MA1AY010 Class Notes

Master I ISIFAR

20025-09-05

Table of contents

Warm up

A short undergraduate course on Probability theory in an applied mathematics curriculum

This *Probability* course is a core part of `Master ISIFAR`, a curriculum that delivers training in Statistics, Mathematical Finance, and Computer Science at Université Paris Cité. The *Probability* course is related to other courses from the same curriculum.

1. Statistique M1
2. Mathématiques financières M1
3. Mathématiques financières M2
4. Apprentissage statistique M2

This list is not exhaustive.

This course is geared towards applications. We borrow examples and applications of probability theory from statistics, computer science, big data engineering. As we have a limited amount of time, we sweep a lot of dust under the carpet. We take for granted key results from integration and measure theory. Nevertheless, we build on rigorous definitions, state and invoke theorems in a consistent way.

- 🔗 Homepage of the course.
- 🔗 Moodle page of the course.

Prerequisites

This course builds on two Licence-level courses:

- Probabilités Licence 3
- Intégration

Chapter 1

Introduction

In this chapter we survey the basic definitions of Probability Theory starting from a simple modeling problem from computer science. The notions are formally defined in next chapters. The simple context allows us to carry out computations and to outline the kind of results we will look for during the course: moments, tail bounds, law of large numbers, central limit theorems, and possibly other kind of weak convergence results.

1.1 Hashing

Hashing is a computational technique that is used in almost every area of computing, from databases to compilers through (big) datawarehouses. Every book on algorithms contain a discussion of hashing, see for example `Introduction to Hashing by Jeff Erickson`.

Under idealized conditions, hashing n items to m values consists of applying a function picked uniformly at random among the m^n functions from $1, \ldots, n$ to $1, \ldots, m$. The performance of a hashing method (how many cells have to be probed during a search operation?) depends on the typical properties of such a random function.

It is convenient to think of the values in $1, \ldots, m$ as numbered bins and of the items as n numbered balls. Picking a random function amounts to throw independently the n balls into the m bins. The probability that a given ball falls into a given bin is $1/m$.

Questions around the random functions can be rephrased.

- How many empty bins on average?
- Distribution of the number of empty bins?
- How many bins with r balls?
- What is the maximum number of balls in a single bin?

Have a look at the http://stephane-v-boucheron.fr/post/2019-09-02-idealizedhashing/ and download the notebook from there.

This toy yet useful model is an opportunity to recall basic notions of probability theory.

In the sequel, we call this framework the ***random alllocations*** experiment.

Table 1.1: An outcome of the random allocation experiment with n=10 and m=5

	1	2	3	4	5	6	7	8	9	10
ω	2	3	1	2	4	4	2	3	2	5

In Table 1.1, line ω represents the outcome of a random allocation with $n = 10, m = 5$: $\omega_4 = 2, \omega_5 = 4, \ldots$

1.2 A Probability space

The set of outcomes is called the ***universe***. In the random allocations setting it is the set of $1, \ldots, m$-valued sequences of length m. This is also a function mapping $\{1, \ldots, n\}$ to $\{1, \ldots, m\}$. We denote a generic outcome by ω. The i^{th} element of ω is denoted by ω_i. This universe is denoted by Ω, here it is finite with cardinality m^n.

In this simple setting, the uniform probability distribution on the universe assigns to each subset A of Ω the probability $|A|/|\Omega|$. When the universe is finite or countable, all subsets of the universe are ***events***, assigning a probability to every subset of the universe is not an issue.

Recall that a ***probability distribution*** P maps a collection $\mathcal{F}$ of subsets of the universe ($\mathcal{F} \subseteq 2^\Omega$) to $[0, 1]$ and satisfies:

1. $P(\emptyset) = 0$
2. $P(\Omega) = 1$
3. for any countable collection of pairwise disjoint events $A_1, A_2, \ldots, n, \ldots$, $P(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty P(A_n)$

See Section 2.3.

This entails $P(A_1 \cup A_2 \cup \ldots \cup A_k) = \sum_{i=1}^k P(A_i)$ for all finite collection of pairwise disjoint subsets $A_1, \ldots, A_k$.

For the domain of P to be well-defined, the collection of subsets $\mathcal{F}$ has to be closed under countable unions, countable intersections and complementation, to contain the empty set $\emptyset$ and the universe Ω. In words, it has to be a ***σ-algebra***, see Section 2.2.

Note that other probability distributions make sense on this simple universe. See for example the ***balanced allocations*** scenario.

In the ballanced allocations scenario, the random functions from $1, \ldots, n$ to $1, \ldots, m$ are constructed sequentially. We first construct ω_1 by picking a number uniformly at random from $1, \ldots, n$. Now, assume we have constructed $\omega_1, \ldots, \omega_i$ for some $i < n$. In order to determine ω_{i+1}, we pick uniformly at random two numbers from $1, \ldots, n$, say j and k. We compute

$$c_j = \Big|\{\ell : 1 \leq \ell \leq i, \omega_\ell = j\}\Big| \qquad \text{and} \qquad c_k = \Big|\{\ell : 1 \leq \ell \leq i, \omega_\ell = k\}\Big| .$$

If $c_j < c_k$, $\omega_{i+1} = j$ otherwise $\omega_{i+1} = k$.

This iterative construction defines a (unique) probability distribution over $\{1, \ldots, m\}^n$ that differs from the uniform probability distribution. It is non-trivial to show that it achieves a non-trivial balancing guarantee for the size of the preimages induced by ω.

1.3 Random variables and independence

Consider the real valued functions from Ω to $\mathbb{R}$ defined by:

$$X_{i,j}(\omega) = \begin{cases} 1 & \text{if } \omega_i = j \\ 0 & \text{otherwise} . \end{cases}$$

This function is a special case of a ***random variable*** see Section 2.5.

In the toy example outlined in Table 1.1, we have $X_{4,1}(\omega) = 1, X_{5,1}(\omega) = 0, \ldots$.

> Note that the definition of the random variable has nothing to do with the probability distribution we have considered so far. There is nothing random in a random variable. Moreover, a random variable is not a variable, it is a function. You may question this terminology, but it has been sanctified by tradition.

In the probability space $(\Omega, 2^\Omega, \text{Pr})$, the distribution of the random variable $X_{i,j}$ is a ***Bernoulli*** distribution with parameter $1/m$.

$$\text{Pr}\left\{X_{i,j} = 1\right\} = \frac{1}{m} \qquad \text{Pr}\left\{X_{i,j} = 0\right\} = 1 - \frac{1}{m}\,,$$

see Section 4.1 for more on Bernoulli distributions. This comes from

$$\text{Pr}\left\{\omega : X_{i,j}(\omega) = 1\right\} = \frac{\left|\{\omega : X_{i,j}(\omega) = 1\}\right|}{m^n} = \frac{m^{n-1}}{m^n} = \frac{1}{m}\,.$$

Recall that $\text{Pr}\left\{X_{i,j} = 1\right\}$ is a shorthand for $\text{Pr}\left\{\omega : X_{i,j}(\omega) = 1\right\}$.

For a while, we fix some $j \in \{1, \ldots, m\}$ and consider the collection of random variables $(X_{i,j})_{i \leq n}$.

For each i, we can define events (subsets of Ω) from the value of $X_{i,j}$:

$$\left\{\omega : X_{i,j}(\omega) = 1\right\}$$
$$\left\{\omega : X_{i,j}(\omega) = 0\right\}$$

and together with $\Omega, \emptyset$ they form the collection $\sigma(X_{i,j})$ of events that are definable from $X_{i,j}$.

Recall the definition of ***independent events*** or rather the definition of a ***collection of independent events***.

A collection of events $E_1, E_2, \ldots, E_k$ from $(\Omega, 2^\Omega)$ is independent with respect to Pr if for all $I \subseteq \{1, \ldots, n\}$,

$$\text{Pr}\left\{\cap_{i \in I} E_i\right\} = \prod_{i \in I} \text{Pr}\{E_i\}$$

One can check that for each fixed $j \leq m$, $(X_{i,j})_{i \leq n}$ is a ***collection of independent random variables*** under Pr. By this we mean that each collection $E_1, E_2, \ldots, E_n$ of events where $E_i \in \sigma(X_{i,j})$ for each $i \in \{1, \ldots, n\}$, $E_1, E_2, \ldots, E_n$ is an independent collection of events under Pr.

The notion of ***independence*** is a cornerstone of probability theory, see Chapter **?@sec-independence**.

Concretely, this means that for any sequence $b_1, \ldots, b_n \in \{0, 1\}^n$ (a possible outcome for the sequence of random variables $X_{1,j}, X_{2,j}, \ldots, X_{n,j}$), we have

$$\begin{aligned}\Pr\left\{\bigwedge_{i=1}^{n} X_{i,j}(\omega)=b_i\right\} &= \prod_{i=1}^{n}\Pr\left\{X_{i,j}(\omega)=b_i\right\}\\ &= \prod_{i=1}^{n}\left(\frac{1}{m}\right)^{b_i}\left(1-\frac{1}{m}\right)^{1-b_i}\\ &= \left(\frac{1}{m}\right)^{\sum_{i=1}^{n} b_i}\left(1-\frac{1}{m}\right)^{n-\sum_{i=1}^{n} b_i}\ .\end{aligned}$$

Observe that the outcome of the sequence $X_{i,j}$ for $i \in 1,\ldots,n$ is $b_1,\ldots,b_n$ only depends on $\sum_{i=1}^{n} b_i = Y_j$. This greatly simplifies computations.

We are interested in the number of elements from $1,\ldots,n$ that are mapped (allocated) to j through the random function ω. Let us define

$$Y_j(\omega)=\sum_{i=1}^{n} X_{i,j}(\omega)\,.$$

In the toy example described in Table 1.1, $Y_3(\omega)=4$ while $Y_5(\omega)=1$ and $Y_4(\omega)=0$:

Table 1.2: Occupancy scores for the random allocation experiment with n=10 and m=5

j	1	2	3	4	5
Y_j	1	4	2	2	1

Occupancy scores

In the probability space $(\Omega, 2^{\Omega}, \Pr)$, the random variable Y_j is distributed as a sum of independent, identically distributed Bernoulli random variables, that is, according to a ***Binomial distribution***, see Section 4.1.

$$\Pr\left\{Y_j=r\right\}=\binom{n}{r}p^r(1-p)^{n-r}\qquad\text{with}\quad p=\frac{1}{m}$$

for $r \in 0,\ldots,n$.

Indeed, recall

$$\begin{aligned}\Pr\left\{Y_j=r\right\} &= \sum_{\omega:Y_j(\omega)=r}\Pr\left\{\omega\right\}\\ &= \sum_{\omega:Y_j(\omega)=r}\left(\frac{1}{m}\right)^{r}\left(1-\frac{1}{m}\right)^{n-r}\\ &= \left|\left\{\omega:\omega\in\Omega, Y_j(\omega)=r\right\}\right|\times\left(\frac{1}{m}\right)^{r}\left(1-\frac{1}{m}\right)^{n-r}\\ &= \binom{n}{r}\left(\frac{1}{m}\right)^{r}\left(1-\frac{1}{m}\right)^{n-r}\ .\end{aligned}$$

For large n, m, this Binomial distribution tends to be concentrated around its mean value or ***expectation***

$$\mathbb{E} Y_j = \sum_{r=0}^{n} r \times \Pr\left\{ Y_j = r \right\} = \frac{n}{m} \, .$$

See Chapter 3 for a systematic approach to expectation, variance and higher moments, based on Integration theory.

The last chapter **?@sec-chapConcentration** is dedicated the development of *tail bounds* for random variables like Y_j that are *smooth functions of independent random variables*.

For the moment recall that on a countable probability space, the expectation of random variable Z can be defined as

$$\mathbb{E} Z = \sum_{\omega \in \Omega} \Pr\{\omega\} \times Z(\omega)$$

provided the series is absolutely convergent.

This is illustrated by Figure 1.1. In principle, a binomial random variable with parameters $n = 5000$ and $p = .001$ can take any value between 0 and 5000. However, most (more than 95%) of the probability mass is supported by $\{1, \ldots, 10\}$.

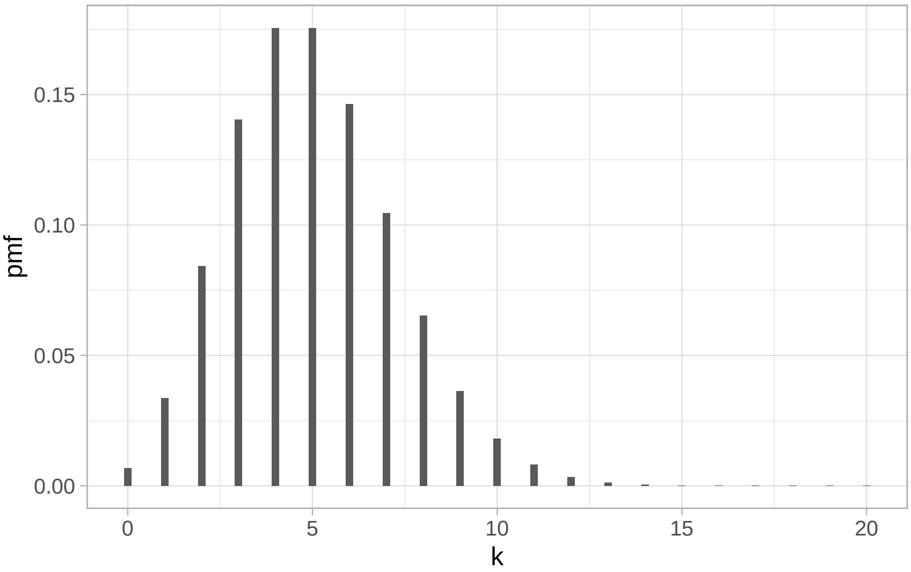

Figure 1.1: Probability mass function of Binomial(5000,0.001)

1.4 Convergences

If we let n, m tend to infinity while n/m tends toward $c > 0$, we observe that, for each fixed $r \geq 0$ the sequence $\Pr\left\{ Y_j = r \right\} = \binom{n}{r} (1/m)^r (1 - 1/m)^{n-r}$ tends towards

$$\mathrm{e}^{-c} \frac{c^r}{r!}$$

which is the probability that a Poisson distributed random variable with expectation $c$equals r (see Section 4.2 for more on Poisson distributions).

This is an instance of the *law of rare events*, a special case of *convergence in distribution* see Chapter 10.

The ability to approximate a Poisson distribution using an appropriate Binomial distribution is illustrated in Figure 1.2. The difference between the probability mass functions of the Binomial distributions with parameters $n = 250, m = 0.02$, and $n = 2500, m = 0.002$ and the Poisson distribution with parameter 5 is small. If we chose parameters $n = 2500, m = 0.002$, the difference between Binomial and Poisson is barely visible.

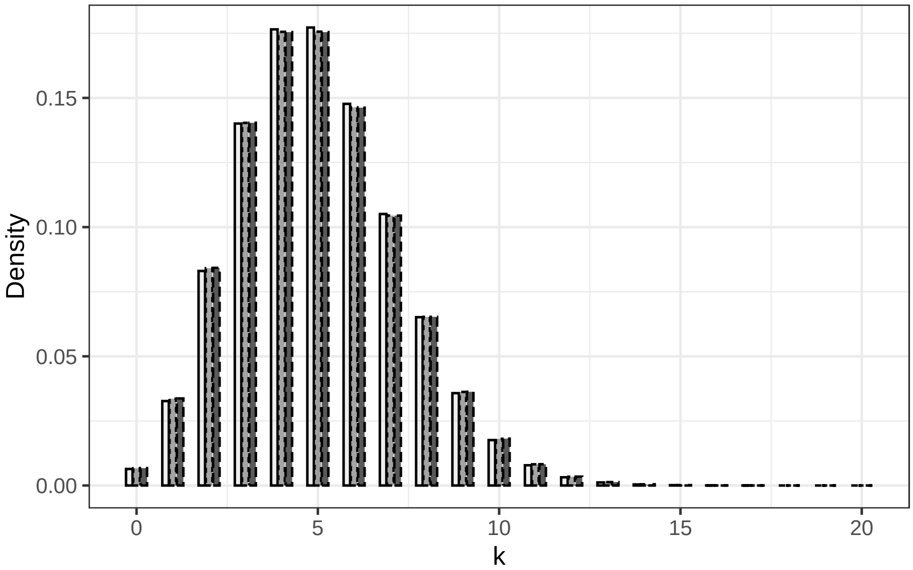

Figure 1.2: Probability mass functions of Binomial(250,0.02) (left), Binomial(2500,0.002) (middle) and Poisson(5) (right)

The proximity between Binomial$(n, \lambda/n)$ and Poisson(λ) can be quantified in different ways. A simple one consists in computing

$$\sum_{x \in \mathbb{N}} \left| p_{n,\lambda/n}(x) - q_\lambda(x) \right|$$

where $p_{n,\lambda/n}$ (resp. q_λ) stands for Binomial (resp. Poisson). This quantity is called the variation distance between the two probability distributions. A general definition is provided in Chapter 10. In Figure 1.3, this distance between Binomial distribution with parameters $n, 5/n$ and Poisson(5) is plotted against n (beware logarithmic scales). This plot suggests that the variation distance decays like $1/n$. This is checked in Chapter 10.

