X|Y]$ can be built as the orthogonal projection of X onto the vector space

$$L^2(Y) = \left\{ \text{ random variables of the form } g(Y) \text{ such that } \mathbb{E}[g(Y)^2] < +\infty \right\}.$$

As we saw in Section 2.6, the orthogonal projection can be seen as the solution of a minimizing problem. This gives the following interpretation of $\mathbb{E}[X|Y]$.



Theorem 5.4 (*Conditional expectation as the best predictor*)

Let $X, Y \in L^2$. The conditional expectation $\mathbb{E}[X|Y]$ is the best predictor of X among all possible predictors which are functions of Y:

$$\mathbb{E}\Big[\,(X - \mathbb{E}[X|Y])^2\,\Big] \leq \mathbb{E}\Big[\,(X - g(Y))^2\,\Big]$$

for every $g(Y) \in L^2$.

5.2 Conditional distributions

Definition/Theorem 5.5

Let (X, Y) have joint density $f_{(X,Y)}(x, y)$, denote by $f_Y(y)$ the marginal density of Y. For $y \in \mathbb{R}$, the conditional density *of X given $Y = y$ (this is an abuse of notation since "$Y = y$" is an event of measure zero) is the function*

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

(we set $f_{X|Y=y}(x) = 0$ if $f_Y(y) = 0$). For each fixed y, $x \mapsto f_{X|Y=y}(x)$ is a probability density.

Remark :

- It is easy to check that $f_{X|Y=y}$ is a density:

$$\int_x f_{X|Y=y}(x)dx = \int_x \frac{f_{(X,Y)}(x, y)}{f_Y(y)}dx = \frac{1}{f_Y(y)}\int_x f_{(X,Y)}(x, y)dx = \frac{1}{f_Y(y)}f_Y(y) = 1.$$

- An important case is when X, Y are independent:

$$f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)} \overset{\text{(by indep.)}}{=} \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x),$$

as expected.

Example : We resume example of page 43: X is uniform in $[0,1]$, and Y uniform in $[0,X]$. This means

$$f_X(x) = \mathbb{1}_{0\leq x\leq 1}$$
$$f_{Y|X=x}(y) = \frac{1}{x}\mathbb{1}_{0\leq y\leq x}.$$

The formula given in Definition/Theorem 5.5 gives us the joint distribution of (X,Y):

$$f_{(X,Y)}(x,y) = f_X(x)f_{Y|X=x}(y) = \mathbb{1}_{0\leq x\leq 1}\times\frac{1}{x}\mathbb{1}_{0\leq y\leq x} = \frac{1}{x}\mathbb{1}_{0\leq y\leq x\leq 1}.$$

Then we can find the density of Y:

$$f_Y(y) = \int_{x=0}^{1}\frac{1}{x}\mathbb{1}_{0\leq y\leq x}dx = \int_{x=y}^{1}\frac{1}{x}dx = [\log(x)]_{x=y}^{x=1} = -\log(y).$$

▷ Bayes's formula

Recall the Bayes formula for events:

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

There a similar formula for densities:

Proposition 5.6 (*Bayes's Formula for conditional densities*)

We have

$$f_{X|Y=y} = \frac{f_{Y|X=x}f_X(x)}{f_Y(y)}.$$

Proof :

$$\begin{aligned}\frac{f_{Y|X=x}f_X(x)}{f_Y(y)} &= \frac{f(x,y)f_X(x)}{f_X(x)f_Y(y)}\\ &= \frac{f(x,y)\cancel{f_X(x)}}{\cancel{f_X(x)}f_Y(y)}\\ &= f_{X|Y=y}.\end{aligned}$$

■

6 More on random variables

6.1 Concentration inequalities

Here is the context: let $X_1, \dots, X_n$ be i.i.d. random variables with expectation $\mu = \mathbb{E}[X_1]$ and variance σ^2. We expect that, loosely speaking,

$$X_1 + \cdots + X_n \approx n\mu.$$

To be more precise, we can use Chebyshev's inequality. Set $S_n = X_1 + \cdots + X_n$,

$$\mathbb{P}\left(|S_n/n - \mu| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\mathrm{Var}(S_n)}{n^2\varepsilon^2} = \frac{n\mathrm{Var}(X_1)}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}. \qquad (\otimes)$$

(we used the properties of page 30). We see that the probability on the left-hand side decays *at least* as fast as $1/n$. The aim of this section is to show that under suitable assumptions on X one can obtain much better bounds.

▷ Concentration inequality for to the binomial: Hoeffding's inequality

We first consider the case where each $X_i = 0/1$ with probability $1/2$, *i.e.* $S_n = \mathrm{Binom}(n, 1/2)$.

Theorem 6.1 (*Hoeffding's inequality for the Binomial*)

Let S_n be a $\mathrm{Binom}(n, 1/2)$. *For every real $A > 0$, we have*

$$\mathbb{P}\left(|S_n - n/2| \geq A\right) \leq 2\exp\left(-\frac{A^2}{2n}\right).$$

Comparison with Bienaymé-Tchebychev : In the case of the $\mathrm{Binom}(n, 1/2)$, we observe that Hoeffding's inequality overpasses Bienaymé-Tchebychev's inequality. In this case $\sigma^2 = 1/2$ and eq.$(\otimes)$ reads

$$\mathbb{P}\left(|S_n/n - 1/2| \geq \varepsilon\right) \leq \frac{1}{2n\varepsilon^2}.$$

On the other hand, Hoeffding's inequality gives

$$\mathbb{P}\left(|S_n/n - 1/2| \geq \varepsilon\right) = \mathbb{P}\left(|S_n - n/2| \geq n\varepsilon\right) \leq 2\exp(-(n\varepsilon)^2/2n) = 2\exp(-n\varepsilon^2/2).$$

Proof of Theorem 6.1 : First notice that

$$\mathbb{P}\left(|S_n - n/2| \geq A\right) = \mathbb{P}(S_n \geq n/2 + A) + \mathbb{P}(S_n \leq n/2 - A),$$

and the two terms in the RHS are in fact equal. Thus we only have to bound $\mathbb{P}(S_n \geq n/2 + A)$.

The idea, due to Chernov, is to use the simple fact that, for every parameter $\lambda > 0$, the map $x \mapsto e^{\lambda x}$ is one-to-one and increasing. Thus we have

$$\mathbb{P}(S_n \geq n/2 + A) = \mathbb{P}\big(\exp(\lambda S_n) \geq \exp(\lambda(n/2 + A))\big).$$

Now, by the Markov inequality applied to the non-negative random variable $e^{\lambda S_n} = e^{\lambda X_1} \dots e^{\lambda X_n}$, we have

$$\begin{aligned}\mathbb{P}\big(\exp(\lambda S_n) \geq \exp(\lambda(n/2+A))\big) &\leq \frac{\mathbb{E}\left[e^{\lambda X_1} \dots e^{\lambda X_n}\right]}{\exp(\lambda(n/2+A))} \\ &\leq \frac{\mathbb{E}\left[e^{\lambda X_1}\right]^n}{\exp(\lambda(n/2+A))}.\end{aligned}$$

Besides,

$$\mathbb{E}\left[e^{\lambda X_1}\right] = \tfrac{1}{2}e^{\lambda\times 1} + \tfrac{1}{2}e^{\lambda\times 0} = \tfrac{1}{2}\left(e^{\lambda}+1\right) \leq \exp(\lambda/2+\lambda^2/8).$$

(The last inequality can be checked by computer or first-year calculus). We obtain

$$\mathbb{P}(S_n \geq n/2+A) \leq \frac{\exp(n(\lambda/2+\lambda^2/8))}{\exp(\lambda(n/2+A))} = \exp(n\lambda^2/8 - \lambda A) = \exp(\varphi(\lambda)),$$

where we put $\varphi(\lambda) = n\lambda^2/8 - \lambda A$.

Note that the left-hand side does not depend on λ. Chernov's idea is to find the $\lambda_\star$ that minimizes the right-hand side. Since

$$\varphi'(\lambda) = \tfrac{n}{4}\lambda - A,$$

function φ is minimal for $\lambda_\star = 4A/n$. Finally,

$$\mathbb{P}(S_n \geq n/2+A) \leq \exp(\varphi(\lambda_\star)) = \exp\left(n(4A/n)^2/8 - (4A/n)\times A\right) = \exp(-A^2/2n).$$

This concludes the proof:

$$\mathbb{P}\left(|S_n - n/2| \geq A\right) = \mathbb{P}(S_n \geq n/2+A) + \mathbb{P}(S_n \leq n/2-A) \leq \exp(-A^2/2n) + \exp(-A^2/2n).$$

■

More generally, Hoeffding's inequality can be used for bounded random variables.

Theorem 6.2 (*Hoeffding's inequality for bounded random variables*)

Let $X_1, \dots, X_n$ be i.i.d. random variables and assume that there exist a, b such that almost surely, $a \leq X_1 \leq b$. Set $S_n = X_1 + \dots + X_n$, and $\mu = \mathbb{E}[X_1]$. For every real $A > 0$, we have

$$\mathbb{P}\left(|S_n - n\mu| \geq A\right) \leq 2\exp\left(-\frac{2A^2}{n(b-a)^2}\right).$$

The proof can be found at `https://en.wikipedia.org/wiki/Hoeffding's_inequality`. An important thing to note is that if A is much larger than $\sqrt{n}$, then Hoeffding's inequality implies that the right-hand side decays exponentially when $n \to +\infty$.

6.2 Order statistics

In this section, we discuss the following family of problems. Let $X_1, \dots, X_n$ be continuous i.i.d. random variables, what can we say about the law of the sample generated by *re-ordering* the X_k's in the increasing order?

A usual notation is to denote by $X_{(1)}$ the smallest value, more generally $X_{(k)}$ stands for the kth-smallest value. In short, we have

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

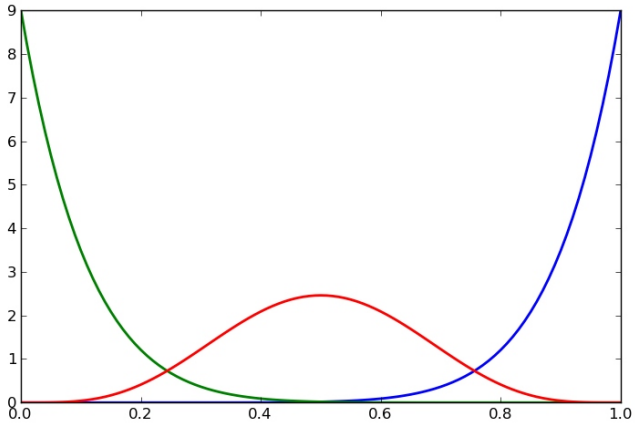
and $\{X_1, \dots, X_n\} = \{X_{(1)}, \dots, X_{(n)}\}$.

Proposition 6.3 (*Law of the minimum, maximum and median*)

Let $X_1, \dots, X_n$ be i.i.d. with density f and cdf F.

- *$X_{(1)}$ has density*
$$nf(t)(1-F(t))^{n-1}.$$
- *$X_{(n)}$ has density*
$$nf(t)(F(t))^{n-1}.$$
- *If $n = 2k+1$ is odd, the median $X_{(k+1)}$ has density*
$$\binom{2k+1}{k}(k+1)F(t)^k f(t)(1-F(t))^k.$$

Here is a plot of the density of the min, max and median of a sample of 9 uniform random variables (resp. $t \mapsto 9t^8$, $t \mapsto 630t^4(1-t)^4$, $t \mapsto 9(1-t)^8$):



Proof: We start by the density of $X_{(1)}$. By definition

$$\begin{aligned}\mathbb{P}(X_{(1)} > t) &= \mathbb{P}(X_1 > t, X_2 > t, \dots, X_n > t) \\ &= \mathbb{P}(X_1 > t)\mathbb{P}(X_2 > t)\mathbb{P}(X_n > t) \qquad \text{(by ind.)} \\ &= (1-F(t))^n\,.\end{aligned}$$

Therefore the cdf of $X_{(1)}$ is $t \mapsto 1-(1-F(t))^n$ and by differentiating we get the desired result. The formula for $X_{(n)}$ is obtained by the same method.