In the probability space $(\Omega, 2^\Omega, \Pr)$, the random variables $Y_j, Y_j', j \neq j'$ are not independent. In order to show that $Y_j, Y_j', j \neq j'$ are not independent, it suffices to check that two events $E_j, E_{j'}$ are not independent with $\omega \in E_j$ being a function of Y_j and $\omega \in E_{j'}$ being a function of $Y_{j'}$ (later, we will concisely say $E_j \in \sigma(X_j)$ or E_j being Y_j-measurable). Choose $E_j = \{\omega : Y_j(\omega) = r\}$ and $E_{j'} = \{\omega : Y_{j'}(\omega) = r\}$.

$$\Pr(E_j) = \binom{n}{r} \left(\frac{1}{m}\right)^r \left(1 - \frac{1}{m}\right)^{n-r}$$

$$\Pr(E_j \cap E_{j'}) = \binom{n}{r} \times \binom{n-r}{r} \left(\frac{1}{m}\right)^{2r} \left(1 - \frac{2}{m}\right)^{n-2r}$$

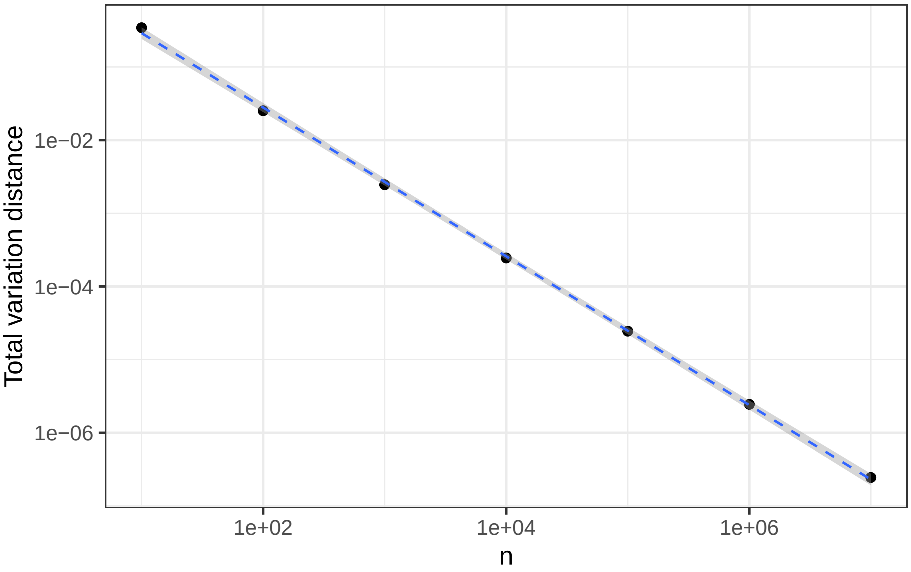

Figure 1.3: Law of rare events: distance between Binomial($n, 5/n$) and Poisson(5) as a function of n

$$\frac{\Pr(E_j \cap E_{j'})}{\Pr(E_j) \times \Pr(E_{j'})} = \frac{\left(1-\frac{2}{m}\right)^{n-2r}}{\left(1-\frac{1}{m}\right)^{2n-2r}} \frac{((n-r)!)^2}{n!(n-2r)!} \neq 1\,.$$

Hence, if we define

$$K_{n,r}(\omega) = \sum_{j=1}^{m} \mathbb{I}_{Y_j(\omega)=r}$$

as the number of elements of $1, \ldots, m$ that occur exactly r times in ω, the random variable $K_{n,r}$ is not described as a sum of independent random variables. Nevertheless, it is possible to gather a lot of information about its moments and distribution. If we let again n, m tend to infinity while n/m tends toward $c > 0$, we observe that the distribution of $K_{n,r}/m$ tends to concentrate around $\mathrm{e}^{-c}\frac{c^r}{r!}$. This is an example of ***convergence in probability***, see Chapter 9.

Now, if we consider the sequence of recentered and rescaled random variables $(K_{n,r} - \mathbb{E}K_{n,r})/\sqrt{\operatorname{var}(K_{n,r})}$, we observe that its ***distribution function*** (see Section 2.7) converges pointwise towards the distribution function of the Gaussian distribution.

Table 1.3: Profile of example 2.1, count of empty urns is omitted

$K_{n,1}$	$K_{n,2}$	$K_{n,4}$
2	2	1

1.5 Summary

In this chapter, we investigated a toy stochastic model: ***random allocations***. This toy model was motivated by the analysis of hashing, a widely used technique from Computer science. To perform the analysis, we introduced notation and notions from probability theory:

- Universe,
- Events,
- σ-algebras,
- Probability distributions,
- Preimages,
- Random variables,
- Expectation,
- Variance,
- Independence of events,
- Independence of random variables,
- Binomial distribution,
- Poisson distribution.

Through numerical simulations, we got a feeling of several important phenomena:

- Law of rare events: approximation of Poisson distribution by certain Binomial distributions.
- Law of large numbers for normalized sums of identically distributed random variables that are not independent.
- Central limit theorems for normalized and centered sums of identically distributed random variables that are not independent

At that point, our elementary approach did not provide us with the notions and tools that make possible the rigorous analysis of these phenomena.

Chapter 2

A modicum of measure theory

2.1 Roadmap

Performing stochastic modeling in a comfortable way requires consistent foundations and notation. In this chapter, we set the stage for further development. Probability theory started as the interaction between combinatorics and games of chance (XVIIth century). At that time, the set of outcomes was finite, and it was legitimate to think that any set of outcomes had a well-defined probability. When mathematicians started to perform stochastic modeling in different branches of sciences (astronomy, thermodynamics, genetics, ...), they had to handle uncountable sets of outcomes. Designing a sound definition of what a probability distribution is, took time. Progress in integration and measure theory during the XIXth century and the early decades of the XXth century led to the modern, measure-theoretical foundation of probability theory.

2.2 Universe, powerset and σ- algebras

A ***universe*** is a set (of possible ***outcomes***) we decide to call a universe. The universe is often denoted by Ω. Generic elements of Ω (outcomes) are denoted by ω.

Example 2.1. If we think of throwing a dice as a random phenomenon, the set of outcomes is the set of labels on the faces $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we are throwing two dices, the set of outcomes is made of couples of labels $\Omega' = \{(1, 1), (1, 2), (1, 3), \ldots , (6, 6)\} = \Omega^2$.

Example 2.2. In the idealized hashing problem (Section 1.1), the universe is the set of functions from $1, \ldots , n$ to $1, \ldots , m$. The size of the universe is m^n.

A universe may or may not be finite or countable. If the universe is countable, all its subsets may be called ***events***. Events are assigned probabilities. If the universe is countable, it is possible to assign a probability to each of its subsets. When the universe is not countable (for example $\mathbb{R}$), Assigning a probability to all subsets is not possible. We have to restrict the collection of subsets in order to assign probabilities to the collection members in a consistent way.

In the sequel 2^Ω denotes the collection of all subsets of Ω (the powerset of Ω).

A sensible collection of events has to be a σ-algebra.

Definition 2.1 (σ-algebra”). Given a set Ω, a collection $\mathcal{G}$ of subsets of Ω ($\mathcal{G} \subseteq 2^\Omega$) is called a σ-algebra (a sigma algebra) iff

- $\mathcal{G}$ is closed under $countable$ union
- $\emptyset \in \mathcal{G}$
- $\mathcal{G}$ is closed under complementation ($A \in \mathcal{G} \Rightarrow A^c = \Omega \setminus A \in \mathcal{G}$)

What the smallest σ-algebra (with respect to set inclusion) that contains subset A of Ω?

The next proposition shows that σ-algebras are stable under countable set-theoretical operations. We could have replaced countable union by countable intersection in the definition of σ-algebras. This is consequence of De Morgan's laws:

$$(A \cup B)^c = A^c \cap B^c \qquad \text{and}(A \cap B)^c = A^c \cup B^c$$

A σ-algebra of subsets is closed under countable intersections.

Proof. For $A \subseteq \Omega$, let $A^c = \Omega \setminus A$. Let $A_1, \ldots, A_n, \ldots$ belong to σ-algebra $\mathcal{G}$ of subsets of Ω. For each n, $A_n^c \in \mathcal{G}$, by definition of σ-algebra,

$$\begin{aligned}\cap_n A_n &= \Big(\big(\cap_n A_n \big)^c \Big)^c \\ &= \Big(\cup_n A_n^c \Big)^c \qquad \text{De Morgan}\,.\end{aligned}$$

By definition of a σ-algebra, $\cup_n A_n^c \in \mathcal{G}$, and for the same reason, $\Big(\cup_n A_n^c \Big)^c \in \mathcal{G}$. □

The next proposition allows us to talk about the smallest σ-algebra containing a collection of subsets, this leads to the notion of generated σ-algebra.

The intersection of two σ-algebras of subsets of Ω is a σ-algebra of subsets of Ω.

Proof. Let $\mathcal{G}$ and $\mathcal{G}'$ be two σ-algebras of subsets of Ω. The intersection of the two σ-algebras is

$$\Big\{ A : A \subseteq \Omega, A \in \mathcal{G}, A \in \mathcal{G}' \Big\}\,.$$

□

Indeed, the intersection of a possibly uncountable collection of σ-algebras is a σ-algebra (check this). Because of this property, the notion of a σ-algebra generated by a collection of subsets is well-founded.

Generated σ- algebra

Given a collection $\mathcal{C}$ of subsets of Ω, there exists a ***unique smallest σ-algebra*** containing all subsets in $\mathcal{C}$, it is called the σ-algebra generated by $\mathcal{H}$ and denoted by $\sigma(\mathcal{C})$.

Exercise 2.1. Check the preceding proposition.

Example 2.3. Consider we are throwing a dice, $\Omega = \{1, \ldots, 6\}$, let

$$\mathcal{H} = \Big\{ \{1, 3, 5\} \Big\}\,.$$

This is a collection made of one event (the outcome is odd). The algebra generated by $\mathcal{H}$ is

$$\sigma(\mathcal{H}) = \Big\{ \{1, 3, 5\}, \{2, 4, 6\}, \emptyset, \Omega \Big\}\,.$$

Two kinds of σ-algebras play a prominent role in a basic probability course:

1. the powerset of countable or finite sets.
2. the Borel σ-algebras of topological spaces.

Definition 2.2 (Borel sigma-algebra)**.** The Borel σ-algebra over $\mathbb{R}$ is the σ-algebra generated by ***open*** sets. It is denoted by $\mathcal{B}(\mathbb{R})$.

This definition works for every topological space. Recall that a topology on a set E is defined by a collection $\mathcal{E}$ of ***open sets***. This collection is defined by the following list of properties:

- $\emptyset, E \in \mathcal{E}$
- A (***possibly uncountable***) union of elements of $\mathcal{E}$ (open sets) belongs to $\mathcal{E}$ (is an open set)
- A finite intersection of open sets is an open set.

In the usual topology on $\mathbb{R}$, a set A is open if for any $x \in A$, there exists some $r > 0$ such that $]x - r, x + r[\subseteq A$. Any interval of the form $]a, b[$ is open (these are the so-called open intervals).

This topology can be generalized to any finite dimension $\mathbb{R}^d$.

Exercise 2.2. Consider the σ-algebra generated by open-intervals of $\mathbb{R}$. Is it the Borel σ-algebra?

Exercise 2.3. Consider the σ-algebra generated by open-intervals of $\mathbb{R}$ with rational bounds. Is it the Borel σ-algebra?

Exercise 2.4. Consider any metric space (E, d). The metric d defines a topology on E. Does the Borel σ-algebra on (E, d) coincide with the σ-algebra generated by open balls $B(x, r) = \Big\{y : y \in E, d(x, y) < r\Big\}$?

We are now ready to set the stage of stochastic modeling. The playground always consists of a measurable space.

Definition 2.3 (Measurable space)**.** A universe Ω endowed with a σ-algebra of subsets $\mathcal{F}$ is called a measurable space. It is denoted by $(\Omega, \mathcal{F})$.

Example 2.4.

- If Ω is a countable or finite set, then $(\Omega, 2^\Omega)$ is a measurable space.
- If $\Omega = \mathbb{R}$, then $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable space.

So far, we have not talked about probability theory, but, we are now equipped to define probability distributions and to manipulate them.

2.3 Probability distributions

A ***probability distribution*** maps a σ-algebra to $[0, 1]$. It is an instance of a more general concept called a ***measure***. We state or recall important concept of measure theory. The key idea underneath the elaboration of measure theory is that we should refrain from trying to measure ***all*** subsets of a universe (unless this universe is countable). Attempts to measure all subsets of $\mathbb{R}$ lead to paradoxes and of little practical use. Measure theory starts by recognizing the desirable properties any useful measure should possess, then measure theory builds objects satisfying these properties on as large as possible σ-algebras of events, for example on Borel σ-algebras.

This motivates the definition of σ-additivity.

Definition 2.4 (Sigma-additivity)**.** Given Ω and $\mathcal{A} \subseteq 2^{\Omega}$, a function μ mapping $\mathcal{A}$ to $[0, \infty)$ is said to be ***σ-additive*** on $\mathcal{A}$ if for any countable collection of ***pairwise disjoint*** subsets $(A_n)_{n \in \mathbb{N}} \in \mathcal{A}$, with $\cup_n A_n \in \mathcal{A}$ we have

$$\mu(\cup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mu(A_n) .$$

Note that if $\mathcal{F}$ is a σ-algebra, $\left(\cup_{n \in \mathbb{N}} A_n \right) \in \mathcal{F}$. σ-additivity fits well with σ-algebras, but it makes sense to define σ-additivity with respect to more general collections of subsets.

Proposition 2.1. *Given Ω, a σ-algebra $\mathcal{A} \subseteq 2^{\Omega}$, a σ-_additive_function μ mapping $\mathcal{A}$ to $[0, \infty)$ satisfies*

a. for any increasing sequence $(A_n)_{n \in \mathbb{N}}$ of elements of $\mathcal{A}$

$$\lim_n \mu(A_n) = \mu\left(\cup_n A_n\right)$$

b. for any decreasing sequence $(A_n)_{n \in \mathbb{N}}$ of elements of $\mathcal{A}$

$$\lim_n \mu(A_n) = \mu\left(\cap_n A_n\right)$$

c.

$$\mu(\emptyset) = 0$$

Proof. a.) Let $B_1 = A_1$, and $B_{n+1} = A_{n+1} \setminus A_n$ for each n, then $(B_n)_n$ is a sequence of pairwise disjoints elements of $\mathcal{A}$. We have $\cup_n B_n = \cup_n A_n$ and and by σ-additivity, $\mu(\cup_n A_n) = \sum_n \mu(B_n)$

$$\sum_{n \leq m} \mu(B_n) = \mu\left(\cup_{n \leq m} B_m\right) = \mu(A_m)$$

Hence $\lim_{m \to \infty} \mu(A_m) = \sum_{n \in \mathbb{N}} \mu(B_n) = \mu(\cup_n A_n)$

b.) The second statement is proved in a similar way.

c.) Let $(A_n)_n$ be such that $A_n = \emptyset$ for each n, this is a sequence of pairwise disjoint elements of $\mathcal{A}$, by σ-additivity, we have

$$\sum_{n \in \mathbb{N}} \mu(\emptyset) = \mu(\emptyset)$$

which implies $\mu(\emptyset) = 0$.

□ □

Definition 2.5 (Positive measure)**.** Given a measurable space $(\Omega, \mathcal{F})$, a σ-additive function μ mapping $\mathcal{F}$ to $[0, \infty)$ is called a ***positive measure*** over $(\Omega, \mathcal{F})$.

The tuple $(\Omega, \mathcal{F}, \mu)$ is called a measure space.

By Proposition 2.1, for any positive measure μ, we have $\mu(\emptyset) = 0$. When $\mu(\Omega)$ is finite, μ is said to be ***finite*** positive measure.

Exercise 2.5. Let $\Omega = \{0,1\}^*$ the set of infinite sequences of 0 and 1 (indexed from 1). Let $\mathcal{F}_n \subseteq 2^\Omega$ be the σ-algebra generated by events of the following form: $\{\omega : \omega \in \Omega, \omega_i = 1\}$ for $1 \leq i \leq n$.

- Define a σ-additive function on $(\Omega, \mathcal{F}_n)$.
- What is the σ-algebra generated by $\cup_{n\geq 1}\mathcal{F}_n$?
- Can you define a σ-additive function on $(\Omega, \sigma(\cup_{n\geq 1}\mathcal{F}_n))$.

A positive measure μ is not necessarily a probability distribution. For example, the ***counting measure*** μ on $\mathbb{N}$ satisfies $\mu(A) = |A|$ for all $A \subseteq \mathbb{N}$, so we have $\mu(\mathbb{N}) = \infty$.

Definition 2.6 (Probability distribution)**.** Given a measurable space $(\Omega, \mathcal{F})$, a function μ mapping $\mathcal{F}$ to $[0,\infty)$ is a probability distribution over $(\Omega, \mathcal{F})$ if

1. μ is a positive measure on $(\Omega, \mathcal{F})$ and
2. $\mu(\Omega) = 1$.

Exercise 2.6. If you think $(\mathbb{R}, 2^{\mathbb{R}})$ is a measurable space, define a σ-additive measure on it. Try even to define a probability measure.

Remark 2.1. The notion of σ-additivity is strictly stronger than finite additivity. Assuming the ***Axiom of Choice*** (as usual when working in Analysis or Probability), there exists a function μ that map $2^{\mathbb{N}}$ to $[0,1]$, that is additive ($\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B, A \cap B = \emptyset$), zero on all finite subsets of $\mathbb{N}$ and such that $\mu(\mathbb{N}) = 1$. Such a function is not σ-additive.

2.4 Lebesgue measure

We take the existence of Lebesgue's measure for granted. This is the content of the next theorem.

Theorem 2.1 (Existence of Lebesgue's measure)**.** *There exists a unique σ-additive measure ℓ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\ell((a,b]) = b - a$ for all finite $a < b$.*

Theorem 2.1 is typical of statements of measure theory. It defines a complex object (a measure) by its trace on a simple collection of sets (intervals).

The proof of Theorem 2.1 can be cut in several meaningful pieces. First define a length function on intervals. Show that this function can be extended to an additive function on finite union, finite intersection and complements of intervals. Then check that the extension is in fact σ-additive on the closure of intervals under finite set-theoretical operations (which is not a σ-algebra).

Once this additive extension is constructed, use Carathéodory's extension theorem below to prove that the length function can be extended to a σ-additive function on the σ-algebra generated by intervals (the Borel σ-algebra).

Then it remains to check that the extension is unique. This can be done by a generating set argument, for example the ***monotone class Lemma*** Lemma 2.4.

Theorem 2.2 (Carathéodory's extension theorem")**.** *Let $\mathcal{A} \subseteq 2^\Omega$. Assume $\mathcal{A}$ contains $\emptyset, \Omega$, and is closed under finite unions, and complementation. Assume $\rho : \mathcal{A} \to [0,\infty]$ is σ-additive on $\mathcal{A}$.*

Then there exists a measure μ on $\sigma(\mathcal{A})$ such that $\mu(A) = \rho(A)$ for all $A \in \mathcal{A}$.

The Lebesque measure existence theorem guarantees that we can define the uniform probability distribution over a finite interval $[a, b]$. If we denote Lebesgue measure by ℓ, the uniform probability distribution over $[a, b]$ assign probability

$$P(A) = \frac{\ell(A)}{b-a} = \frac{\ell(A)}{\ell([a,b])}$$

to any $A \in \mathcal{B}(\mathbb{R}) \cap [a, b]$.