We only sketch the proof for the median. Let ε be a "small" number, we aim at computing the probability that $X_{(k+1)}$ lies in $(t, t+\varepsilon)$. Indeed, one should have

$$\mathbb{P}\left(X_{(k+1)} \in (t, t+\varepsilon)\right) = \int_t^{t+\varepsilon} m(x)dx \approx \varepsilon m(t),$$

where m is the density of $X_{(k+1)}$.

Now, the event $\left\{X_{(k+1)} \in (t, t+\varepsilon)\right\}$ occurs if exactly k values lie in $(-\infty, t)$, one in $(t, t+\varepsilon)$, and the k remaining values in $(t+\varepsilon, +\infty)$. There are $\binom{2k+1}{k} \times (k+1)$ choices, so we get

$$\mathbb{P}\left(X_{(k+1)} \in (x, x+\varepsilon)\right) = \binom{2k+1}{k} (F(t))^k \times (k+1) \times \varepsilon \times (1-F(t+\varepsilon))^k,$$

which gives the desired formula. ■

▷ Number of records

We now establish a simple and useful result: we estimate the number of *records* in a sequence of n i.i.d. random variables, *i.e.* the number of k's for which X_k is bigger than all the preceding values.

Proposition 6.4 (*Records in an i.i.d. sequence*)

Let $X_1, \dots, X_n$ be i.i.d. random variables with density f. Then, regardless of f, for each $1 \leq k \leq n$,

$$\mathbb{P}(X_k = \max\{X_1, \dots, X_n\}) = 1/n.$$

Moreover, let N_n be the number of records *in the sequence $\{X_1, \dots, X_n\}$,* i.e.

$$N_n = \operatorname{card}\{k \leq n, X_k = \max\{X_1, \dots, X_k\}\}.$$

Then

$$\mathbb{E}[N_n] = \frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{n} \overset{n\to+\infty}{\sim} \log(n).$$

Proof: First, recall (Prop.3.10) that almost surely all X_k's are pairwise distinct. Therefore

$$\begin{aligned} 1 = \quad & \mathbb{P}(X_1 = \max\{X_1, \dots, X_n\}) \\ & + \mathbb{P}(X_2 = \max\{X_1, \dots, X_n\}) \\ & \dots \\ & + \mathbb{P}(X_n = \max\{X_1, \dots, X_n\}). \end{aligned}$$

Now, each of the n terms on the right-hand side is equal to

$$\int_{\mathbb{R}^n} \mathbb{1}_{x_1 = \max\{x_1, \dots, x_n\}} f(x_1) \dots f(x_n) dx_1 \dots dx_n,$$

and since they sum up to one, each term equals $1/n$.

It is now easy to prove the second assertion: by linearity

$$\mathbb{E}[N_n] = \sum_{k=1}^{n} \mathbb{E}[\mathbb{1}_{X_k \text{ is a record}}] = \sum_{k=1}^{n} \mathbb{P}(X_k = \max\{X_1, \dots, X_k\}) = \sum_{k=1}^{n} \frac{1}{k}.$$

The fact that $\sum_{k=1}^{n} \frac{1}{k} \sim \log(n)$ is well-known[(i)]. ■

▷ Joint distribution of the order statistics

Finally we state (without proof) the following formula:

Theorem 6.5

Let $X_1, \dots, X_n$ be i.i.d. with density f. Then $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ has density

$$n! f(x_1) f(x_2) \dots f(x_n) \mathbb{1}_{x_1 < \dots x_n}.$$

Example : Let X_1, X_2, X_3 be three independent Uniform$(0,1)$. Then $(X_{(1)}, X_{(2)}, X_{(3)})$ has joint density $6 \times \mathbb{1}_{0<x_1<x_2<x_3<1}$.

6.3 Mixture of densities

Mixture of densities arise naturally in many models, when the population contains two (or more) subpopulations.

Definition/Theorem 6.6 (*Mixture of two densities*)

Let f_1, f_2 be two densities, and let $w_1, w_2 \in (0,1)$, such that $w_1 + w_2 = 1$. The mixture *of densities f_1, f_2 with weights w_1, w_2 is the density given by*

$$x \mapsto w_1 f_1(x) + w_2 f_2(x).$$

The mixture density has the following interpretation: if X, Y have density f_1, f_2, if $U \sim$ Ber(w_1) and if X, Y, U are independent, then

$$UX + (1-U)Y \text{ has density } w_1 f_1 + w_2 f_2.$$

The previous assertion says that a mixture of f_1, f_2 is obtained when you pick $U \in \{0,1\}$ and, conditional on U, you pick X at random according to density f_1 or f_2.

Proof : We prove that $UX + (1-U)Y$ has density $w_1 f_1 + w_2 f_2$. Let ϕ be bounded and continuous, we use

$$\mathbb{E}\left[\phi(UX + (1-U)Y)\right] = \mathbb{E}\Big[\mathbb{E}[\phi(UX + (1-U)Y)|U]\Big].$$

[(i)] `https://en.wikipedia.org/wiki/Harmonic_series_(mathematics)`

Note that

$$\begin{cases} UX + (1-U)Y = X & \text{if } U = 1, \\ UX + (1-U)Y = Y & \text{if } U = 0. \end{cases}$$

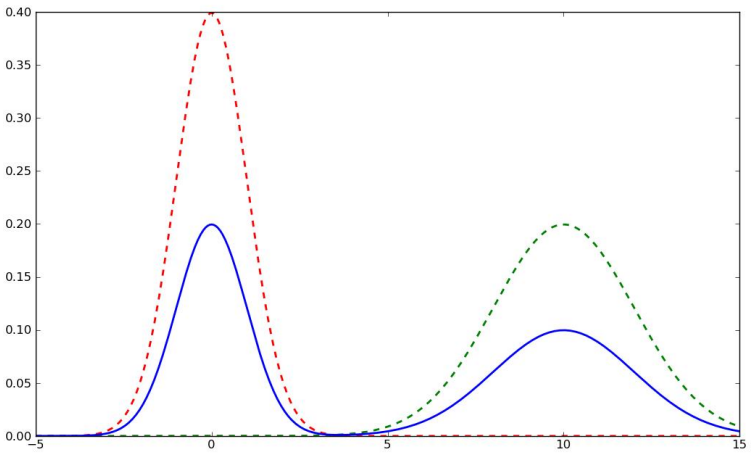
Therefore

$$\begin{cases} \mathbb{E}\left[\phi(UX + (1-U)Y)|U\right] = \mathbb{E}[\phi(X)|U] = \int \phi(x) f_1(x) dx \text{ if } U = 1, \\ \mathbb{E}\left[\phi(UX + (1-U)Y)|U\right] = \mathbb{E}[\phi(Y)|U] = \int \phi(y) f_2(y) dy \text{ if } U = 0. \end{cases}$$

We obtain

$$\begin{aligned} \mathbb{E}\Big[\mathbb{E}[\phi(UX + (1-U)Y)|U]\Big] &= \int \phi(x) f_1(x) dx \times w_1 + \int \phi(y) f_2(y) dy \times w_2 \\ &= \int \phi(x) \left(w_1 f_1(x) + w_2 f_2(x)\right) dx. \end{aligned}$$

■



The densities of $\mathcal{N}(0,1), \mathcal{N}(10,2)$ *(dashed lines) and the density of the mixture with weights* $(0.5; 0.5)$ *(solid line).*

7 Convergences of random variables

7.1 $\neq$ kinds of convergences of random variables

Let $(X_n)_{n\geq 1}$ be a sequence of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let X be another random variable defined on Ω.

Definition 7.1

- *The sequence* $(X_n)_{n\geq 1}$ *converges to* X in probability *if for all real number* $\varepsilon > 0$

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

  One writes $(X_n) \overset{\text{prob.}}{\to} X$.

- *The sequence* $(X_n)_{n\geq 1}$ *converges to* X in L^p *if*

$$\mathbb{E}\left[\ |X_n - X|^p\ \right] \overset{n\to+\infty}{\to} 0$$

  (of course it is equivalent to $\|X_n - X\|_p \to 0$*). One writes* $(X_n) \overset{L^p}{\to} X$*, and one also says that* (X_n) *converges to* X *in* p*-th mean.*

- *The sequence* $(X_n)_{n\geq 1}$ *converges to* X almost surely *if*

$$\mathbb{P}\left(X_n \overset{n\to\infty}{\longrightarrow} X\right) = \mathbb{P}\left(\omega \text{ such that } X_n(\omega) \overset{n\to\infty}{\longrightarrow} X(\omega)\right) = 1.$$

  One writes $(X_n) \overset{\text{a.s.}}{\to} X$.

For statistical applications, convergences in probability and in L^p are the most useful, so we will focus on them in the sequel.
Here is an example that shows that these kinds of convergence are NOT equivalent:

Example : $\left(\overset{\text{prob.}}{\to} \neq \overset{L^p}{\to}\right)$ Take a sequence of independent random variables $X_1, X_2, \ldots$ such that

$$X_n = \begin{cases} \sqrt{n} & \text{with probability } \frac{1}{n}, \\ 0 & \text{with probability } 1 - \frac{1}{n}. \end{cases}$$

When n goes large, X_n is more and more likely to be zero, so we expect $(X_n)_{n\geq 1}$ to converge (at least in some sense) to zero.

Let us first check that $(X_n) \overset{\text{prob.}}{\to} 0$: fix a small $\varepsilon > 0$, we have

$$\mathbb{P}\left(|X_n - 0| > \varepsilon\right) = \mathbb{P}\left(X_n = \sqrt{n}\right) = 1/n \overset{n\to+\infty}{\to} 0.$$

This proves that $(X_n) \overset{\text{prob.}}{\to} 0$.

Let us now consider the convergence to 0 in L^p.

$$\mathbb{E}\left[\ |X_n - 0|^p\ \right] = \mathbb{E}[(X_n)^p] = 0 \times (1 - \tfrac{1}{n}) + (\sqrt{n})^p \times \tfrac{1}{n} = n^{p/2-1}.$$

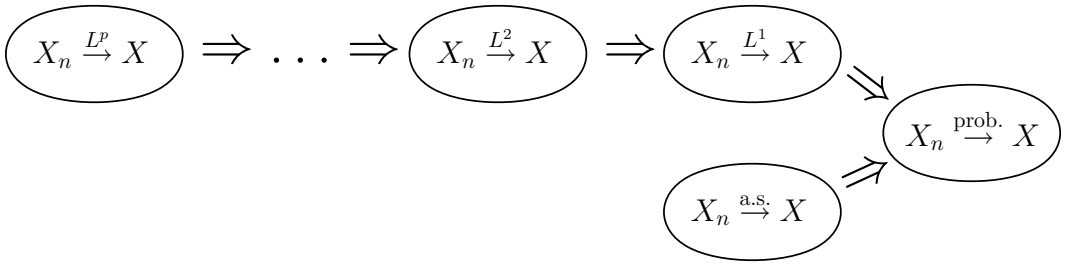
This goes to zero for $p < 2$: $(X_n) \overset{L^p}{\to} X$ for all $1 \leq p < 2$.