The uniform distribution over $([0, 1], \mathcal{B}([0, 1]))$ looks like an academic curiosity with no practical utility. This superficial opinion should be dispelled. Using a generator for the uniform distribution, it is possible to build a generator for any probability distribution over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. This can be done using a device called the ***quantile transform***. In this sense, the uniform distribution is the mother of all distribution.

An outcome ω of the uniform distribution is a real number. How does a typical outcome look? A real number $\omega \in [0, 1]$ has binary expansions: $\omega = \sum_{i=1}^{\infty} b_i 2^{-1}$ with $b_i \in \{0, 1\}$. What is the probability there is a unique binary expansion? First, check whether this probability is well-defined. Assuming the binary expansion is unique, ω is said to be ***normal*** if $\lim_n \frac{1}{n} \sum_{i=1}^{n} b_i(\omega) = 1/2$. Is the probability of obtaining a normal number well-defined? If yes, compute it.

Check that $\mathcal{B}(\mathbb{R}) \cap [a, b] = \Big\{ A \cap [a, b] : A \in \mathcal{B}(\mathbb{R}) \Big\}$ is the σ-algebra generated by the trace of the usual topology if $\mathbb{R}$ on $[a, b]$.

The Lebesgue existence theorem can be extended. Indeed, any sensible definition of the length of an interval can serve as a starting point.

Recall that a real function is CADLAG if it is right-continuous everywhere, and has left-limits everywhere.

The next Theorem can be established in a way that parallels the construction of Lebesgue's measure.

Theorem 2.3. *Any non-decreasing CADLAG function F on $\mathbb{R}$ defines a σ-additive measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that satisfies:*

$$\mu((a, b]) = F(b) - F(a)$$

We recover Lebesgue's existence Theorem by taking $F(x) = x$.

If we focus on functions F that satisfy $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, Theorem Theorem 2.3 defines probability distributions through their ***cumulative distribution functions*** (more on this topic in Section 2.7).

Do we really to assume that the function F has left-limits in Theorem Theorem 2.3?

2.5 Measurable functions and random variables

So far, we only talked probability and measure of sets (events). As stochastic modeling is at the root of quantitative analysis, we introduce the notion of measurable function. This allows us handle numerical functions that map outcomes to $\mathbb{R}$ or $\mathbb{R}^d$.

Not every numerical function is measurable. To define what we call a measurable function, we need the notion of *inverse image* or *preimage*.

Definition 2.7 (Preimage). Let f be a function from $\mathcal{X}$ to $\mathcal{Y}$, we denote by f^{-1} the function that maps $2^{\mathcal{Y}}$ to $2^{\mathcal{X}}$ defined by

$$\begin{aligned} f^{-1}: \qquad & 2^{\mathcal{Y}} \to 2^{\mathcal{X}} \\ & B \mapsto f^{-1}(B) = \Big\{ x : x \in \mathcal{X}, f(x) \in B \Big\} . \end{aligned}$$

The set $f^{-1}(B)$ is called the ***preimage*** or ***inverse image*** of B under f.

Note that f^{-1} does not denote the inverse of function f which may not be injective. In this course, f^{-1} is a set function from the powerset of the codomain of f to the powerset of the domain of f. The inverse function if it exists (or the generalized inverse function) is denoted by $f^{\leftarrow}$. The inverse function, when it exists, maps $f(\mathcal{X}) \subseteq \mathcal{Y}$ to $\mathcal{X}$.

Recall the idealized hashing setting from Section 1.1. Let Ω denote the set of functions from $1, \ldots, n$ to $1, \ldots, m$ (assume $n \leq m$). For $\omega \in \Omega$ (ω is function, but it is also a $1, \ldots, m$-valued sequence of length n), let $f(\omega)$ be the number of values in $1, \ldots, m$ that have no occurrence in ω (the number of empty bins in the allocation defined by ω). The function f is a numerical function that maps $\Omega = \{1, \ldots, m\}^n$. For $B \in \mathbb{N}$, $f^{-1}(B)$ is the subset of allocations which have k empty bins, $k \in B$.

The preimage operation works well with set-theoretical operations.

Elementary properties of measurable functions follow from properties of inverse images. Inverse image preserves set-theoretical operations.

Proposition 2.2. *Let $f : E \mapsto F$, then for $A, B, A_1, \ldots, A_n, \ldots \subseteq F$,*

$$\begin{aligned} f^{-1}(A \cup B) &= f^{-1}(A) \cup f^{-1}(B) \\ f^{-1}(A \cap B) &= f^{-1}(A) \cap f^{-1}(B) \\ f^{-1}(\cup_{n \in \mathbb{N}} A_n) &= \cup_{n \in \mathbb{N}} f^{-1}(A_n) \\ f^{-1}(\cap_{n \in \mathbb{N}} A_n) &= \cap_{n \in \mathbb{N}} f^{-1}(A_n) \\ f^{-1}(F \setminus A) &= f^{-1}(F) \setminus f^{-1}(A) \end{aligned}$$

Check Proposition Proposition 2.2 Section 2.7

Taking the preimages of elements of a σ-algebra defines a σ-algebra.

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{G})$ be two measurable spaces. Let f map Ω to Ω', prove that

$$\mathcal{H} = \Big\{ f^{-1}(B) : B \in \mathcal{G} \Big\}$$

is a σ-algebra of subsets of $f^{-1}(\Omega')$.

Definition 2.8 (Measurable functions)**.** Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{G})$ be two measurable spaces. A function $f : \Omega \to \Omega'$ is said to be $\mathcal{F}/\mathcal{G}$-measurable iff for $B \in \mathcal{G}$, $f^{-1}(B) \in \mathcal{F}$.

Under which condition on $\mathcal{H}$ is f $\mathcal{F}/\mathcal{G}$-measurable?

Recall the idealized hashing scenario from Section 1.1.

Check that if Ω is a topological space and $\mathcal{F}$ the associated Borelian σ-algebra, then any continuous function from Ω to $\mathbb{R}$ is measurable.

If $\Omega = \mathbb{R}^d$ is the Borel σ-algebra the smallest σ-algebra that makes all continuous functions measurable?

Proposition 2.3. *The pointwise limit of measurable functions is a measurable function: if $(f_n)_n$ is a sequence of measurable functions from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{G})$, and $f_n \to f$ pointwise, then f is a measurable function.*

Prove Proposition Proposition 2.3

Proposition 2.4. *The sum of measurable functions is a measurable function: if f, g are measurable functions from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then $af + bg$ is a measurable function for all $a, b \in \mathbb{R}$.*

Prove Proposition Proposition 2.4

Proposition 2.5. *The composition of measurable functions is a measurable function: if f is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{G})$, and g is a measurable function from $(\mathcal{X}, \mathcal{G})$ to $(\mathcal{Y}, \mathcal{H})$, then $g \circ f$ ($g \circ f(\omega) = g(f(\omega))$ for all ω) is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{H})$.*

Prove Proposition Proposition 2.5

2.6 The Monotone class theorem

The monotone class theorem or lemma is a powerful example of the generating class arguments that can be used to prove that two probability measures or maybe two σ-finite measures are equal.

Definition 2.9 (π-class)**.** A collection $\mathcal{G}$ of subsets of Ω is said to be a π-class if:

- $\Omega \in \mathcal{G}$
- it is stable/closed by finite intersection

$$A, B \in \mathcal{G} \Rightarrow A \cap B \in \mathcal{G} .$$

A σ-algebra is a π class, but the converse is false.

Definition 2.10 (Monotone class)**.** A collection $\mathcal{M}$ of subsets of Ω is said to be a monotone class or a λ-system if it satisfies the following properties:

- $\Omega \in \mathcal{M}$
- If $A, B \in \mathcal{M}$, and $A \subseteq B$ then $B \setminus A \in \mathcal{M}$
- If $A_n \in \mathcal{M}$ and $A_n \subseteq A_{n+1}$ for every $n \in \mathbb{N}$ then $\lim_n A_n = \cup_{n \in \mathbb{N}} A_n \in \mathcal{M}$.

A σ-algebra is a λ-system.

The intersection of a collection of λ-systems is a λ-system. Hence, it makes sense to talk about the smallest λ-system containing a collection of sets.

The next easy proposition makes λ-system very useful when we want to check that two probability distributions are equal.

Proposition 2.6. *The class of sets over which two probability distributions coincide is a λ-system.*

Proof. Let $(\Omega, \mathcal{F})$ be a measurable space. Let P, Q be two probability distributions over $(\Omega, \mathcal{F})$. Let $\mathcal{C} \subseteq \mathcal{F}$ be defined by

$$\mathcal{C} = \Big\{ A : A \in \mathcal{F}, P(A) = Q(A) \Big\} .$$

By the very definition of measures we have $P(\Omega) = Q(\Omega)$, hence $\Omega \in \mathcal{C}$.

If $A \subseteq B$ both belong to $\mathcal{C}$, again by the very definition of measures,

$$P(B \setminus A) = P(B) - P(A) = Q(B) - Q(A) = Q(B \setminus A)\,,$$

hence, $B \setminus A \subseteq \mathcal{C}$.

Let $A_1 \subseteq A_2 \subseteq A_n \subseteq \ldots$ be a non-decreasing sequence of elements of $\mathcal{C}$, again by the very definition of measures,

$$P(\cup_n A_n) = \lim_n \uparrow P(A_n) = \lim_n \uparrow Q(A_n) = Q(\cup_n A_n)\,.$$

Hence $\mathcal{C}$ is closed by monotone limits. □

Exercise 2.7. What happens if we consider the collections of measurable sets over which two measures are equal? What happens if we assume that the two measures are finite?

Definition 2.11 (σ-finite measures)**.** A measure μ on $(\Omega, \mathcal{F})$ is σ-finite iff there exists $(A_n)_n$ with $\Omega \subseteq \cup_n A_n$ and $\mu(A_n) < \infty$ for each n.

Finite measures (this encompasses probability measures) are σ-finite. Lebesgue measure is σ-finite. The counting measure on $\mathbb{R}$ is not σ-finite.

What happens if we only assume that the two measures are σ-finite?

Theorem 2.4 (Monotone class lemma)**.** *If $\mathcal{A}$ is a π-systen in Ω and $\mathcal{M}$ a λ-system in Ω such that $\mathcal{A} \subseteq \mathcal{M}$, then the σ-algebra generated by $\mathcal{A}$, $\sigma(A)$, is the smallest λ-system larger than $\mathcal{A}$:*

$$\sigma(\mathcal{A}) \subseteq \mathcal{M}\,.$$

Proof. Let $\mathcal{M}$ denote the intersection of all monotone classes that contain tyhe π-system $\mathcal{A}$. As a σ-algebra is a monotone class (a λ-system), we have $\mathcal{M} \subseteq \sigma(\mathcal{A})$, the only point that has to be checked is $\sigma(\mathcal{A}) \subseteq \mathcal{M}$. It is enough to check that $\mathcal{M}$ is indeed a σ-algebra.

In order to check that $\mathcal{M}$ is a σ-algebra, it is enough to check that it is closed under finite union or equivalently under finite intersection.

For each $A \in \mathcal{A}$, let $\mathcal{M}_A$ be defined by

$$\mathcal{M}_A = \Big\{B : B \in \mathcal{M}, A \cap B \in \mathcal{M}\Big\}\,.$$

Remember that $\mathcal{A}$ is a π-system, and $\mathcal{A} \subseteq \mathcal{M}$, we have $\mathcal{A} \subseteq \mathcal{M}_A$. To show that $\mathcal{M} = \mathcal{M}_A$, it suffices to show that $\mathcal{M}_A$ is a monotone class.

If $(B_n)_n$ is an increasing sequence of elements of $\mathcal{M}_A$, then

$$(\cup_n B_n) \cap A = \cup_n \Big(\underbrace{B_n \cap A}_{\in \mathcal{M}}\Big)\,,$$

the right-hand-side belongs to $\mathcal{M}$ since $\mathcal{M}$ is monotone. Hence $\mathcal{M}_A$ is closed by monotone increasing limit.

To check closure by complementation, let $B \subseteq C$ with $B, C \in \mathcal{M}_A$. As

$$A \cap (C \setminus B) = \Big(\underbrace{A \cap C}_{\in \mathcal{M}}\Big) \setminus \Big(\underbrace{A \cap B}_{\in \mathcal{M}}\Big))$$

the closure of $\mathcal{M}$ under complementation entails $A \cap (C \setminus B) \in \mathcal{M}$ and $C \setminus B \in \mathcal{M}_A$.

Now, let $\mathcal{M}^\circ$ be defined as

$$\mathcal{M}^\circ = \Big\{A : A \in \mathcal{M}, \forall B \in \mathcal{M}, A \cap B \in \mathcal{M}\Big\}\,.$$

We just established that $\mathcal{A} \subseteq \mathcal{M}^\circ$. Using the same line of reasoning allows us to check that $\mathcal{M}^\circ$ is also a monotone class. This shows that $\mathcal{M}^\circ = \mathcal{M}$.

We are done. □

Combining Proposition 2.6 and the Monotone Class Lemma (Theorem 2.4) leads to the next useful corollary.

If two probabilities P, Q on $(\Omega, \mathcal{F})$ coincide on a π-system $\mathcal{A}$ that generates $\mathcal{F}$:

$$\mathcal{A} \subseteq \{A : A \in \mathcal{F} \text{ and } P(A) = Q(A)\} \qquad \text{and} \qquad \mathcal{F} \subseteq \sigma(\mathcal{A})$$

then P, Q coincide on $\mathcal{F}$.

2.7 Probability distributions on the real line

A probability distribution is a complex object: it maps a large collection of sets (a σ-algebra) to $[0, 1]$. Fortunately, it is possible to characterize a probability distribution by simpler object. If we focus on probability distributions over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, they can be characterized by real functions on $\mathbb{R}$.

Definition 2.12 (Distribution function). Given a probability distribution P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the distribution function F of P maps $\mathbb{R}$ to $[0, 1]$, it is defined by

$$x \mapsto F(x) = P(-\infty, x].$$

A probability distribution defines a unique distribution function. What is perhaps surprising is that a distribution function defines a unique probability distribution function.

Proposition 2.7. *Let F be a function from $\mathbb{R}$ to $[0, 1]$.*

The function F is the distribution function of a probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, iff the following five properties are satisfied:

1. *F is non-decreasing,*
2. *F is right-continuous*
3. *$\lim_{y \nearrow x} F(y)$ exists at every $x \in \mathbb{R}$ (F has left-limits everywhere)*
4. *$\lim_{x \to -\infty} F(x) = 0$*
5. *$\lim_{x \to \infty} F(x) = 1$.*

This is a rephrasing of Theorem 2.3.

Figure Figure 2.1 shows the cumulative distribution function of Poisson distributions for different values of the parameter (see Sections Section 1.4 and Section 4.2 for more on Poisson distributions). For parameter μ, $F_\mu(x) = \sum_{k \leq x} \mathrm{e}^{-\mu} \frac{\mu^k}{k!}$.

2.8 General random variables

A real random variable is neither a variable, nor random. A real random variable is a measurable function from some measurable space to the real line endowed with the Borel σ-algebra. There is nothing random in a random variable.

Definition 2.13 (Real valued random variable). Given a measurable space $(\Omega, \mathbb{F})$, a mapping X from Ω to $\mathbb{R}$ is a real valued random variable such that for every $B \in \mathcal{B}(\mathbb{R})$ the inverse image of B:

$$X^{-1}(B) = \{\omega : \omega \in \Omega, X(\omega) \in B\}$$

belongs to $\mathcal{F}$

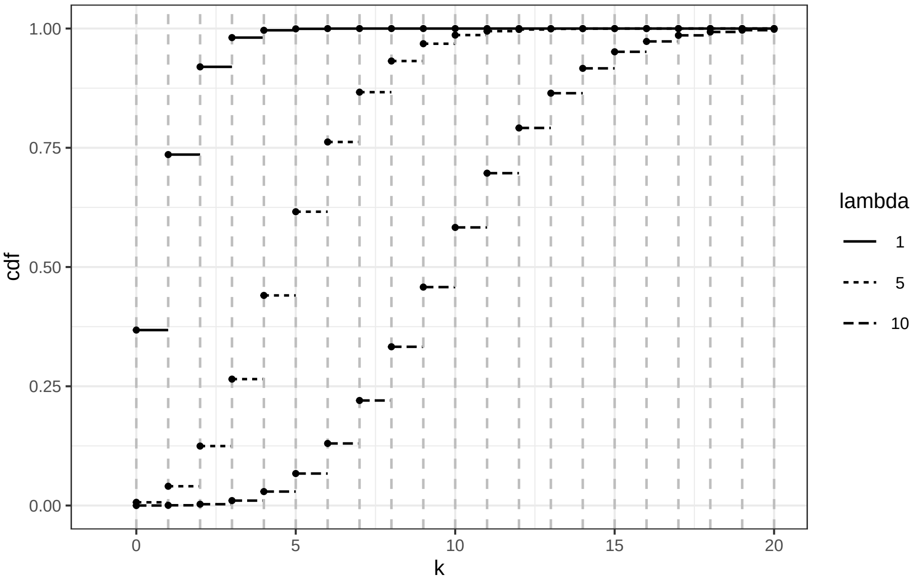

Figure 2.1: Cumulative distribution functions for Poisson distributions with different parameters. Observe that, apparently, $\mu \leq \nu \Rightarrow F_\mu \geq F_\nu$. How would you establish this domination property?

Once a measurable space is endowed with a probability distribution, is it possible to define the (probability) distribution of a random variable.

Definition 2.14. Given $(\Omega, \mathcal{F}, P)$ and a real valued random variable X, the law or probability distribution of X, denoted by $P \circ X^{-1}$, is the probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$(P \circ X^{-1})(B) = P(X^{-1}(B)) \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

Random variables may be vector-valued, function-valued, *etc*. General random variables are defined as measurable functions between measurable spaces.

Definition 2.15 (Random variable)**.** Given two measurable spaces $(\Omega, \mathcal{F})$, and $(\Omega', \mathcal{G})$ a mapping X from Ω to Ω' is a $\mathcal{F}/\mathcal{G}$-random variable if for every $B \in \mathcal{G}$ the inverse image of B:

$$X^{-1}(B) = \{\omega : \omega \in \Omega, X(\omega) \in B\}$$

belongs to $\mathcal{F}$.