Convergence in probability is in fact the "weakest" of all:

Proposition 7.2

- *If* $(X_n) \overset{L^p}{\to} X$, *then* $(X_n) \overset{prob.}{\to} X$.
- *If* $(X_n) \overset{\text{a.s.}}{\to} X$, *then* $(X_n) \overset{prob.}{\to} X$.
- *if* $(X_n) \overset{L^q}{\to} X$ *for some* q, *then* $(X_n) \overset{L^p}{\to} X$ *for every* $p < q$.

In short:



Proof of $\left(\overset{L^p}{\to} \Rightarrow \overset{\textbf{prob.}}{\to}\right)$**:** Assume that $(X_n) \overset{L^p}{\to} X$ and fix $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(|X_n - X| > \varepsilon\right) &= \mathbb{P}\left(|X_n - X|^p > \varepsilon^p\right), && \text{(this is the same event)} \\ &\leq \frac{\mathbb{E}[\ |X_n - X|^p\ ]}{\varepsilon^p}, && \text{(by Markov's inequality)} \end{aligned}$$

which goes to zero by assumption (ε is fixed and $n \to +\infty$). ■

Proof of $\left(\overset{\textbf{a.s.}}{\to} \Rightarrow \overset{\textbf{prob.}}{\to}\right)$**:** This part is admitted. ■

Proof of $\left(\overset{L^q}{\to} \Rightarrow \overset{L^p}{\to}\right)$**:** Recall (see page 19) that for $p < q$:

$$\|X_n - X\|_p \leq \underbrace{\|X_n - X\|_q}_{\to 0}.$$

If the right-hand side goes to zero, then so does the left-hand side. ■

The following proposition is sometimes useful:

Proposition 7.3 (*Convergence preservation*)

Let $(X_n), (Y_n)$ *be sequences of random variables on the same sample space* Ω.

- *Let* $g : \mathbb{R} \to \mathbb{R}$ *be a continuous function. If* $X_n \overset{prob.}{\to} X$ *then* $g(X_n) \overset{prob.}{\to} g(X)$.
- *If* $X_n \overset{prob.}{\to} X$, $Y_n \overset{prob.}{\to} Y$ *then* $X_n + Y_n \overset{prob.}{\to} X + Y$ *and* $X_n Y_n \overset{prob.}{\to} XY$.

7.2 Law(s) of Large Numbers

Basically, the *Law of Large Numbers* (LLN) says that the average of i.i.d. random variables $X_1, X_2, \ldots, X_n$ gets closer and closer to $\mathbb{E}[X]$. We need a precise statement.

Theorem 7.4 (*The weak Law of Large Numbers*)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. integrable random variables with expectation $\mu = \mathbb{E}[X_1]$ and such that $\mathrm{Var}(X_1) < +\infty$,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \overset{n\to+\infty}{\to} \mu,$$

where the convergence holds in L^2, and thus also in probability.

The "weak" refers to the fact that the convergence only holds in L^2 and in probability, we will see just below a "strong" LLN.

Proof: Set $S_n = X_1 + X_2 + \cdots + X_n$ and recall that $\mathbb{E}[S_n/n] = \frac{\mathbb{E}[X_1]+\cdots+\mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu$. We have to prove that the following goes to zero:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{S_n}{n} - \mu\right)^2\right] &= \mathbb{E}\left[\left(\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right)^2\right] \\
&= \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(S_n) && \text{(recall formula (\$) page 16)} \\
&= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right) && \text{(by independence)} \\
&= \frac{1}{n^2} n \mathrm{Var}(X_1) = \frac{1}{n}\mathrm{Var}(X_1) \to 0,
\end{aligned}$$

and the L^2 convergence is proved. ■

Example: We toss a fair coin infinitely many times, and denote by H_n the number of *Heads* seen in the first n tosses. Then

$$\mathbb{P}\left(0.49n \le H_n \le 0.51n\right) = \mathbb{P}\left(\left|\tfrac{H_n}{n} - 0.5\right| \le 0.01\right) \to 1,$$

applying convergence in probability in the weak LLN with $\varepsilon = 0.01$.

Theorem 7.5 (*The strong Law of Large Numbers*)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. integrable random variables with expectation μ,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \overset{\text{a.s.}}{\to} \mu.$$

In fact, the strong LLN is a much deeper result than the weak one (and with less assumptions since X_n's do not need to have a finite variance.) and we omit the proof.

Remark : Let us turn back to the example of coin tosses. The strong LLN shows that the frequency $\frac{H_n}{n}$ of Heads converges to 1/2, almost surely. Here you may see the meaning of "almost surely": $\Omega = \{H, T\}^{\mathbb{N}}$, and there are $\omega \in \Omega$ such that $\frac{H_n(\omega)}{n}$ does not go to 1/2, for instance

$$\omega = (H, H, H, H, H, \dots),$$

for which $\dfrac{H_n(\omega)}{n} = 1$. The strong LLN shows that such ω's form a set of measure zero.

8 Convergences of distributions

We now discuss a very different kind of convergence: convergence of distributions of random variables instead of convergence of random variables themselves. Let $X_1, X_2, \dots$ be random variables (unlike in the previous section, they may be defined in different probability spaces).

Definition 8.1

One says that $(X_n)_{n\geq 1}$ converges in distribution *(or* in law*) to X if for every bounded and continuous function ϕ we have*

$$\lim_{n\to+\infty} \mathbb{E}[\phi(X_n)] = \mathbb{E}[\phi(X)].$$

One writes $(X_n) \overset{(d)}{\to} X$ or $(X_n) \overset{(law)}{\to} X$.

A few words about notations: *This kind of convergence is very different in that it regards laws rather than random variables. To see why, observe that if $X_1, X_2, \dots$ are identically distributed, then for all n, X_n has the same law as X_1 and as X_2, so that*

$$(X_n) \overset{(d)}{\to} X_1 \text{ but also } (X_n) \overset{(d)}{\to} X_2,$$

and you see that the limit is not unique. In fact, it would be more appropriate to write that the law *of X_n converges to* the law *of X: you will sometimes find the notation*

$$\mathbb{P}_{X_n} \overset{(d)}{\to} \mathbb{P}_X;$$

and now the limit is unique.
One also says that $\mathbb{P}_{X_n}$ converges weakly *to $\mathbb{P}_X$.*

Convergence in distribution is in fact the "weakest" of all kinds of convergence:

Proposition 8.2

Let X and $(X_n)_{n\geq 1}$ be random variables defined on the same probability space. If $(X_n) \overset{prob.}{\to} X$ then $(X_n) \overset{(d)}{\to} X$.

Proof : (You may skip the proof).

Take a continuous function ϕ bounded by some $A > 0$. For a sake of simplicity, we will assume furthermore that ϕ is not only bounded and continuous but also Lipschitz: there exists $c > 0$ such that for all $x, y \in \mathbb{R}$

$$|\phi(x) - \phi(y)| \leq c|x - y|.$$

Fix $\varepsilon > 0$ and write

$$\begin{array}{rcccl}
|\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X)]| & \leq & \mathbb{E}[|\phi(X_n) - \phi(X)|] & & \\
& = & \mathbb{E}\Big[\underbrace{|\phi(X_n) - \phi(X)|}_{\leq c|X_n - X| \text{ since } \phi \text{ is Lip.}} \mathbb{1}_{|X_n - X| \leq \varepsilon} \Big] & + & \mathbb{E}\Big[\underbrace{|\phi(X_n) - \phi(X)|}_{\leq |\phi(X_n)| + |\phi(X)| \leq 2A} \mathbb{1}_{|X_n - X| > \varepsilon} \Big] \\
& \leq & \mathbb{E}\left[c|X_n - X| \mathbb{1}_{|X_n - X| \leq \varepsilon} \right] & + & \mathbb{E}\left[2A \mathbb{1}_{|X_n - X| > \varepsilon} \right] \\
& \leq & \mathbb{E}[c\varepsilon \mathbb{1}_{|X_n - X| \leq \varepsilon}] & + & 2A\mathbb{E}\left[\mathbb{1}_{|X_n - X| > \varepsilon} \right] \\
& \leq & c\varepsilon & + & 2A\mathbb{P}(|X_n - X| > \varepsilon),
\end{array}$$

and the last probability goes to zero by assumption. This proves that

$$\lim_{n \to +\infty} |\mathbb{E}[\phi(X_n)] - \mathbb{E}[\phi(X)]| \leq c\varepsilon$$

for any $\varepsilon > 0$, so the limit is zero. ■

In practice, we often do not compute $\mathbb{E}[\phi(X_n)]$, but rather use one of the two following criteria:

Theorem 8.3

The three following conditions are equivalent:

1. $(X_n) \overset{(d)}{\to} X$,
2. $F_{X_n}(t) \to F_X(t)$ *for every real* t *such that* F_X *is continuous at* t,
3. $\Phi_{X_n}(t) \to \Phi_X(t)$ *for every real* t.

We use items 2. or 3. in the Theorem according to how much it is easy to compute $F_{X_n}(t)$ or $\Phi_{X_n}(t)$.

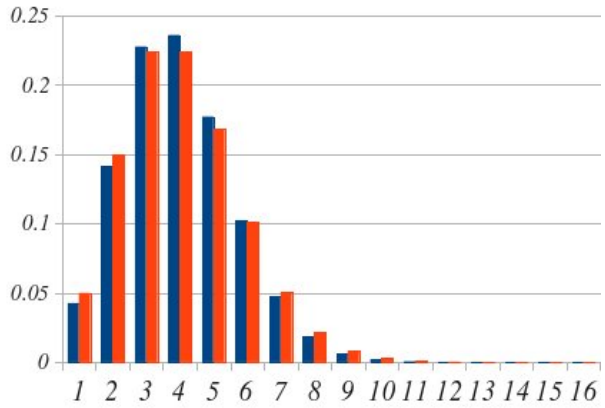
As an application, let us prove the *Law of rare events*: a sum of independent Bernoulli random variables with a small parameter is approximately distributed as a Poisson random variable.

Proposition 8.4 (*The Law of rare events*)

Let $\lambda > 0$,

$$\text{Binom}(n, \lambda/n) \overset{(d)}{\to} \text{Poisson}(\lambda).$$

This proposition partly explains why the Poisson distribution is so interesting for modeling.



Probabilities of the number of successes in 30 *Bernoulli trials with* 10% *success are well approximated by the* Poisson(3) *(left:* Binom$(30, 3/30)$*, right:* Poisson(3)*).*

Proof of Proposition 8.4 : Let $B_n \sim \text{Binom}(n, \lambda/n)$, we use characteristic functions.

$$\begin{aligned}\mathbb{E}[\exp(itB_n)] = \sum_{k=0}^{n} e^{itk}\mathbb{P}(B_n = k) &= \sum_{k=0}^{n} e^{itk}\binom{n}{k}(\tfrac{\lambda}{n})^k(1-\tfrac{\lambda}{n})^{n-k} \\ &= \sum_{k=0}^{n}\binom{n}{k}(e^{it}\tfrac{\lambda}{n})^k(1-\tfrac{\lambda}{n})^{n-k} \\ &= \left(1-\tfrac{\lambda}{n}+e^{it}\tfrac{\lambda}{n}\right)^n \text{ (by the binomial identity).}\end{aligned}$$

Now, for each t

$$\mathbb{E}[\exp(itB_n)] = \left(1+\frac{\lambda e^{it}-\lambda}{n}\right)^n \to \exp\left(\lambda e^{it}-\lambda\right),$$

where we used $(1+u/n)^n \to \exp(u)$, which is true for real and also (but this is not so easy to prove) for complex numbers. Now it remains to prove that $\exp(\lambda e^{it}-\lambda)$ is the CF of a r.v. X having the Poisson distribution with parameter λ:

$$\mathbb{E}[\exp(itX)] = \sum_{k\geq 0} e^{itk}e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k\geq 0} e^{-\lambda}\frac{(e^{it}\lambda)^k}{k!} = \exp\left(\lambda e^{it}-\lambda\right).$$

■

▷ Convergence preservation

Here are two useful properties regarding convergence in distribution.