2.9 Bibliographic remarks

There are many beautiful books on Probability Theory. They are targetted at different audiences. Some may be more suited to the students of the dual curriculum Mathématiques-Informatique. I found the following ones particularly useful.

Youssef (2019) is a clear and concise collection of class notes designed for a Master I-level Probability course that is substantially more ambitious than this minimal course.

Dudley (2002) delivers a self-contained course on Analysis and Probability. The book can serve both as an introduction and a reference book. Beyond cautious and transparent proofs, it contains historical notes that help understand the connections between landmark results.

Pollard (2002) introduces measure and integration theory to an audience that has been exposed to discrete probability theory and that is familiar with probabilistic reasoning.

Chapter 3

A modicum of integration

3.1 Roadmap

We start by reviewing basic definitions and results from integration theory. We follow the measure-theoretic approach. First, we define ***simple functions,*** a subclass of piecewise measurable functions in Section 3.2). Defining the integral of a simple function with respect to a measure in Section 3.3) is straightforward. Some more work allows us to derive useful properties: linearity, monotonicity, to name a few. In Section 3.3), we define the integral of a non-negative measurable function as a supremum of integrals of simple functions. This definition is theoretically sound and it lends itself to computations. Section 3.4) states three convergence theorems culminating with the ***dominated convergence theorem***.

In Section 3.6), we relate the notion of ***expectation*** of a random variable and the notion of integral. The ***Transfer Theorem*** (Theorem 3.4) is a key instrument in the characterization of image distributions.

3.2 Simple functions

The integral of a $\{0,1\}$-valued measurable function f with respect to a measure μ is defined

$$\int_\Omega f \mathrm{d}\mu = \mu\Big(f^{-1}(\{1\})\Big)\,,$$

alternatively

$$\int_\Omega \mathbb{I}_A \mathrm{d}\mu = \mu(A) \qquad \text{for any measurable set } A\,.$$

The next step consists in defining the integral of finite linear combinations of $\{0,1\}$-valued measurable function f.

Definition 3.1 (Simple function). Let $(\Omega, \mathcal{F})$ be a measurable space. The function $f:\Omega\to\mathbb{R}$ is said to be ***simple*** iff

- f takes finitely many values: $\Big|\{f(x) : x\in\Omega\}\Big| < \infty$
- For each $y\in f(\Omega)\subset\mathbb{R}$, $f^{-1}(\{y\})\in\mathcal{F}$.

A simple function defines a partition of Ω into finitely many measurable classes. The simple function is constant on each class.

If f is a simple function, then the σ-algebra $f^{-1}(\mathcal{B}(\mathbb{R}))$ is finite.

Simple functions are finite linear combinations of set characteristic (indicator) functions.

- For each $A \in \mathcal{F}$, $\mathbb{I}_A$ is simple
- For any finite collection $A_1, \ldots, A_n$ of measurable subsets of Ω, any sequence $c_1, \ldots, c_n$ of real numbers, $\sum_{i \leq n} c_i \mathbb{I}_{A_i}$ is a simple function
- For any measurable function $f : \Omega \to \mathbb{R}$, and $n \in \mathbb{N}$, the function g_n defined by

$$g_n(\omega) = n \wedge (-n \vee \lfloor f(\omega) \rfloor)$$

is simple.

The definition of the integral of a simple function with respect to a measure is straightforward: it is a finite sum.

Definition 3.2 (Integral of a simple function)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measured space. Let $f : \Omega \to \mathbb{R}$ be a non-negative simple function which is defined by a finite partition of Ω into measurable sets $A_1, A_2, \ldots, A_n$ and numbers $f_1, \ldots, f_n$:

$$f(\omega) = \sum_{i \leq n} f_i \mathbb{I}_{A_i}(\omega) .$$

The integral of f with respect to μ is defined by

$$\int_\Omega f \mathrm{d}\mu = \sum_{i \leq n} f_i \mu(A_i) .$$

Note that if measure μ is not finite, the integral of a simple non-negative function may be infinite.

If $\mu(A_i) = \infty$ and $f_i = 0$, we agree on $f_i \mu(A_i) = 0$.

If we turn to signed simple functions, it is enough to notice than if f is simple, so are $(f)_+$ and $(f)_-$ and to define $\int_\Omega f \mathrm{d}\mu$ as

$$\int_\Omega (f)_+ \mathrm{d}\mu - \int_\Omega (f)_- \mathrm{d}\mu$$

provided at leat one of the two summands is finite.

Although they are simple, simple functions have interesting approximation capabilities. Any non-negative measurable function can be approximated from below by non-negative simple functions.

Proposition 3.1 (Approximation of measurable functions)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space. Any non-negative measurable function $f : \Omega \to \mathbb{R}$ is the monotone pointwise limit of simple functions: there exists a sequence of simple function $f_1, \ldots, f_n, \ldots$ such that for each $\omega \in \Omega$, the following holds:*

$$f_1(\omega) \leq f_2(\omega) \leq \ldots \leq f_n(\omega) \leq \ldots \leq f(\omega)$$

and

$$\lim_n f_n(\omega) = f(\omega) .$$

Proof. Define f_n as

$$f_n(\omega) = n \wedge \left(2^{-n} \lfloor 2^n f(\omega) \rfloor \right) .$$

As

$$\lfloor 2^n f(\omega) \rfloor \leq 2^n f(\omega)$$

we have $f_n(\omega) \leq f(\omega)$ for all ω.

The range of function f_n is $i \times 2^{-n}$ for $i = 0, \ldots, n \times 2^n$. For each $i \in 0, \ldots, (n-1) \times 2^n$

$$f_n^{-1}\Big(\{i \times 2^{-n}\}\Big) = f^{-1}\Big(\Big[\frac{i}{2^n}, \frac{i+1}{2^n}\Big)\Big)$$

which is in $\mathcal{F}$ because f is measurable and $\Big[\frac{i}{2^n}, \frac{i+1}{2^n}\Big) \in \mathcal{B}(\mathbb{R})$.

Likewise $f_n^{-1}\Big(\{n\}\Big) = f^{-1}([n, \infty))$ belongs to $\mathcal{F}$.

To check that $f_n \leq f_{n+1}$, we consider two cases.

1. $f_{n+1}(\omega) \geq n$. This entails $f(\omega) \geq n$ and thus $f_n(\omega) = n < f_{n+1}(\omega)$
2. $f_{n+1}(\omega) = k + i2^{-n-1}$ for $k < n$ and $i < 2^{n+1}$. This entails $f_n(\omega) = k + \lfloor i/2 \rfloor 2^{-n} \leq f_{n+1}(\omega)$.

Finally if $f(\omega) \leq n$, $0 \leq f(\omega) - f_n(\omega) \leq 2^{-n}$. This implies that $\lim_n f_n(\omega) = f(\omega)$ for all ω. □

Figure Figure 3.1 illustrates the approximation capabilities of simple functions.

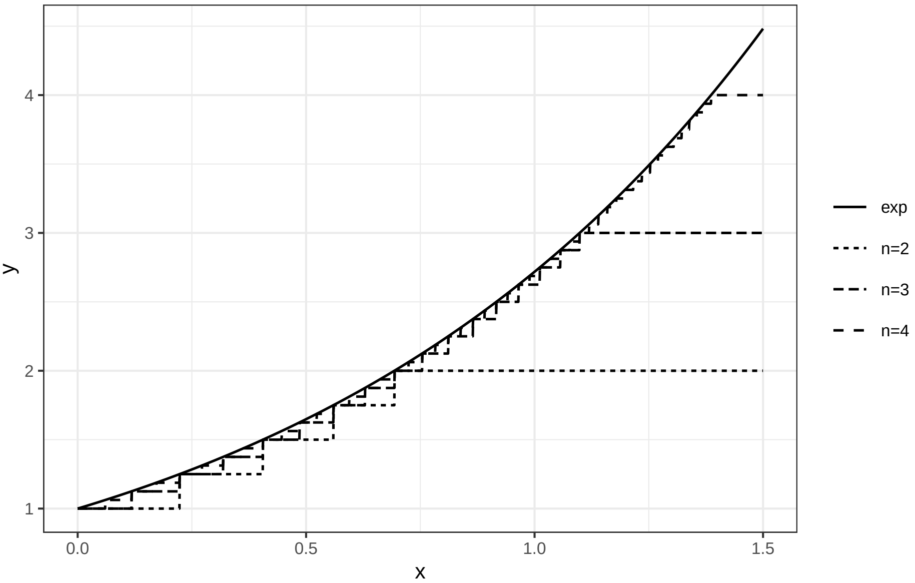

Figure 3.1: Approximation of the exponential function by simple functions $n \wedge \Big(2^{-n}\lfloor 2^n \exp(\omega)\rfloor\Big)$ for $n = 2, 3, 4$.

If f, g are two non-negative simple functions on $(\Omega, \mathcal{F})$, then for all $a, b \in [0, \infty)$, $af + bg$ and fg are non-negative simple functions.

Check the proposition.

Proposition 3.2 (Monotonicity of integration of simple functions). *If f, g are two non-negative simple functions and μ a non-negative measure on $(\Omega, \mathcal{F})$ such that*

$$\mu\Big\{\omega : f(\omega) > g(\omega)\Big\} = 0 \,.$$

(f is less of equal than g μ-almost everywhere), then

$$\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu \,.$$

Check Proposition 3.2

Proposition 3.3 (Linearity of integration of simple functions). ***If*** f, g ***are two non-negative simple functions and*** μ ***a non-negative measure on*** $(\Omega, \mathcal{F})$***, then for all*** $a, b \in [0, \infty)$***,***

$$\int af + bg \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu + b \int g \, \mathrm{d}\mu \, .$$

Check Proposition Proposition 3.3

3.3 Integration

Let $\mathcal{S}_+$ denote the set of non-negative simple functions on $(\Omega, \mathcal{F})$.

Definition 3.3 (Integration with respect to a measure). Let f be a non-negative measurable function on $(\Omega, \mathcal{F}, \mu)$, then for any $A \in \mathcal{F}$, the integral of f over A with respect to measure μ is defined by:

$$\int_A f \, \mathrm{d}\mu = \sup_{s \in \mathcal{S}_+ : s \leq f} \int_A s \, \mathrm{d}\mu$$

If the supremum is finite, the function is said to be ***integrable*** with respect to μ, or to be μ-integrable.

Proposition 3.4 (Monotonicity of integration). ***If*** f, g ***are two non-negative measurable functions and*** μ ***a non-negative measure on*** $(\Omega, \mathcal{F})$ ***such that***

$$\mu\Big\{\omega : f(\omega) > g(\omega)\Big\} = 0 \, .$$

(f ***is less of equal than*** g μ***-almost everywhere), then***

$$\int f \, \mathrm{d}\mu \leq \int g \, \mathrm{d}\mu \, .$$

Prove Proposition Proposition 3.4.

Proposition 3.5 (Linearity of integration). ***If*** f, g ***are two non-negative measurable functions and*** μ ***a non-negative measure on*** $(\Omega, \mathcal{F})$***, then for all*** $a, b \in [0, \infty)$***,***

$$\int af + bg \, \mathrm{d}\mu = a \int f \, \mathrm{d}\mu + b \int g \, \mathrm{d}\mu \, .$$

Prove Proposition Proposition 3.5.

The integral of a signed measurable functions is defined by a decomposition argument. Let f be a measurable function and $f = (f)_+ - (f)_-$, then

$$\int_\Omega f \mathrm{d}\mu = \int_\Omega (f)_+ \mathrm{d}\mu - \int_\Omega (f)_- \mathrm{d}\mu$$

provided at least one of $\int_\Omega (f)_+ \mathrm{d}\mu$ and $\int_\Omega (f)_- \mathrm{d}\mu$ is finite.

3.4 Limit theorems

In this section, measurable functions are meant to be real-valued, and $\mathbb{R}$ is endowed with the Borel σ-algebra ($\mathcal{B}(\mathbb{R})$).

Theorems Theorem 3.1, Theorem 3.2, Theorem 3.3 below are the three pillars of integral calculus.

Theorem 3.1 (Monotone convergence theorem). *Let $(\Omega, \mathcal{F}, \mu)$ be a measured space. Let $(f_n)_n$ be a non-decreasing sequence of non-negative measurable functions converging towards f. Then*

$$\int \lim_n \uparrow f_n \, \mathrm{d}\mu = \lim_n \uparrow \int f_n \, \mathrm{d}\mu.$$

The proof of the monotone convergence theorem boils down to the definition of positive measure and property $\mu(\lim_n \uparrow A_n) = \lim_n \uparrow \mu(A_n)$.

Proof. Let function f be defined by $f(\omega) = \lim_n \uparrow f_n(\omega)$ for all $\omega \in \Omega$. Note that if $f(\omega) = 0$, then $f_n(\omega) = 0$ for all $n \in \mathbb{N}$.

The function f is positive measurable. In order to prove the monotone convergence theorem it is enough to check that for every non-negative simple function g such that $g \leq f$ everywhere, for any $a \in [0, 1)$, the following holds:

$$a \int g \, \mathrm{d}\mu \leq \lim_n \uparrow \int f_n \, \mathrm{d}\mu \,. \tag{3.1}$$

For each $n \in \mathbb{N}$, define

$$E_n = \Big\{\omega : f_n(\omega) \geq ag(\omega)\Big\}.$$

Note that as $(f_n)_n$ is non-decreasing, the sequence (E_n) is non-decreasing. Moreover, if $f(\omega) > 0$ as $\lim_n \uparrow f_n(\omega) = f(\omega) > af(\omega) \geq ag(\omega)$. Hence for all $\omega \in \Omega$, $\mathbb{I}_{E_n}(\omega) = 1$ for all sufficiently large n (beware there is no uniformity guarantee). We have

$$\lim_n \uparrow E_n = \Omega \,.$$

Combining the different remarks, we have for all n, $\mathbb{I}_{E_n} ag \leq f_n$ everywhere. Monotonicity of integration entails, for all n

$$\int \mathbb{I}_{E_n} ag \, \mathrm{d}\mu \leq \int f_n \, \mathrm{d}\mu \qquad \forall n \,.$$

Now, for each n, $\mathbb{I}_{E_n} ag$ is a non-negative simple function, and the sequence $(\mathbb{I}_{E_n} ag)_n$ is a non-decreasing sequence of non-negative simple functions converging towards simple function ag.

Let $g = \sum_{i \leq k} c_i \mathbb{I}_{A_i}$ where $(A_i)_{i \leq k}$ is a finite partition of Ω into measurable subsets.

$$\mathbb{I}_{E_n} g = \sum_{i \leq k} c_i \mathbb{I}_{A_i \cap E_n} \,.$$

Hence

$$\begin{aligned} \textstyle\int \mathbb{I}_{E_n} ag \, \mathrm{d}\mu & = \textstyle\sum_{i \leq k} c_i \int \mathbb{I}_{A_i \cap E_n} \, \mathrm{d}\mu \\ & = \textstyle\sum_{i \leq k} c_i \mu(A_i \cap E_n) \,. \end{aligned}$$

For each $i \leq k$, we have $\lim_n \uparrow c_i \mu(A_i \cap E_n) = c_i \mu(A_i)$. We have:

$$\int \lim_n \uparrow \mathbb{I}_{E_n} ag \, \mathrm{d}\mu = \lim_n \uparrow \int \mathbb{I}_{E_n} ag \, \mathrm{d}\mu \, .$$

This proves that Equation 3.1 holds for all $a \in [0,1)$ and $g \in \mathcal{S}_+$ with $g \leq f$:

$$\forall g \in \mathcal{S}_+ \text{ with } \forall a \in [0,1),$$

□

The non-negativity assumptiom on f_n is not necessary. It is enough to assume $\int f_1 \mathrm{d}\mu > -\infty$. Prove this.

Let $(f_n)_n$ be a monotone decreasing sequence of non-negative measurable functions. Let $f = \lim_n \downarrow f_n$ (check the existence of f).

Is it true that $\int \lim_n \downarrow f_n \mathrm{d}\mu = \lim_n \downarrow \int f_n \mathrm{d}\mu$?.

Answer the same question assuming $\int f_1 \mathrm{d}\mu < \infty$.

Answer the same question if μ is assumed to be a probability measure.

Theorem 3.2 (Fatou's Lemma). *Let $(\Omega, \mathcal{F}, \mu)$ be a measured space. Let $(f_n)_n$ be a sequence of non-negative measurable functions. Then*

$$\int \liminf_n f_n \mathrm{d}\mu \leq \liminf_n \int f_n \mathrm{d}\mu.$$

Proof. Define $h_n(\omega) = \inf_{m \geq n} f_n(\omega)$. Each h_n is also non-negative and measurable. By monotonicity,

$$\int h_n \mathrm{d}\mu \leq \inf_{m \geq n} \int f_m \mathrm{d}\mu \, .$$

The sequence h_n is non-decreasing. And $\lim \uparrow h_n(\omega) = \liminf f_n(\omega)$ for all $\omega \in \Omega$. For each n, by the monotone convergence theorem (Theorem 3.1):

$$\int \lim_n \uparrow h_n \mathrm{d}\mu = \lim_n \uparrow \int h_n \mathrm{d}\mu$$

so that

$$\int \liminf_n f_n \mathrm{d}\mu = \lim_n \uparrow \int h_n \mathrm{d}\mu$$

and

$$\int \liminf_n f_n \mathrm{d}\mu \leq \lim_n \inf_{m \geq n} \int f_m \mathrm{d}\mu = \liminf_n \int f_n \mathrm{d}\mu$$

□

Theorem 3.3 (Dominated convergence theorem). *Let $(\Omega, \mathcal{F}, \mu)$ be a measured space. Let $(f_n)_n$ be a sequence of measurable functions that converges pointwise towards function f. Assume that there exists a integrable function g that dominates $(f_n)_n$: for all n, all $\omega \in \Omega$, $|f_n(\omega)| \leq g(\omega)$. Then f is integrable and*

$$\int f \mathrm{d}\mu = \int \lim_n f_n \mathrm{d}\mu = \lim_n \int f_n \mathrm{d}\mu.$$

Proof. Let us first check that f is integrable.

Observe that $\lim_n |f_n| = |f|$ and thus $\liminf |f_n| = |f|$.