Proposition 8.5 (*Convergence of a function of X_n*)

Let (X_n) be a sequence of random variables, and let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. If $X_n \overset{(d)}{\to} X$ then $g(X_n) \overset{(d)}{\to} g(X)$.

Proposition 8.6 (*Slutsky's Lemma*)

Let $(X_n), (Y_n)$ be sequences of random variables, and let $c \in \mathbb{R}$ be a constant. Assume that $X_n \overset{(d)}{\to} X$, $Y_n \overset{prob.}{\to} c$, then

$$(X_n, Y_n) \overset{(d)}{\to} (X, c).$$

In particular, this implies

$$(X_nY_n) \overset{(d)}{\to} Xc, \qquad (X_n + Y_n) \overset{(d)}{\to} X + c, \qquad \left(\frac{X_n}{Y_n}\right) \overset{(d)}{\to} X/c.$$

8.1 The Central Limit Theorem

Let $X_1, X_2, \dots$ be i.i.d. random variables with zero mean and finite variance σ^2. What can we say about $S_n = X_1 + X_2 + \cdots + X_n$ when n is large?
We know that $\frac{S_n}{n}$ goes to zero almost surely, but we still don't know if S_n is of order $\sqrt{n}, \sqrt[3]{n}, \log(n), \dots$

We can make an educated guess. Let $\alpha > 0$ and let us compute[(ii)]

$$\mathbb{E}\left[\left(\frac{S_n}{n^\alpha}\right)^2\right] = \frac{1}{n^{2\alpha}}\mathbb{E}[(S_n)^2] = \frac{1}{n^{2\alpha}}\mathrm{Var}(S_n) = \frac{1}{n^{2\alpha}}n\mathrm{Var}(X_1) = \frac{\sigma^2}{n^{2\alpha-1}}.$$

This goes to zero if $2\alpha - 1 > 0$, to infinity if $2\alpha - 1 < 0$, then something interesting seems to happen for $\alpha = 1/2$, *i.e.* for the sequence $S_n/\sqrt{n}$.

Theorem 8.7 (*The Central Limit Theorem*)

Let $(X_n)_{n\geq 1}$ be i.i.d. random variables with finite variance. Set $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathrm{Var}(X_1)$,

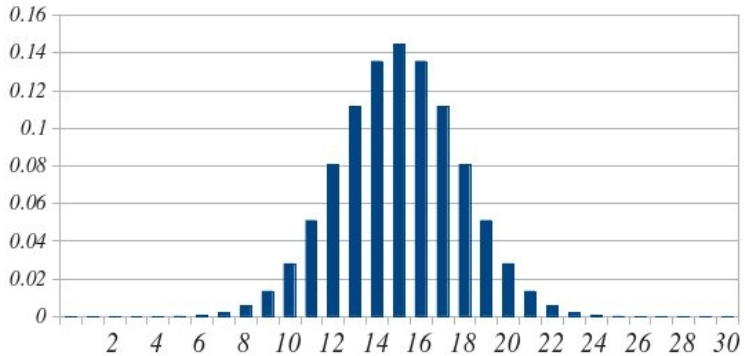
$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \overset{(d)}{\to} \mathcal{N}(0,1),$$

where $S_n = X_1 + X_2 + \cdots + X_n$.

One can (loosely) interpret the Theorem as

$$S_n \approx n\mu + \sigma\sqrt{n}Z,$$

where $Z \sim \mathcal{N}(0,1)$: the CLT says that S_n has *gaussian fluctuations* around is mean $n\mu$.



Probabilities of a Binomial$(30, 1/2)$.

[(ii)] This approach should remind you of the proof of the weak LLN.

Proof of the Central Limit Theorem : For simplicity, we make the proof in the particular case where

$$\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = 1/2,$$

we have $\mu = 0$, $\sigma^2 = 1$. The proof with an arbitrary distribution for the X_n's is very similar. We will prove that for all t

$$\Phi_{S_n/\sqrt{n}}(t) \to \exp(-t^2/2),$$

since $\exp(-t^2/2)$ is the characteristic function of a $\mathcal{N}(0,1)$.

$$\begin{aligned}
\Phi_{S_n/\sqrt{n}}(t) = \mathbb{E}\left[e^{itS_n/\sqrt{n}}\right] &= \mathbb{E}\left[e^{it\frac{(X_1+\cdots+X_n)}{\sqrt{n}}}\right] \\
&= \mathbb{E}\left[e^{it\frac{X_1}{\sqrt{n}}} \times \cdots \times e^{it\frac{X_n}{\sqrt{n}}}\right] \\
&= \mathbb{E}\left[e^{it\frac{X_1}{\sqrt{n}}}\right]^n \qquad \text{(by independence).}
\end{aligned}$$

Now,

$$\mathbb{E}\left[e^{it\frac{X_1}{\sqrt{n}}}\right] = \mathbb{P}(X_1 = +1) \times e^{it\frac{+1}{\sqrt{n}}} + \mathbb{P}(X_1 = -1) \times e^{it\frac{-1}{\sqrt{n}}} = \frac{e^{it/\sqrt{n}} + e^{-it/\sqrt{n}}}{2},$$

and recall that $e^{it} = \cos(t) + i\sin(t)$, so that $\frac{e^{it}+e^{-it}}{2} = \cos(t)$. Thus

$$\mathbb{E}\left[e^{it\frac{X_1}{\sqrt{n}}}\right] = \cos\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \mathrm{o}(1/n)$$

(recall $\cos(u) = 1 - u^2/2 + \mathrm{o}(u^2)$). Finally,

$$\begin{aligned}
\Phi_{S_n/\sqrt{n}}(t) &= \left(1 - \frac{t^2}{2n} + \mathrm{o}(1/n)\right)^n \\
&= \exp\left(n \log\left(1 - \frac{t^2}{2n} + \mathrm{o}(1/n)\right)\right) \\
&= \exp\left(n\left(-\frac{t^2}{2n} + \mathrm{o}(1/n)\right)\right) \qquad \text{(since } \log(1+u) = u + \mathrm{o}(u)\text{)} \\
&\overset{n\to+\infty}{\to} \exp(-t^2/2) = \Phi_Z(t).
\end{aligned}$$

■

8.2 Confidence intervals

▷ Asymptotic confidence interval given by the CLT

Let us flip n times an unfair coin that turns Heads with probability p. Let H_n be the number of Heads in the first n tosses, $H_n \sim \mathrm{Binom}(n,p)$ and we can write

$$H_n = X_1 + \cdots + X_n,$$

where X_k's are i.i.d. with $\mathbb{P}(X_k = 1) = p$, $\mathbb{P}(X_k = 0) = 1 - p$. Do check that $\mathbb{E}[X_k] = p$, $\mathrm{Var}(X_k) = p(1-p)$. Then the CLT says that

$$\frac{H_n - np}{\sqrt{p(1-p)}\sqrt{n}} \overset{(\mathrm{d})}{\to} Z,$$

where $Z \sim \mathcal{N}(0,1)$. Item (ii) in Theorem 8.3 says then that, for any reals a, b,

$$\mathbb{P}\left(a \leq \frac{H_n - np}{\sqrt{p(1-p)}\sqrt{n}} \leq b\right) \overset{n\to+\infty}{\to} \mathbb{P}(a \leq Z \leq b),$$

since $F_Z(t)$ is continuous for every t. Take for instance a such that $\mathbb{P}(-a \leq Z \leq a) = 95\%$ (with a computer one finds $a \approx 1.96$), this formula rewrites

$$\mathbb{P}\left(\frac{H_n}{n} \in \left[p \pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) = \mathbb{P}\left(-1.96 \leq \frac{H_n - np}{\sqrt{p(1-p)}\sqrt{n}} \leq 1.96\right) \overset{n\to+\infty}{\to} 95\%.$$

Observe now that for $0 < p < 1$, one has $p(1-p) \leq 1/4$, so that $1.96\sqrt{p(1-p)} \leq 1.96/2 < 1$. Then

$$\left[p \pm \frac{1}{\sqrt{n}}\right] \supset \left[p \pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right], \text{ and thus}$$

$$\mathbb{P}\left(\frac{H_n}{n} \in \left[p \pm \frac{1}{\sqrt{n}}\right]\right) \geq \mathbb{P}\left(\frac{H_n}{n} \in \left[p \pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \overset{n\to+\infty}{\to} 95\%.$$

$\Rightarrow$ *With more than* 95% *chance, the frequency of Heads after* n *flips is close to* p *within* $\frac{1}{\sqrt{n}}$.

Let us now reverse the question: imagine that p is unknown, this tells us that H_n/n is a good estimation of p. Obviously we have

$$\frac{H_n}{n} \in \left[p \pm \frac{1}{\sqrt{n}}\right] \Leftrightarrow \left|\frac{H_n}{n} - p\right| \leq \frac{1}{\sqrt{n}} \Leftrightarrow p \in \left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right].$$

One says that $[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}]$ is a 95% *confidence interval* for p.

Definition/Theorem 8.8 (*Asymptotic confidence interval for the Binomial*)

Let $H_n \sim \mathrm{Binom}(n,p)$.
The interval $\left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right]$ *is a* confidence interval *for* p *with asymptotic confidence larger than* 95%*:*

$$\lim_{n\to\infty} \mathbb{P}\left(\left[\frac{H_n}{n} \pm \frac{1}{\sqrt{n}}\right] \ni p\right) \geq 0.95.$$

Note that in the proposition the interval depends on H_n/n and therefore is random.

▷ Exact confidence interval: using Hoeffding's inequality

Recall first the Hoeffding inequality (p. 47): since $X_1, \dots, X_n$ are i.i.d. and bounded by $0 \le X_1 \le 1$ then for every real $A > 0$, we have

$$\mathbb{P}\left(|H_n - np| \ge A\right) \le 2\exp\left(-\frac{2A^2}{n(1-0)^2}\right) = 2\exp\left(-2A^2/n\right).$$

Therefore we obtain

$$\begin{aligned}\mathbb{P}\left(\frac{H_n}{n} \notin \left[p \pm \frac{a}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(|H_n - np| \ge n\frac{a}{\sqrt{n}}\right) \\ &\le 2\exp\left(-2(\sqrt{n}a)^2/n\right) = 2\exp\left(-2a^2\right),\end{aligned}$$

where we used Hoeffding's inequality with $A = \sqrt{n}a$. Now, take a such that $2\exp\left(-2a^2\right) = 0.05$, *i.e.* $a \approx 1.3581\ldots$ one obtains the following result:

Definition/Theorem 8.9 (*Exact confidence interval for the Binomial*)

Let $H_n \sim \mathrm{Binom}(n, p)$.
The interval $\left[\frac{H_n}{n} \pm \frac{1.3581}{\sqrt{n}}\right]$ *is an* exact confidence interval *for* p *with confidence larger than* 95%*:*

$$\forall n \ge 1, \qquad \mathbb{P}\left(\ \left[\frac{H_n}{n} \pm \frac{1.3581}{\sqrt{n}}\right] \ni p\ \right) \ge 0.95.$$

It seems at first sight that the result is weaker than Theorem 8.8 (since the interval is larger) but it is valid for every n.

Index