By Theorem 3.2,

$$\int |f| \mathrm{d}\mu = \int \liminf_n |f_n| \mathrm{d}\mu \leq \liminf_n \int |f_n| \mathrm{d}\mu = \int |g| \mathrm{d}\mu < \infty \, .$$

Now define $h_n = \inf_{m\geq n} f_m$ and $j_n = \sup_{m\geq n} f_m$. We have $\lim_n \uparrow h_n = f$ and $\lim_n \downarrow j_n = f$.

Note also that

$$\int h_n \mathrm{d}\mu \leq \int f \mathrm{d}\mu \leq \int j_n \mathrm{d}\mu \, .$$

By monotone convergence $\int h_n \mathrm{d}\mu \uparrow \int f \mathrm{d}\mu$ and $\int j_n \mathrm{d}\mu \downarrow \int f \mathrm{d}\mu$. This entails $\lim \int f_n \mathrm{d}\mu$. □

Let $g : \Omega \times \mathbb{R} \to \mathbb{R}$ be a function of two variables such that for each $t \in \mathbb{R}$, $g(\cdot, t)$ is measurable. Assume that for each $t \in \mathbb{R}$, $g(\cdot, t)$ is μ-integrable and that for each $\omega \in \Omega$, $g(\omega, \cdot)$ is differentiable. Define $G(t) = \int_\Omega g(\omega, t)\mathrm{d}\mu(\omega)$.

Is it always true that G is differentiable at every t?

Provide sufficient conditions for G to be differentiable and

$$G'(t) = \int \frac{\partial g}{\partial s}(\omega, s)_{|s=t} \mathrm{d}\mu(\omega) \, .$$

3.5 Probability distributions defined by a density

Let $(\Omega, \mathcal{F})$ be a measurable space and μ be a σ-finite measure over $(\Omega, \mathcal{F})$. Let f be a non-negative measurable real function over $(\Omega, \mathcal{F})$.

Let $\nu : \mathcal{F} \to [0, \infty)$ be defined by

$$\nu(A) = \int \mathbb{I}_A f \, \mathrm{d}\mu = \int_A f \, \mathrm{d}\mu \, .$$

ν is a measure over $(\Omega, \mathcal{F})$. The function f is said to be a density of ν with respect to μ.

Proof. The fact that $\nu(\emptyset) = 0$ is immediate.

The fact that ν is σ-additive follows from the monotone convergence theorem (Theorem 3.1).

If $A_1, \ldots, A_n, \ldots$ is a collection or pairwise disjoint measurable sets,

$$\begin{aligned}
\nu(\cup_n A_n) &= \int \mathbb{I}_{\cup_n A} f \, \mathrm{d}\mu \\
&= \int \left(\lim_n \sum_{k\leq n} \mathbb{I}_{A_k} \right) f \, \mathrm{d}\mu \\
&= \int \left(\lim_n \sum_{k\leq n} \mathbb{I}_{A_k} f \right) \mathrm{d}\mu \\
&= \lim_n \sum_{k\leq n} \int \mathbb{I}_{A_k} f \, \mathrm{d}\mu \\
&= \lim_n \sum_{k\leq n} \int \mathbb{I}_{A_k} f \, \mathrm{d}\mu \\
&= \lim_n \sum_{k\leq n} \nu(A_k) \\
&= \sum_{k=1}^{\infty} \nu(A_k) \, .
\end{aligned}$$

The fourth equality is justified by the monotone convergence theorem, others equalities follow from the fact that we are handling non-negative series. □

Let $(A_n)_n$ be such that $A_n \in \mathcal{F}, \mu(A_n) < \infty$ for each n and $\cup_n A_n = \Omega$. For each n, we have $\nu(A_n) = \int_{A_n} f \, \mathrm{d}\mu \leq \int_\Omega f \, \mathrm{d}\mu < \infty$. This proves that if μ is σ-finite, so is ν.

Check that if $\mu(A) = 0$, then $\nu(A) = 0$ for every $A \in \mathcal{F}$.

TODO.

3.6 Expectation

The expectation of a real random variable is a (Lebesgue) integral with respect to a probability measure. We have to get familiar with probabilistic notation.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The random variable X defined on $(\Omega, \mathcal{F})$ is P-integrable if the measurable function $|X| : \omega \mapsto |X(\omega)|$ is P-integrable: we agree on

$$\mathbb{E}X = \mathbb{E}_P X = \int_{\mathcal{X}} X(\omega)\mathrm{d}P(\omega) = \int X\mathrm{d}P.$$

Check the consistency of this definition with the definition used in the discrete setting.

The next statement called the ***transfer formula*** can be used to compute the density of an image distribution or to simplify the computation of an expectation.

Theorem 3.4 (Transfer formula). ***Let*** $(\mathcal{X}, \mathcal{F}, P)$ ***be a probability space,*** $(\mathcal{Y}, \mathcal{G})$ ***a measurable space,*** f ***a measurable function from*** $(\mathcal{X}, \mathcal{F})$ ***to*** $(\mathcal{Y}, \mathcal{G})$***. Let*** Q ***denote the probability distribution that is the image of*** P ***by*** f***:*** $Q = P \circ f^{-1}$***.***

Then for all measurable functions h ***from*** $(\mathcal{Y}, \mathcal{G})$ ***to*** $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

$$\mathbb{E}[h(Y)] = \int_{\mathcal{Y}} h(y)\mathrm{d}Q(y) = \int_{\mathcal{X}} h \circ f(x)\mathrm{d}P(x) = \mathbb{E}h \circ f(X)$$

if either integral is defined.

Proof. Assume first that $h = \mathbb{I}_B$ where $C \in \mathcal{G}$. Then

$$\begin{aligned} \mathbb{E}h(Y) &= \textstyle\int_{\mathcal{Y}} \mathbb{I}_B(y)\,\mathrm{d}Q(y) \\ &= Q(B) \\ &= P \circ f^{-1}(B) \\ &= P\Big\{x : f(x) \in B\Big\} \\ &= P\Big\{x : h \circ f(x) = 1\Big\} \\ &= \textstyle\int_{\mathcal{X}} h \circ f(x)\mathrm{d}P(x) \\ &= \mathbb{E}h \circ f(X)\,. \end{aligned}$$

Then, by linearity, the transfer formula holds for all simple functions from $\mathcal{Y}$ to $\mathbb{R}$. By the definition of the Lebesgue integral, the transfer formula holds for non-negative measurable functions. The usual decomposition argument completes the proof. □

It is clear that the expectation of a random variable only depends on the probability distribution of the random variable.

3.7 Jensen's inequality

The tools from integration theory we have reviewed so far serve to compute or approximate integrals and expectations. The next theorem circumvents computations and allows us to compare expectations.

Jensen's inequality is a workhorse of Information Theory, Statistics and large parts of Probability Theory. It embodies the interaction between convexity and expectation.

We first introduce a modicum of convexity theory and notation.

Definition 3.4 (Lower semi-continuity). A function f from some metric space $\mathcal{X}$ to $\mathbb{R}$ is lower semi-continuous at $x \in \mathcal{X}$, if

$$\liminf_{x_n \to x} f(x_n) \geq f(x) \, .$$

A continuous function is lower semi-continuous. But the converse is not true. If $A \subseteq \mathcal{X}$ is an open set, then $\mathbb{I}_A$ is lower semi-continuous but, unless it is constant, it is not continuous at the boundary of A.

Definition 3.5 (Convex subset). Let $\mathcal{X}$ be a vector space, a subset $C \subseteq \mathcal{X}$ is said to be convex if for all $x, y \in C$, all $\lambda \in [0, 1]$:

$$\lambda x + (1 - \lambda) y \in C \, .$$

Let C be a convex subset of some (topological real) vector space, let $\overline{C}$ be the closure of C. Prove that $\overline{C}$ and $\overline{C} \setminus C$ are convex.

A convex set may be neither closed nor open. Provide examples.

In the next definition, we consider functions from some vector space to $\mathbb{R} \cup \{+\infty\}$.

Definition 3.6 (Convex functions). Let $\mathcal{X}$ be a (topological) vector space. Let $C \subseteq \mathcal{X}$ be a convex subset. A function f from $\mathcal{C}$ to $\mathbb{R} \cup \{\infty\}$ is convex if for $x, y \in C$, all $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y) \, .$$

The *domain* of f $\operatorname{Dom}(f)$ is the subset of C where f is finite.

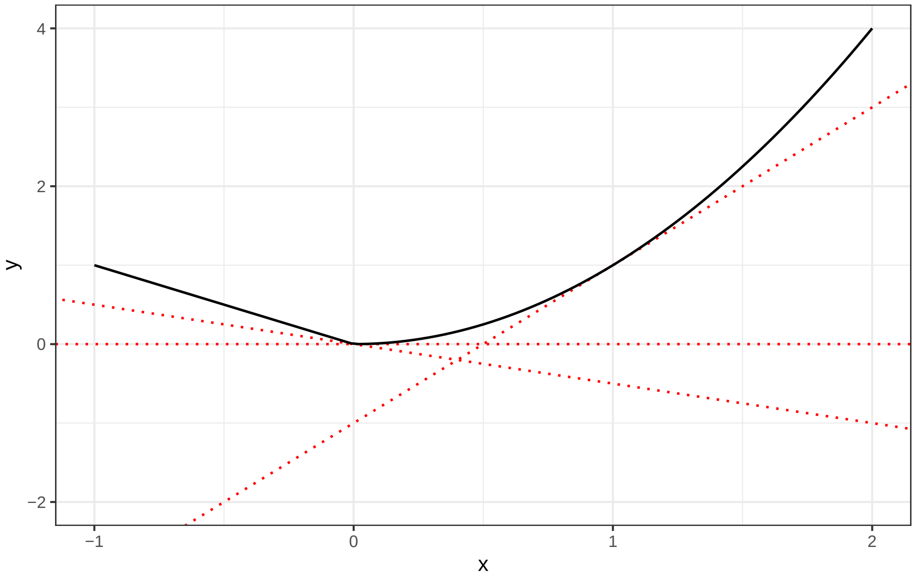

Figure 3.2: The function $f : x \mapsto \mathbb{I}_{x<0}|x| + \mathbb{I}_{x \geq 0} x^2$ is convex, continuous. It is differentiable everywhere except at $x = 0$. The dotted lines define affine functions that are below the cruve $y = f(x)$. The dotted lines define supporting hyperplanes for the epigraph of f.

Check that a convex function f is lower semi-continuous iff sets $\{x : f(x) \leq t\}$ are closed intervals for all $t \in \mathbb{R}$.

The next result warrants that any convex lower semi-continuous has a dual representation. This dual representation is a precious tool when comparing expectation of random variables.

Theorem 3.5 (Fenchel-Legendre duality). ***Let f be a convex lower-semi-continuous function on $\mathbb{R}$ with a closed domain.***

The dual function f^* of f is defined over $\mathbb{R}$ by

$$f^*(y) = \sup_{x \in \mathit{Dom}(f)} xy - f(x)\,.$$

Then

- ***f^* is convex***
- ***f^* is lower-semi-continuous***
- ***If $f^*(y) = xy - f(x)$ then y is a sub-gradient of f at x.***
- ***If y is a sub-gradient of f at x, $f^*(y) = xy - f(x)$.***
- ***$f = (f^*)^*$, the dual function of the dual function equals the original function: $f(x) = \sup_y xy - f^*(y)$.***

The next dual pairs will be used in several places.

- if $f(x) = \frac{|x|^p}{p}$ $(p > 1)$, then $f^*(y) = \frac{|y|^q}{q}$ where $q = p/(p-1)$.
- if $f(x) = |x|$, then $f^*(y) = 0$ for $y \in [-1, 1]$ and ∞ for $|y| > 1$.
- if $f(x) = \exp(x)$ then $f^*(y) = y \log y - y$ for $y > 0$, $f^*(y) = \infty$ for $y < 0$

Proof. The fact that f^* is $\mathbb{R} \cup \{\infty\}$-valued and convex is immediate.

To check lower semi-continuity, assume $y_n \to y$, with $y_n \in \mathrm{Dom}(f^*)$ and $f^*(y) > \liminf_n f^*(y_n)$.

Assume first that $y \in \mathrm{Dom}(f^*)$. Then for some sufficiently large m and some $x \in \mathrm{Dom}(f)$

$$f^*(y) \geq xy - f(x) - \frac{1}{m} > \liminf_n f^*(y_n) \geq \liminf_n y_n x - f(x) = yx - f(x)$$

which is contradictory.

Assume now that $y \notin \mathrm{Dom}(f^*)$ and $\liminf_n f^*(y_n) < \infty$. Extract a subsequence $(y_{m_n})_n$ such that $\lim_n f^*(y_{m_n}) = \liminf_n f^*(y_n)$. There exists $x \in \mathrm{Dom}(f)$ such that

$$f^*(y) > xy - f(x) > \liminf_n f^*(y_n) = \lim_n f^*(y_{m_n}) \geq \lim_n xy_{m_n} - f(x) = xy - f(x)$$

which is again contradictory.

The fact that y is a sub-gradient of f at x if $f^*(y) = xy - f(x)$ is a rephrasing of the definition of sub-gradients.

Note that if $x \in \mathrm{Dom}(f)$ and $y \in \mathrm{Dom}(f^*)$ then $f(x) + f^*(y) \geq xy$.

This observation entails that $(f^*)^*(x) \leq f(x)$ for all $x \in \mathrm{Dom}(f)$. If there existed some $x \in \mathrm{Dom}(f)$ with $(f^*)^*(x) > x$, there would exist some $y \in \mathrm{Dom}(f^*)$ with $xy - f^*(y) > f(x)$ which is not possible.

In order to prove that that $(f^*)^*(x) \geq f(x)$ for all $x \in \mathrm{Dom}(f)$, we rely on the convexity, lower semi-continuity of f and f^* and the closure of $\mathrm{Dom}(f)$. Under these conditions, every point x in $\mathrm{Dom}(f)$ has a sub-gradient y and this entails $f(x) + f^*(y) = xy$. □

Extend the notion of Fenchel-Legendre duality to lower-semi-continuous convex functions over $\mathbb{R}^k$.

Are all convex functions lower-semi-continuous? measurable?

Are all convex lower-semi-continuous functions measurable?

It is possible to define f^* as $f^*(y) = \sup_x xy - f(x)$ even if f is not convex and lower semi-continuous. The function f^* retains the convexity and lower semi-continuity properties. But $f \neq (f^*)^*$, we only get $f \geq (f^*)^*$. Indeed $(f^*)^*$ is the largest convex minorant of f.

Theorem 3.6 (Jensen's inequality). ***Let X be a real-valued random variable and $f : \mathbb{R} \to \mathbb{R}$ be*** convex, lower-semi-continuous ***such that the closed set Dom(f) $\subseteq$ supp($\mathcal{L}(X)$) and $\mathbb{E}|f(X)| < \infty$., then***

$$f(\mathbb{E}X) \leq \mathbb{E}f(X) .$$

In view of the definition of convexity and of the fact that taking expectation extends the idea of taking a convex combination, Jensen's inequality is not a surprise.

Proof.

$$\begin{aligned}\mathbb{E}f(X) &= \mathbb{E}(f^*)^*(X) \\ &= \mathbb{E}\Big[\sup_y \Big(yX - f^*(y)\Big)\Big] \\ &\geq \sup_y \Big(y\mathbb{E}X - f^*(y)\Big) \\ &= (f^*)^*\Big(\mathbb{E}X\Big) \\ &= f\Big(\mathbb{E}X\Big) .\end{aligned}$$

□

In the argument above, it is not ***a priori*** obvious that $\sup_y \Big(yX - f^*(y)\Big)$ is measurable, since the supremum is taken over a non-countable collection. Check that this is not an issue.

We will see many applications of Jensen's inequality:

- comparison of sampling with replacement with sampling without replacement (comparison of binomial and hypergeometric tails)
- Cauchy-Schwarz and Hölder's inequalities
- Derivation of maximal inequalities
- Non-negativity of relative entropy
- Derivation of Efron-Stein-Steele's inequalities
- ...

3.8 Variance

The variance (when it is defined) is an index of dispersion of the distribution of a random variable.

Proposition 3.6 (Characterizations of variance). ***Let X be a random variable over some probability space. The variance of X is finite iff $\mathbb{E}X^2 < \infty$ and it may be defined using the netx three equalities:***

$$\begin{aligned}\text{var}(X) &= \mathbb{E}\left[(X - \mathbb{E}X)^2\right] \\ &= \inf_{a\in\mathbb{R}} \mathbb{E}\left[(X-a)^2\right] \\ &= \mathbb{E}X^2 - (\mathbb{E}X)^2 .\end{aligned}$$

We need to check that three right-hand-side are finite if one of them is, and that when they are finite, they are all equal.

Proof. Assume $\mathbb{E}X^2 < \infty$, as $|X| \leq \frac{X^2}{2} + \frac{1}{2}$, this entails $\mathbb{E}|X| < \infty$. If $\mathbb{E}X^2 < \infty$ then so is $\mathbb{E}|X|$. The right-hand-side on the third line is finite if $\mathbb{E}X^2 < \infty$. As $(x-b)^2 \leq 2x^2 + 2b^2$ for all x, b, The right-hand-side on the first line, the infimum on the second line are finite when $\mathbb{E}X^2 < \infty$.

As $X^2 \leq 2(X - \mathbb{E}X)^2 + 2(\mathbb{E}X)^2$, $\mathbb{E}X^2 < \infty$ if $\mathbb{E}\left[(X-\mathbb{E}X)^2\right] < \infty$.

Assume now that $\mathbb{E}X^2 < \infty$.

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

As $(a-\mathbb{E}X)^2 \geq 0$, we have established that $\mathbb{E}\left[(X-\mathbb{E}X)^2\right] = \inf_{a\in\mathbb{R}} \mathbb{E}\left[(X-a)^2\right]$. Moreover, the infimum is a minimum, it is achieved at a single point $\mathbb{E}X$. □

The first and second characterizations of variance assert that the expectation minimizes the average quadratic error. A fact of great importance in Statistics.

Check that if $P\{X \in [a,b]\} = 1$, then $\operatorname{var}(X) \leq \frac{(b-a)^2}{4}$, .

3.9 Higher moments

In this Section we relate $\mathbb{E}|X|^p$ with $\mathbb{E}|X|^q$ for different values of $p, q \in \mathbb{R}_+$. Our starting point is small technical result in real analysis.

Proposition 3.7 (Young's inequality). ***Let** $p, q > 1$ **be** conjugate ($1/p + 1/q = 1$), **and** $x, y > 0$, **then***

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q} .$$

Proof. Note that if p and q are conjugate, then $q = p/(p-1)$ and $(p-1)(q-1) = 1$.

It suffices to check that for all $x, y > 0$,

$$\frac{x^p}{p} \geq xy - \frac{y^q}{q} .$$

Fix $x > 0$, consider the function over $[0, \infty)$ defined by

$$z \mapsto xz - \frac{z^q}{q} .$$

This function is differentiable with derivative $x - z^{q-1} = x - z^{1/(p-1)}$. It achieves its maximum at $z = x^{p-1}$ and the maximum is equal to

$$xx^{p-1} - \frac{x^{q(p-1)}}{q} = x^p - \frac{x^p}{q} = \frac{x^p}{p} .$$

□

Figure 3.3 displays a graphic proof of Young's inequality.

A special case of Young inequality is obtained by taking $p = q = 2$.

We are now in a position to prove three fundamental inequalities: Cauchy-Schwarz, Hölder and Minkowski.

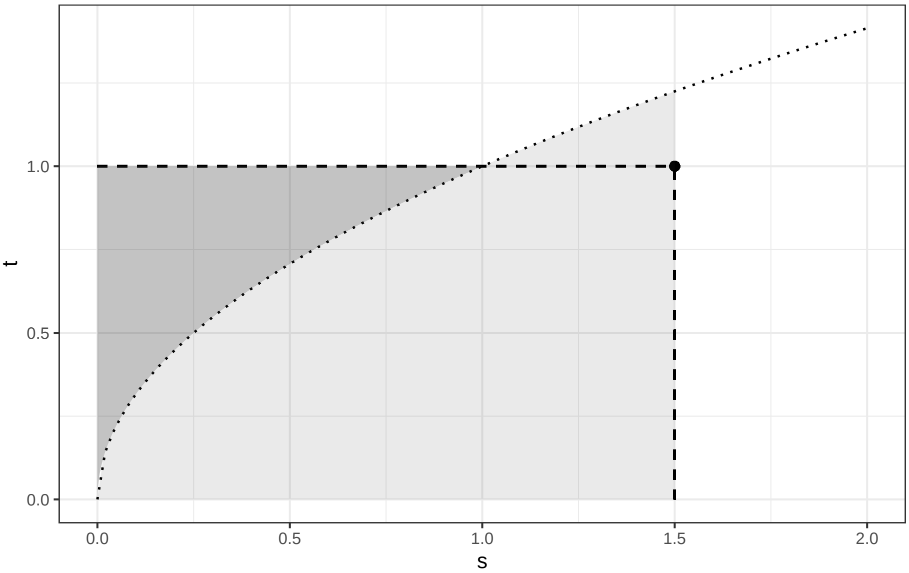

Figure 3.3: Graphical illlustration of Young's inequality. We choose $p =` rp`$ and $q =` rq`$, $x =` rx`$ and $y =` ry`$. The black point is located at $(x,y)^T$. The product xy equals the area of the rectangle located between the origin and $(x,y)^T$ (delimited by the dashed segments). The dotted line represents function $s \mapsto s^{p-1}$, and interchanging the axes, the function $t \mapsto t^{q-1} = t^{1/(p-1)}$. The area of the light grey surface under the dotted line equals $\frac{x^p}{p} = \int_0^x s^{p-1}\mathrm{d}s$, while the area of the darker grey surface below line $y = 1$ and above the dotted line, equals $\frac{y^q}{q} = \int_0^y t^{q-1}\mathrm{d}t$. The union of the two disjoint surfaces covers the rectangle located between the origin and $(x,y)^\top$. Equality occurs when the dotted line passes though $(x,y)^\top$, that is when $y = x^{p-1}$.

Theorem 3.7 (Cauchy-Schwarz). *Let X and Y be two random variables on the same probability space. Assume both $\mathbb{E}X^2$ and $\mathbb{E}Y^2$ are finite. Then*

$$\mathbb{E}[XY] \leq \sqrt{\mathbb{E}X^2} \times \sqrt{\mathbb{E}Y^2}\,.$$

Proof. If either $\sqrt{\mathbb{E}X^2} = 0$ or $\sqrt{\mathbb{E}Y^2} = 0$, the inequality is trivially satisfied.

So, without loss of generality, assume $\sqrt{\mathbb{E}X^2} > 0$ and $\sqrt{\mathbb{E}Y^2} > 0$. Then, because $ab \leq a^2/2 + b^2/2$, for all real a, b, everywhere,

$$\frac{|XY|}{\sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2}} \leq \frac{|X|^2}{2\mathbb{E}X^2} + \frac{|Y|^2}{2\mathbb{E}Y^2}\,.$$

Taking expectation on both sides leads to the desired result. □

Why is the inequality trivially satisfied if $\sqrt{\mathbb{E}X^2} = 0$?

Theorem 3.7 tells us that if X and Y are square-integrable, then XY is integrable.

Hölder's inequality generalizes Cauchy-Schwarz inequality. Indeed, Cauchy-Schwarz inequality is just Hölder's inequality for $p = q = 2$ (2 is its own conjugate).

Theorem 3.8 (Hölder's inequality). *Let X and Y be two random variables on the same probability space. Let $p, q > 1$ be* conjugate *($1/p + 1/q = 1$), assume both $\mathbb{E}|X|^p$ and $\mathbb{E}|Y|^q$ are finite. Then we have*

$$\mathbb{E}[XY] \leq \left(\mathbb{E}|X|^p\right)^{1/p} \times \left(\mathbb{E}|Y|^q\right)^{1/q} .$$

Proof. If either $\mathbb{E}|X|^p = 0$ or $\mathbb{E}|Y|^q = 0$, the inequality is trivially satisfied.

Assume that $\mathbb{E}|X|^p > 0$ and $\mathbb{E}|Y|^q > 0$.

Follow the proof of Cauchy-Schwarz inequality, but replace $2ab \leq a^2 + b^2$ by Young's inequality:

$$ab \leq \frac{|a|^p}{p} + \frac{|b|^q}{q} \qquad \forall a, b \in \mathbb{R}$$

if $1/p + 1/q = 1$.

The inequality below is a consequence of Young's inequality and of the monotonicity of expectation:

$$\begin{aligned}\frac{\mathbb{E}|XY|}{\mathbb{E}[|X|^p]^{1/p}\mathbb{E}[|Y|^q]^{1/q}} & = \mathbb{E}\left[\frac{|X|}{\mathbb{E}[|X|^p]^{1/p}} \frac{|Y|}{\mathbb{E}[|Y|^q]^{1/q}}\right] \\ & \leq \mathbb{E}\left[\frac{|X|^p}{p\mathbb{E}[|X|^p]} + \frac{|Y|^q}{q\mathbb{E}[|Y|^q]}\right] \\ & = \frac{1}{p} + \frac{1}{q} \\ & = 1 .\end{aligned}$$

□

For $1 \leq p < q$,

$$\mathbb{E}\Big[|X|^p\Big]^{1/p} \leq \mathbb{E}\Big[|X|^q\Big]^{1/q} .$$

For $p \in [0, \infty)$ $X \mapsto (\mathbb{E}|X|^p)^{1/p}$ defines a semi-norm on the set of random variables for which $(\mathbb{E}|X|^p)^{1/p}$ is finite. Minkowski's inequality asserts that $X \mapsto (\mathbb{E}|X|^p)^{1/p}$ satisfies the triangle inequality.

Theorem 3.9 (Minkowski's inequality). *Let X, Y be two real-valued random variables defined on the same probability space. Let $p \in [1, \infty)$. Assume that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^p < \infty$. Then we have:*

$$\left(\mathbb{E}[|X+Y|^p]\right)^{1/p} \leq \left(\mathbb{E}[|X|^p]\right)^{1/p} + \left(\mathbb{E}[|Y|^p]\right)^{1/p}$$

which entails $\mathbb{E}|X+Y|^p < \infty$.

The proof of Theorem 3.9 follows from Hölder's inequality (Theorem 3.8).

Proof. The inequality below also follows from triangle inequality on $\mathbb{R}$, monotonicity. The last equality follows from linearity of expectation:

$$\begin{aligned}\mathbb{E}\Big[|X+Y|^p\Big] & \leq \mathbb{E}\Big[(|X| + |Y|) \times |X+Y|^{p-1}\Big] \\ & = \mathbb{E}\Big[|X| \times |X+Y|^{p-1}\Big] + \mathbb{E}\Big[|Y| \times |X+Y|^{p-1}\Big] .\end{aligned}$$

This is enough to handle the case $p = 1$.

From now on, assume $p > 1$. Hölder's inequality entails the next inequality and a similar upper bound for $\mathbb{E}\Big[|Y| \times |X+Y|^{p-1}\Big]$.

$$\mathbb{E}\Big[|X| \times |X+Y|^{p-1}\Big] \leq \mathbb{E}\Big[|X|^p\Big]^{1/p} \times \mathbb{E}\Big[|X+Y|^p\Big]^{(p-1)/p}$$

Summing the two upper bounds, we obtain

$$\mathbb{E}\Big[|X+Y|^p\Big] \quad \leq \left(\mathbb{E}\Big[|X|^p\Big]^{1/p} + \mathbb{E}\Big[|Y|^p\Big]^{1/p}\right) \times \mathbb{E}\Big[|X+Y|^p\Big]^{(p-1)/p}.$$

This prove's Minkowski's inequality for $p > 1$. □

3.10 Median and interquartile range

Robust and non-robust indices of location.

Let X be a real random variable over some probability space. Let F be the cumulative distribution function of X. The median of the distribution of X is $F^{\leftarrow}(1/2)$.

The median minimizes the mean absolute deviation.

If m is such that $P\{X > m\} = P\{X < m\}$ then m is median of the distribution of X, and if X is integrable:

$$\mathbb{E}\Big|X - m\Big| = \min_{a \in \mathbb{R}} \mathbb{E}\Big|X - a\Big|$$

Proof. Assume $a < m$,

$$\begin{aligned}\mathbb{E}\left[\Big|X-a\Big| - \Big|X-m\Big|\right] \quad & = -(m-a)P(-\infty,a] + \int_{(a,m]}(2X-(a+m))\mathrm{d}P(x) + (m-a)P(m,\infty) \\ & \geq -(m-a)P(-\infty,a] - (m-a)P(a,m] + (m-a)P(m,\infty) \\ & = (m-a)\Big(P(m,\infty) - P(-\infty,m]\Big) \\ & = 0\,.\end{aligned}$$

The same line of reasoning allows to handle the case $a > m$ and to conclude. □

Combining three of the inequalities we have just proved, allows us to establish an interesting connection between expectation, median and standard deviation.

Theorem 3.10 (Lévy's inequality)**.** *Let m be the median of the distribution of X, a square-integrable random variable over some probability space. Then*

$$\Big|m - \mathbb{E}X\Big| \leq \sqrt{\mathrm{var}(X)}\,.$$

The robust and non-robust indices of location differ by at most the standard deviation, which may be infinite.

Proof. By convexity of $x \mapsto |x|$, we have

$$\begin{aligned}\Big|m - \mathbb{E}X\Big| \quad & \leq \mathbb{E}\Big|m - X\Big| \\ & \text{by Jensen's inequality} \\ & \leq \mathbb{E}\Big|\mathbb{E}X - X\Big| \\ & \text{the median minimizes the mean absolute error} \\ & \leq \left(\mathbb{E}\Big|\mathbb{E}X - X\Big|^2\right)^{1/2} \\ & \text{by Cauchy-Schwarz inequality.}\end{aligned}$$

□

The mean and the median may differ. First the median is always defined, while the mean may not. Think for example of the standard Cauchy distribution which has density $\frac{1}{\pi}\frac{1}{1+x^2}$ over $\mathbb{R}$. If X is Cauchy distributed, then $\mathbb{E}|X| = \infty$. The mean is not defined. But as the density is a pair function, X is symmetric (X and $-X$ are distributed the same way), and this implies that the median of (the distribution) of X is 0.

Consider the exponential distribution with density $\exp(-x)$ over $[0, \infty)$, it has mean 1, median $\log(2)$, and variance 1. If we turn to exponential distribution with density $\lambda \exp(-\lambda x)$, it has mean $1/\lambda$, median $\log(2)/\lambda$, and variance $1/\lambda^2$. Lévy's inequality does not tell more that what we can compute with bare hands.

Finally consider Gamma distributions with shape parameter p and intensity parameter λ. It has mean p/λ, variance p/λ^2. The median is not easily computed though we can easily check that it is equal to $g(p)/\lambda$ where $g(p)$ is the median of the Gamma distribution with parameters p and 1. Lévy's inequality tells us that $|g(p) - p| \leq \sqrt{p}$.

3.11 $\mathcal{L}_p$ and L_p spaces

Let $p \in [1, \infty)$. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Define $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ (often abbreviated to $\mathcal{L}_p(P)$ or even $\mathcal{L}_p$ when there is no ambiguity) as

$$\mathcal{L}_p(\Omega, \mathcal{F}, P) = \Big\{ X : X \text{ is a real random variable over } (\Omega, \mathcal{F}, P), \quad \mathbb{E}|X|^p < \infty \Big\} .$$

Let $\|X\|_p$ be defined by $\|X\|_p = \Big(\mathbb{E}|X|^p\Big)^{1/p}$.

Let $\mathcal{L}_0(\Omega, \mathcal{F}, P)$ denote the vector space of random variables over $(\Omega, \mathcal{F}, P)$.

We first notice that sets $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ form a nested sequence.

Proposition 3.8. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, then for $1 \leq p \leq q < \infty$:*

1. $\|X\|_p < \|X\|_q$.
2. $\mathcal{L}_q(\Omega, \mathcal{F}, P) \leq \mathcal{L}_p(\Omega, \mathcal{F}, P)$.

Proof. Assume $1 \leq p \leq q < \infty$, as $x \mapsto x^{q/p}$ is convex on $[0, \infty)$ by Jensen's inequality (Theorem 3.6), we have

$$\mathbb{E}[|X|^p]^{q/p} \quad \leq \mathbb{E}[|X|^q] .$$

This establishes 1.) And 2.) is an immediate consequence of 1. □

Proposition 3.8 is a about inclusion of sets. The next theorem summarizes several points: that sets $\mathcal{L}_p$ are linear subspaces of $\mathcal{L}_0$, and that they are complete as pseudo-metric (pseudo-normed) spaces.

For $p \in [1, \infty)$, let $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ and $\|\cdot\|_p$ be defined as above. Then,

1. $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ is a linear subspace of the space of real random variables.
2. $\|\cdot\|_p$ is a pseudo-norm on $\mathcal{L}_p(\Omega, \mathcal{F}, P)$.
3. If $(X_n)_n$ is a sequence in $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ that satisfies

$$\lim_n \sup_{m \geq n} \Big| X_n - X_m \Big|_p = 0$$

then there exists $X \in \mathcal{L}_p(\Omega, \mathcal{F}, P)$ such that $\lim_n \|X_n - X\|_p = 0$.

4. There exists a subsequence $(X_{m_n})_n$ such that $X_{m_n} \to X$ P-almost surely.

In a pseudo-metric space, to prove that a Cauchy sequence converges, it is enough to check convergence of a subsequence. Picking a convenient subsequence, and possibly relabeling elements, we may assume $\left\|X_n - X_m\right\|_p \leq 2^{-n\wedge m}$ for all n, m.

First Borell-Cantelli Lemma

Let $(A_n)_n$ be a sequence of events from some probability space $(\Omega, \mathcal{F}, P)$. Assume $\sum_n P(A_n) < \infty$ then, with probability 1, only finetely many events A_n are realized:

$$P\left\{\omega : \sum_n \mathbb{I}_{A_n}(\omega) < \infty\right\} = 1\,.$$

Proof. The event $\left\{\omega : \sum_n \mathbb{I}_{A_n}(\omega) = \infty\right\}$ coincides with $\cap_n \cup_{m\geq n} A_n$:

$$P\left\{\sum_n \mathbb{I}_{A_n}(\omega) = \infty\right\} = P(\cap_n \cup_{m\geq n} A_n)\,.$$

Now, the sequence $(\cup_{m\geq n} A_n)_n$ is monotone decreasing: $\lim_n \downarrow \cup_{m\geq n} A_n = \cap_n \cup_{m\geq n} A_n$.

By Fatou's Lemma,

$$\begin{aligned} \mathbb{E} \lim_m \mathbb{I}_{\cup_{m\geq n} A_m} & = \mathbb{E} \liminf_n \mathbb{I}_{\cup_{m\geq n} A_m} \\ & \leq \liminf_n \mathbb{E}\mathbb{I}_{\cup_{m\geq n} A_m} \\ & \leq \liminf_n \sum_{m\geq n} P(A_m) \\ & = 0\,. \end{aligned}$$

The last equation comes from the fact that the remainders of a convergent series are vanishing. □

Proof. Points 1) and 2) follow from Minkowski's inequality. This entails that $\|\cdot\|_p$ defines a pseudo-norm on $\mathcal{L}_p$. If two random variables X, Y from $\mathcal{L}_p$ satisfy $\|X - Y\|_p = 0$, then $X = Y$ P-a.s.

To establish 3), we need to check that the sequence converges almost surely, and that an almost sure limit belongs to $\mathcal{L}_p$.

Define event A_n by

$$A_n = \left\{\omega : \left|X_n(\omega) - X_{n+1}(\omega)\right| > \frac{1}{n^2}\right\}.$$

By Markov inequality,

$$P(A_n) \leq \mathbb{E}\left[n^{2p}\left|X_n - X_m\right|^p\right] \leq n^{2p} 2^{-np}\,.$$

Hence, $\sum_{n\geq 1} P(A_n) < \infty$. By the first Borel-Cantelli Lemma, on some event E with probability 1, only finitely many A_n are realized.

If $\omega \in E$, the condition $\left|X_n(\omega) - X_{n+1}(\omega)\right| > \frac{1}{n^2}$ is realized for only finitely many indices n. Thus the real-valued sequence $(X_n(\omega))_n$ is a Cauchy sequence. It has a limit we denote $X(\omega)$. If $\omega \notin E$, we agree on $X(\omega) = 0$. On Ω, we have

$$X(\omega) = \lim_n \mathbb{I}_E(\omega) X(\omega)\,.$$

A limit of random variables is a random variable. Hence X is a random variable.

It remains to check that $X \in \mathcal{L}_p$. Note first that

$$\left| \|X_m\|_p - \|X_n\|_p \right| \leq \|X_m - X_n\|_p .$$

Hence $(\|X_n\|_p)_n$ is a Cauchy sequence and converges to some finite limit. As

$$|X(\omega)| \leq \liminf |X_n(\omega)|$$

by Fatou's Lemma

$$\mathbb{E}|X|^p \leq \liminf \mathbb{E}|X_n|^p < \infty .$$

Hence $X \in \mathcal{L}_p$.

Finally we check that $\lim_m \|X_n - X\|_p = 0$. By Fatou's lemma again,

$$\mathbb{E}\left|X - X_m\right|^p \leq \liminf_n \mathbb{E}\left|X_n - X_m\right|^p$$

Hence

$$\lim_m \mathbb{E}\left|X - X_m\right|^p \leq \lim_m \liminf_n \mathbb{E}\left|X_n - X_m\right|^p = 0 .$$

□

Can we extend the almost sure convergence to the whole sequence? This is not the case. Consider $([0,1], \mathcal{B}([0,1]), P)$ where P is the uniform distribution. For $k = j + n(n-1)/2$, $1 \leq j \leq n$, let $X_n = \mathbb{I}_{[(j-1)/n, j/n]}$. The sequence X_n converges to 0 in $\mathcal{L}_p$ for all $p \in [1, \infty)$. Indeed $\|X_k\|_p = n^{-p}$ for $k = j + n(n-1)/2, 1 \leq j \leq n$. For any $\omega \in [0,1]$, the sequence $X_n(\omega)$ oscillates between 0 and 1 infinitely many times.

$\mathcal{L}_p$ provide us with a bridge between probability and analysis. In analysis, the fact that $\|\cdot\|_p$ is just a pseudo-norm leads to consider L_p spaces. L_p spaces are defined from $\mathcal{L}_p$ spaces by taking equivalence classes of random variables. Indeed, define relation $\equiv$ over $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ by $X \equiv X'$ iff $P\{X = X'\} = 1$. This relation is an equivalence relation (reflexive, symmetric and transitive). If $X \equiv X'$ and $Y \equiv Y'$, then $\|X - Y\|_p = \|X' - Y\|_p = \|X' - Y'\|_p$. $L_p(\Omega, \mathcal{F}, P)$ is the quotient space of $\mathcal{L}_p$ by relation $\equiv$. We have the fundamental result.

Theorem 3.11. *For $p \in [1, \infty)$, $L_p(\Omega, \mathcal{F}, P)$ equiped with $\|\cdot\|_p$ is a complete normed space (Banach space).*

This eventually allows us to invoke theorems from functional analysis.

3.12 Bibliographic remarks

Dudley (2002) gives a self-contained and thorough treatment of measure and integration theory with probability theory in mind.

Hiriart-Urruty & Lemaréchal (1993) is an excellent and accessible reference on convexity.

Chapter 4

Families of discrete distributions

The goal of this lesson is getting acquainted with important families of distributions and to get familiar with distributional calculus. Probability distributions will be presented through distribution functions, probability mass functions (discrete distribution), densities (when available).

4.1 Bernoulli and Binomial

Definition 4.1. A Bernoulli distribution is a probability distribution P on $\Omega = \{0, 1\}$. The parameter of P is $P\{1\} \in [0, 1]$.

A Bernoulli distribution is completely defined by its parameter.

Assume $\Omega' = \{0, 1\}^n$.

Definition 4.2 (Binomial distribution)**.** A binomial distribution with parameters $n \in \mathbb{N}, p \in [0, 1]$ (n is size and p is success) is a probability distribution P on $\Omega = \{0, 1, 2, \dots, n\}$, defined by

$$P\{k\} = \binom{n}{k} p^k (1-p)^k$$

The connexion between Bernoulli and Binomial distributions is obvious: a Bernoulli distribution is a Binomial distribution with size parameter equal to 1. This connexion goes further: the sum of independent Bernoulli random variables with same success parameter is Binomial distributed.

Let $X_1, X_2, \dots, X_n$ be independent, identically distributed Bernoulli random variables with success parameter $p \in [0, 1]$, then $Y = \sum_{i=1}^n X_i$ is distributed according to a Binomial disctribution with size parameter n and success probability p.

Proof. For $k \in 0, \dots, n$

$$
\begin{aligned}
P\Big\{\sum_{i=1}^{n} X_i = k\Big\} &= \sum_{x_1,\ldots,x_n\in\{0,1\}^p} \mathbb{I}_{\sum_{i=1}^n x_i = k} P\Big\{\wedge_{i=1}^n X_i = x_i\Big\} \\
&= \sum_{x_1,\ldots,x_n\in\{0,1\}^p} \mathbb{I}_{\sum_{i=1}^n x_i = k} \prod_{i=1}^{n} P\Big\{X_i = x_i\Big\} \\
&= \sum_{x_1,\ldots,x_n\in\{0,1\}^p} \mathbb{I}_{\sum_{i=1}^n x_i = k} \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{x_1,\ldots,x_n\in\{0,1\}^p} \mathbb{I}_{\sum_{i=1}^n x_i = k}\, p^k(1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k}\,.
\end{aligned}
$$

□

This observation facilitates the computation of moments of Binomial distribution.

The expected value of a Bernoulli distribution with parameter p is p! Its variance is $p(1-p)$.

By linearity of expectation, the expected value of the binomial distribution with parameters n and p is np. The variance of a sum of independent random variables is the sum of the variances, hence the variance of he binomial distribution with parameters n and p is $np(1-p)$.

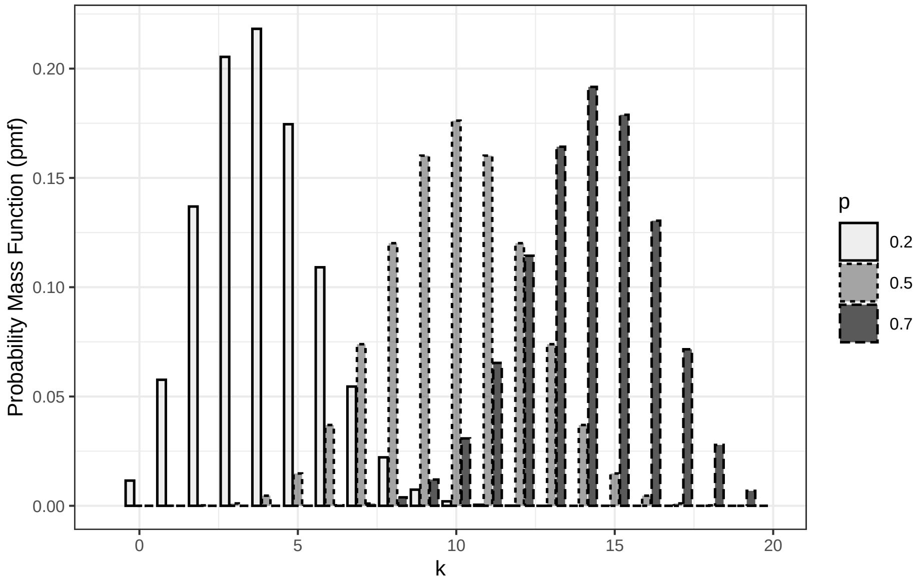

Figure 4.1: Binomial probability mass functions with $n = 20$ and different values of p : $.5, .7, .2$.

More on wikipedia.

Binomial distributions with the same success parameter

Let X, Y be independent over probability space $(\Omega, \mathcal{F}, P)$ and distributed according to $\text{Bin}(n_1, p)$ and $\text{Bin}(n_2, p)$.

Then $X + Y$ is distributed according to $\text{Bin}(n_1 + n_2, p)$.

Check the preceding proposition.

4.2 Poisson

The Poisson distribution appears as a limit of Binomial distributions in a variety of circumstances connected to rare events phenomena.

Definition 4.3. A Poisson distribution with parameter $\lambda > 0$ is a probability distribution P on $\Omega = \mathbb{N}$ with

$$P\{k\} = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$$

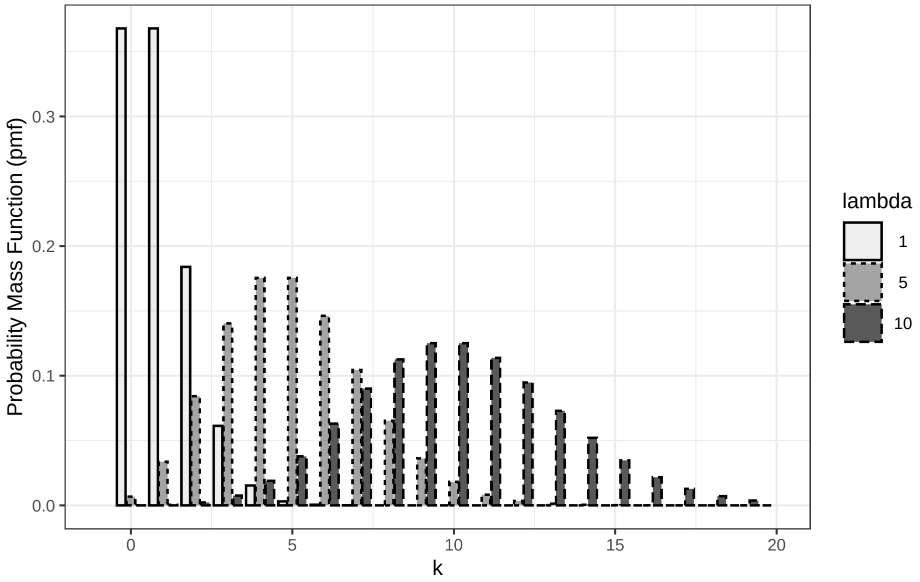

Figure 4.2: Poisson probability mass functions with different values of parameter: $1, 5, 10$. Recall that the parameter of a Poisson distribution equals its expectation and its variance. The probability mass function of a Poisson distribution achieves its maximum (called the mode) close to its expectation.

The expected value of the Poisson distribution with paramenter λ is λ. The variance of a Poisson distribution is equal to its expected value.

$$\begin{aligned}\mathbb{E}X &= \sum_{n=0}^{\infty}\mathrm{e}^{-\lambda}\frac{\lambda^n}{n!}\times n\\ &= \lambda\times\sum_{n=1}^{\infty}\mathrm{e}^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!}\\ &= \lambda\,.\end{aligned}$$

Let X, Y be independent and Poisson distributed over probability space $(\Omega, \mathcal{F}, P)$, then $X + Y$ is Poisson distributed.

Proof. We check the proposition in the simplest and most tedious way. We compute the probability mass function of the distribution of $X+Y$. Assume $X \sim \mathrm{Po}(\lambda), Y \sim \mathrm{Po}(\mu)$.

For each $k \in \mathbb{N}$:

$$\begin{aligned}
\Pr\{X+Y=k\} &= \Pr\{\bigvee_{m=0}^{k} (X=m \wedge Y=k-m)\}\\
&= \sum_{m=0}^{k} \Pr\{X=m \wedge Y=k-m\}\\
&= \sum_{m=0}^{k} \Pr\{X=m\} \times \Pr\{Y=k-m\}\\
&= \sum_{m=0}^{k} \mathrm{e}^{-\lambda}\frac{\lambda^m}{m!}\mathrm{e}^{-\mu}\frac{\mu^{k-m}}{(k-m)!}\\
&= \mathrm{e}^{-\lambda-\mu}\frac{(\lambda+\mu)^k}{k!}\sum_{m=0}^{k}\frac{k!}{m!(k-m)!}\left(\frac{\lambda}{\lambda+\mu}\right)^m\left(\frac{\mu}{\lambda+\mu}\right)^{k-m}\\
&= \mathrm{e}^{-\lambda-\mu}\frac{(\lambda+\mu)^k}{k!}\sum_{m=0}^{k}\binom{k}{m}\left(\frac{\lambda}{\lambda+\mu}\right)^m\left(\frac{\mu}{\lambda+\mu}\right)^{k-m}\\
&= \mathrm{e}^{-\lambda-\mu}\frac{(\lambda+\mu)^k}{k!}\left(\frac{\lambda}{\lambda+\mu}+\frac{\mu}{\lambda+\mu}\right)^k\\
&= \mathrm{e}^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!}
\end{aligned}$$

The last expression if the pmf of $\mathrm{Po}(\lambda+\mu)$ at k. □

Check that the ***mode*** (maximum) of a Poisson probability mass function with parameter λ is achieved at $k=\lfloor\lambda\rfloor$. It is always unique?

Check that the median of a Poisson distribution with integer parameter λ is not smaller than λ.

4.3 Geometric

A geometric distribution is a probability distribution over $\mathbb{N} \subset \{0,1\}$. It depends on a parameter $p>0$.

Assume we are allowed to toss a biased coin infinitely many times. The number of times we have to toss the coin ***until*** we get a ***head*** is geometrically distributed.

Let X be distributed according to a geometric distribution with parameter p. The geometric probability distribution is easily defined by its tail function. In the event $X>k$, the first k outcomes have to be ***tail***.

$$P\{X>k\}=(1-p)^k$$

The probability mass function of the geometric distribution follows:

$$P\{X=k\}=(1-p)^{k-1}-(1-p)^k=p\times(1-p)^{k-1} \qquad \text{for } k=1,2,\ldots$$

On average, we have to toss the coin p times until we get a ***head***:

$$\mathbb{E}X=\sum_{k=0}^{\infty}P\{X>k\}=\frac{1}{p}$$

It is also possible to define geometric random variables as the number of times we have to toss the coin **before** we get a *head*. This requires modifying quantile function, probability mass function, expectation, and so on. This is the convention R uses.

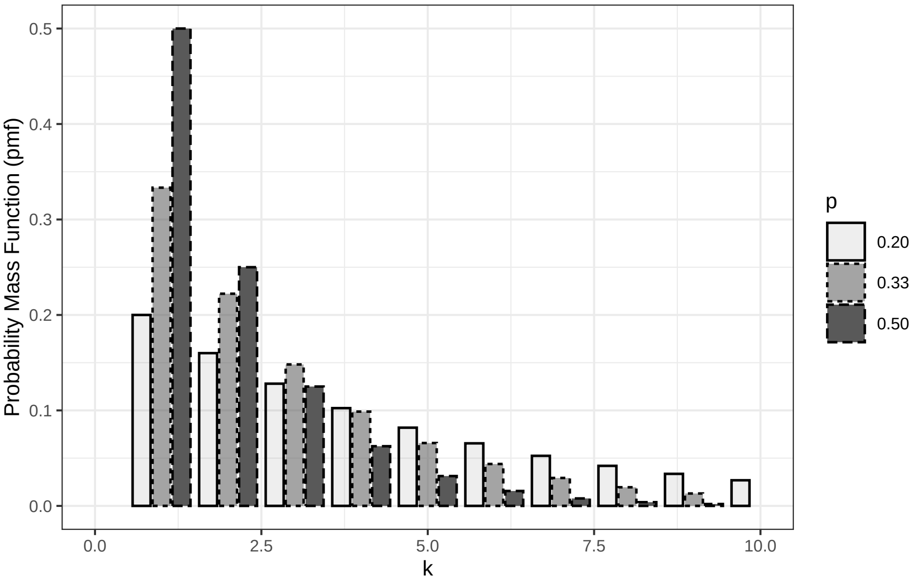

Figure 4.3: Geometric probability mass functions with different values of parameter p: $1/2, 1/3, 1/5$. The probability mass function equals $p \times (1-p)^{k-1}$ at $k \geq 1$. The mode is achieved at $k = 1$ whatever the value of p. The expectation equals $1/p$

Sums of independent geometric random variables are not distributed according to a geometric distribution.

Chapter 5

Discrete Conditioning

5.1 Roadmap

Conditioning is central to probabilistic reasoning. In this lesson, we investigate discrete conditioning. In this setting, the definition of conditional probability is not an issue. The definition of conditional expectation can be deceptively simple. Nevertheless the discrete setting lends itself to intuitive definitions and manipulations.

The simplest notion we meet is conditional probability with respect to a specific event with positive probability (Section 5.2). Conditional probability offers an intuitive interpretation of independence.

In Section 5.3 we state, check and discuss Bayes formula.

In Section 5.4, we define conditional expectation with respect to an atomic σ-algebra. This defines conditional expectation with respect to a discrete random variables.

In Section 5.5, we characterize conditional expectation as an optimal predictor. This characterization is very helpful when defining conditional expectation in the general setting.

5.2 Conditioning with respect to an event

Definition 5.1. Let P be a probability distribution on $(\Omega, \mathcal{F})$. Let $A \in \mathcal{F}$ be such that $P\{A\} > 0$. Let B be another event ($B \in \mathcal{F}$), the ***probability of** B **given** A* is defined as

$$P\{B \mid A\} = \frac{P\{A \cap B\}}{P\{A\}}.$$

If X is a standard Gaussian random variable on $(\Omega, \mathcal{F})$, and event A is defined by $\{\omega : X(\omega) \geq t\}$ for some $t \geq 0$, we may condition on event A and define $P\{B \mid A\}$ for $B = \{\omega : |X(\omega)| \geq 2t\}$.

We get

$$P\{B \mid A\} = \frac{P\{X \geq 2t\}}{P\{X \geq t\}}.$$

We may check the next proposition by considering once again the definition of probability distributions.

Proposition 5.1. *Let P be a probability distribution on $(\Omega, \mathcal{F})$.*

Let $A \in \mathcal{F})$ be such that $P\{A\} > 0$, then $P\{\cdot \mid A\}$ (P given A) defines a probability distribution over $(\Omega, \mathcal{F})$.

Proof. $P(\cdot \mid A)$ maps $\mathcal{F}$ to $[0,1]$.

We have $P(\Omega \mid A) = P(A \cap \Omega)/P(A) = P(A)/P(A) = 1$

Let $(B_n)_n$ be a monotone increasing sequence of events, then

$$\begin{aligned} P(\cup_n B_n \mid A) &= \tfrac{P((\cup_n B_n)\cap A)}{P(A)} \\ &= \tfrac{P(\cup_n (B_n \cap A))}{P(A)} \\ &= \tfrac{\lim_n P(B_n \cap A)}{P(A)} \\ &= \lim_n P(B_n \mid A)\,. \end{aligned}$$

□

We may consider the distribution of random variables on $(\Omega, \mathcal{F})$ under $P\{\cdot \mid A\}$. We compute the expectation of X under $P\{\cdot \mid A\}$:

$$\mathbb{E}_{P\{\cdot|A\}} X = \frac{\mathbb{E}[\mathbb{I}_A X]}{P\{A\}}\,.$$

This is often denoted by $\mathbb{E}[X \mid A]$, we will try to avoid this possibly misleading notation.

Example 5.1. Assume X is standard normally distributed. One may investigate the distribution of X^2 conditionnally on event $A = \{\omega : X(\omega) \geq t\}$. For $t > 1$, we have

$$\begin{aligned} \mathbb{E}_{P\{\cdot|X\geq t\}} X^2 &= \tfrac{\int_t^\infty x^2 \phi(x)\mathrm{d}x}{\int_t^\infty \phi(x)\mathrm{d}x} \\ &\leq \tfrac{t^2}{1-1/t} + 1\,. \end{aligned}$$

where the upper bound is obtained by repeated integration by parts.

The distribution of X given A is not Gaussian. Under A, X is very concentrated in the neighborhood of t, and tends to be more concentrated as t goes to infinity.

Knowing the probability distribution given event A enables to investigate independence of events with respect to A The next trivial proposition is worth reminding.

Proposition 5.2. *If A and $B \in \mathcal{F}$ satisfy $P\{A\} > 0$, then A and B are independent under P iff*

$$P\{B \mid A\} = P\{B\}.$$

5.3 Bayes formula

Bayes formula is sometimes used in probabilistic causation theory. This is a difficult matter. Causality is a subtle notion and we will refrain from making causal interpretations.

Proposition 5.3 (Bayes formula). *Let P be a probability distribution on $(\Omega, \mathcal{F})$, let $(A_i)_{i\in\mathcal{I}\subseteq\mathbb{N}}$ be a collection of pairwise disjoint events, with non-zero probability such that $\cup_{i\in\mathcal{I}} A_i = \Omega$ ($(A_i)_i$ form a complete system of events), let B be an event with non-zero probability, then for all $i \in \mathcal{I}$,*

$$P\{A_i \mid B\} = \frac{P\{A_i\} \times P\{B \mid A_i\}}{\sum_{j\in\mathcal{I}} P\{A_j\} \times P\{B \mid A_j\}}$$

Proof. By definition, $P\{A_i \mid B\} = P\{A_i \cap B\}/P\{B\} = P\{A_i\} \times P\{B \mid A_i\}/P\{B\}$.

Morever

$$\begin{aligned} P\{B\} &= P\{B \cap (\cup_{j\in\mathcal{J}} A_j)\} \\ &= P\{\cup_{j\in\mathcal{J}} (B \cap A_j)\} \\ &= \sum_{j\in\mathcal{J}} P\{B \cap A_j\} \\ &= \sum_{j\in\mathcal{J}} P\{A_j\} \times P\{B \mid A_j\}. \end{aligned}$$

□

In the preceding proposition, $P\{A_i\}$ is called the prior probability of A_i and $P\{A_i \mid B\}$ the posterior probability.

5.4 Conditional expectation with respect to a discrete σ-algebra

While the general notion of conditional expectation requires some abstraction, we can introduce conditioning with respect to a discrete σ-algebra starting from the elementary notion of conditional probability with respect to an event with positive probability.

Definition 5.2. Let Ω be a universe, $\mathcal{F}$ a σ-algebra of events on Ω, P a probability distribution on $(\Omega, \mathcal{F})$, let $(A_i)_{i\in\mathcal{J}\subseteq\mathbb{N}}$ be pairwise disjoint events, with non-zero probability such that $\cup_i A_i = \Omega$. Let $\mathcal{G}$ be the atomic σ-algebra generated by $(A_i)_{i\in\mathcal{J}}$.

Let X be a random variable from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{H})$, the *conditional expectation of X with respect to* $\mathcal{G}$ is the random variable defined as

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{i\in\mathcal{J}} \mathbb{E}_{P_{\{\cdot|A_i\}}}[X] \times \mathbf{1}_{A_i}.$$

While $\mathbb{E}_{P_{\{\cdot|A_i\}}}[X]$ is a real number, $\mathbb{E}[X \mid \mathcal{G}]$ is a $\mathcal{G}$-measurable function from Ω to $\mathcal{X}$:

$$\mathbb{E}[X \mid \mathcal{G}](\omega) = \sum_{i\in\mathcal{J}} \mathbb{E}_{P_{\{\cdot|A_i\}}}[X] \times \mathbf{1}_{A_i}(\omega) \qquad \forall \omega \in \Omega\,.$$

These two kinds of objects should not be confused. We will refrain from using notation $\mathbb{E}[X \mid A_i]$ since it may be confusing: $\mathbb{E}[X \mid A_i]$ might denote either $\mathbb{E}_{P_{\{\cdot|A_i\}}}[X]$ or $\mathbb{E}[X \mid \sigma(A_i)]$ where $\sigma(A_i)$ is the sigma-algebra generated by A_i: $\{A_i, A_i^c, \Omega, \emptyset\}$.

Proposition 5.4. *Let P be a probability distribution on $(\Omega, \mathcal{F})$. Let $(A_i)_{i\in\mathcal{J}\subseteq\mathbb{N}}$ be a collection of pairwise disjoint events, with non-zero probability satisfying $\cup_{i\in\mathcal{J}} A_i = \Omega$. Let $\mathcal{G} = \sigma\Big((A_i)_{i\in\mathcal{J}}\Big)$ denote the sigma-algebra generated by $(A_i)_{i\in\mathcal{J}}$.The random variable X is assumed to be P- integrabe.*

1. *The conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ is a $\mathcal{G}$-measurable random variable, that satisfies*

$$\mathbb{E}\left[YX\right] = \mathbb{E}\left[Y\mathbb{E}[X \mid \mathcal{G}]\right] \qquad \forall Y \in \sigma(\mathcal{G}), Y \textit{ bounded.}$$

2. *If two $\mathcal{G}$-measurable random variables Z, T satisfy $\mathbb{E}\left[YX\right] = \mathbb{E}\left[YZ\right] = \mathbb{E}[YT]$, for all $Y \in \sigma(\mathcal{G}), Y$ bounded, then $Z = T$ almost surely.*

Proof. We need to ckeck points 1.) and 2.):

1. $\mathbb{E}\left[X \mid \mathcal{G}\right]$ satisfies first property in Proposition Proposition 5.4.

2. If Z satisfies Proposition 5.4, then $Z = \mathbb{E}[X \mid \mathcal{G}]$ P-almost-surely.

Checking i.)
If Y is $\mathcal{G}$-measurable, then $Y = \sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}$ for some real-valued sequence $(\lambda_i)_{i\in\mathcal{I}}$.
Then

$$\begin{aligned}
\mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]] \quad &= \mathbb{E}\left[\left(\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}\right)\left(\sum_{j\in\mathcal{I}} \mathbf{1}_{A_j} \frac{\mathbb{E}[\mathbf{1}_{A_j}X]}{P\{A_j\}}\right)\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i} \frac{\mathbb{E}[\mathbf{1}_{A_i}X]}{P\{A_i\}}\bigg)\right] \\
&= \sum_{i\in\mathcal{I}} \lambda_i \mathbb{E}[\mathbf{1}_{A_i}X] \frac{\mathbb{E}\left[\mathbf{1}_{A_i}\right]}{P\{A_i\}} \quad \text{linearity of expectation} \\
&= \sum_{i\in\mathcal{I}} \lambda_i \mathbb{E}[\mathbf{1}_{A_i}X] \\
&= \mathbb{E}\left[\left(\sum_{i\in\mathcal{I}} \lambda_i \mathbf{1}_{A_i}\right) X\right] \\
&= \mathbb{E}[YX] .
\end{aligned}$$

Checking ii.)
Assume Z satisfies Proposition 5.4.
Let us define Y using $Y = \mathbf{1}_{A_i}$, for some index $i \in \mathcal{I}$.
As Z is $\mathcal{G}$-measurable, there exists a real-valued sequence $(\mu_j)_{j\in\mathcal{I}}$, such that $Z = \sum_{j\in\mathcal{I}} \mu_j \mathbf{1}_{A_j}$.
Thus, relying on the fact that events A_j are pairwise disjoint:

$$\mathbb{E}[ZY] \quad = \mathbb{E}\left[\sum_{j\in\mathcal{I}} \mu_j \mathbf{1}_{A_j} \mathbf{1}_{A_i}\right] = \mu_i P\{A_i\}$$

By the defining property of Z, we have

$$\mathbb{E}[ZY] = \mathbb{E}[XY] = \mathbb{E}[X\mathbf{1}_{A_i}] .$$

Finally, for all $i \in \mathcal{I}$, $\mu_i = \mathbb{E}[X\mathbf{1}_{A_i}]/P\{A_i\}$.
We can now conclude $Z = \mathbb{E}[X \mid \mathcal{G}]$.
□ □

5.5 Conditional expectation as prediction

The next proposition reveals the role of conditional expectation in prediction/approximation problems.

Proposition 5.5. *Let Y be a square-integrable random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a discrete sub-σ-algebra of $\mathcal{F}$. The conditional expectation of Y with respect to $\mathcal{G}$ minimizes*

$$\mathbb{E}\left[(Y - Z)^2\right]$$

amongst $\mathcal{G}$-measurable square-integrable random variables.

Recall that a $\mathcal{G}$-measurable random variable is a function that remains constant on each $A_i, i \in \mathcal{I}$.

Proof. If Y is a random variable on $(\Omega, \mathcal{F})$, and if we are trying to predict at best Y from a $\mathcal{G}$-measurable random variable , we are looking for a sequence of coefficients $(b_i)_{i\in\mathcal{I}}$ that minimizes:

$$\begin{aligned}\mathbb{E}_P\left[\left(Y-\sum_{i\in\mathcal{I}} b_i\mathbf{I}_{A_i}\right)^2\right] & = \mathbb{E}_P\left[\left(\sum_{i\in\mathcal{I}}(Y-b_i)\mathbf{I}_{A_i}\right)^2\right] \\ & = \sum_{i\in\mathcal{I}}\mathbb{E}_P\left[(Y-b_i)^2\,\mathbf{I}_{A_i}\right] \\ & = \sum_{i\in\mathcal{I}} P\{A_i\}\,\mathbb{E}_{P\{\cdot\mid A_i\}}\left[(Y-b_i)^2\right]\end{aligned}$$

Thus for each i, b_i must coincide with the expectation of Y under $P\{\cdot \mid A_i\}$. The best prediction of Y, in the sense of the quadratic error, among the $\mathcal{G}$-measurable functions is the conditional expectation of Y with respect to $\mathcal{G}$.

□ □

The properties identified by propositions Proposition 5.4 and @ref(prp:espercondpred) serve as a definition for the conditional expectation with respect to a general σ-algebra.

5.6 Properties of conditional expectation

We state without proof a number of useful properties of conditional expectation with respect to discrete σ-algebras. We shall prove them in full generality later.

Proposition 5.6. *If $X \leq Y$, P-a.s., then*

$$\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}] \qquad P\text{-}p.s.$$

Proposition 5.7.

$$\mathbb{E}[aX+bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]\,.$$

Proposition 5.8. *If $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right] \mid \mathcal{H}\right] = \mathbb{E}\left[X \mid \mathcal{H}\right]\,.$$

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\right] = \mathbb{E}\left[X \mid \mathcal{H}\right]\,.$$

Prove the proposition.

$$\mathbb{E}X = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]]\,.$$

5.7 Application: Galton-Watson processes I

The size of generation $k \geq 0$ is defined recursively by

$$Z_0 = 1, \qquad Z_{k+1} = \sum_{i=1}^{Z_k} X_i^k\,.$$

The σ-algebra $\sigma(Z_k)$ is discrete/atomic, it is generated by the pairwise disjoint events $\left\{Z_k = a\right\}$ for $a \in \mathbb{N}$.

Proposition 5.9. *In a Galton-Watson (homogeneous) branching process with reproduction number μ, the conditional expectation of the size of the size of the $k+1^{th}$ generation with respect to the size of the k^{th} generation is a linear function of the size of the k^{th} generation:*

$$\mathbb{E}\Big[Z_{k+1} \mid \sigma(Z_k)\Big] = \mathbb{E}X_1^0 \times Z_k = \mu \times Z_k$$

Proof. On the event $\Big\{Z_k = a\Big\}$, we can determine the conditional distribution of Z_{k+1}.

$$\begin{aligned}\{Z_{k+1} = b \wedge Z_k = a\Big\} &= \Big\{\textstyle\sum_{i=1}^a X_i^k = b \wedge Z_k = a\Big\} \\ &= \Big\{\textstyle\sum_{i=1}^a X_i^k = b\Big\} \cap \Big\{Z_k = a\Big\}\end{aligned}$$

we have

$$P\Big\{Z_{k+1} = b \mid Z_k = a\Big\} = P\Big\{\textstyle\sum_{i=1}^a X_i^k = b \mid Z_k = a\Big\} = P\Big\{\textstyle\sum_{i=1}^a X_i^k = b\Big\}$$

On the event $\{Z_k = a\}$, Z_{k+1} is distributed like the sum of a independent copies of X_1^0:

$$\begin{aligned}\mathbb{E}\Big[Z_{k+1} \mid \sigma(Z_k)\Big] &= \textstyle\sum_{a=0}^\infty \mathbb{E}_{P(|Z_k=a)}\Big[Z_{k+1}\Big] \times \mathbb{I}_{Z_k=a} \\ &= \textstyle\sum_{a=0}^\infty \mathbb{E}_{P(|Z_k=a)}\Big[\textstyle\sum_{i=1}^a X_i^k\Big] \times \mathbb{I}_{Z_k=a} \\ &= \textstyle\sum_{a=0}^\infty \mathbb{E}\Big[\textstyle\sum_{i=1}^a X_i^k\Big] \times \mathbb{I}_{Z_k=a} \\ &= \textstyle\sum_{a=0}^\infty \textstyle\sum_{i=1}^a \mathbb{E}\Big[X_i^k\Big] \times \mathbb{I}_{Z_k=a} \\ &= \textstyle\sum_{a=0}^\infty a\mathbb{E}X_1^0 \times \mathbb{I}_{Z_k=a} \\ &= \mathbb{E}X_1^0 \times Z_k\,.\end{aligned}$$

□

An immediate corollary is:

$$\mathbb{E}Z_k = (\mathbb{E}X_1^0)^k \qquad \text{forall } k \geq 0\,.$$

The sequence of expected sizes of generations forms a geometric sequence.

A Galton-Watson process is said to be ***sub-critical*** if the expectation of the offspring distribution is smaller than 1.

Proposition 5.10 (Extinction under sub-critical offspring distribution)**.** *The extinction probability of a sub-critical branching process is equal to* 1.

Proof. Denote by E_k the event $\{Z_k = 0\}$. Observe that the sequence $(E_k)_k$ is increasing. Denote by $E_\infty = \cup_{k=0}^\infty E_k$.

$$P\{E_k^c\} = P\{Z_k \geq 1\} \leq \mathbb{E}Z_k\,.$$

Hence $P\{E_k^c\} \downarrow 0$ and $P\{E_k\} \uparrow 1$. By monotone convergence $P(E_\infty) = 1$. □

The expected size of the total progeny of subcritical branching process is equal to

$$\sum_{k=0}^\infty \mathbb{E}Z_k = \sum_{k=0}^\infty (\mathbb{E}X_1^0)^k = \frac{1}{1 - \mathbb{E}X_1^0}\,.$$

Working with discrete conditioning allows us to derive non-trivial statements about the Galton-Watson process without knowing much about the offspring distribution beyond the fact that its expectation is smaller than 1. We still ignore the the details of the distribution of Z_k, let alone of the distribution of $\sum_{k=0}^\infty Z_k$.

Chapter 6

Absolutely continuous probability measures

6.1 Densities and absolute continuity

Beyond discrete distributions, the simplest probability distributions are defined by a density function with respect to a (σ-finite) measure. This encompasses the distributions of the so-called ***continuous random variables***.

Definition 6.1 (Absolute continuity). Let μ, ν be two positive measures on measurable space $(\Omega, \mathcal{F})$, μ is said to be absolutely continuous with respect to ν (denoted by $\mu \trianglelefteq \nu$) iff for every $A \in \mathcal{F}$ with $\nu(A) = 0$, we also have $\mu(A) = 0$.

If μ, ν are two probability distributions, and $\mu \trianglelefteq \nu$, then any event which is impossible under ν is also impossible under μ.

Exercise 6.1. Answer the two questions:

- Is the counting measure on $\mathbb{R}$ absolutely continuous with respect to Lebesgue measure?
- Is the converse true?

Exercise 6.2. Check that absolute continuity is a transitive relationship.

The next theorem has far-reaching practical consequences.

Theorem 6.1 (Radon-Nikodym). *Let μ, ν be two positive measures on measurable space $(\Omega, \mathcal{F})$. Assume ν is σ-finite. If $\mu \trianglelefteq \nu$, then there exists a measurable function f from Ω to $[0, \infty)$ such that for all $A \in \mathcal{F}$,*

$$\mu(A) = \int_A f(\omega) \mathrm{d}\nu(\omega) = \int \mathbb{I}_A f \mathrm{d}\nu .$$

The function f is called a version of the density of μ with respect to ν.

The density is also called the Radon-Nikodym derivative of μ with respect to ν. It is sometimes denoted by $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

Remark 6.1. The σ-finiteness assumption is crucial. If we choose μ as Lebesgue measure and ν as the counting measure, ν is not σ-finite, $\mu(A) > 0$ implies $\nu(A) = \infty$ which we may consider as larger than 0. Nevertheless, Lebesgue measure has no density with respect to the counting measure.

In the next sections, we investigate probability distributions over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that are absolutely continuous with respect to Lebesgue measure.

Proposition 6.1. *If $\rho \trianglelefteq \mu \trianglelefteq \nu$, f is a density of ρ with repsect to μ while g is a density of μ with respect to ν, then fg is a density of ρ with respect to ν.*

Exercise 6.3. Prove proposition 6.1.

6.2 Exponential distribution

The exponential distribution shows up in several areas of probability and statistics. In reliability theory, its memoryless property make it a borderline case. In the theory of point processes, the exponential distribution is connected with Poisson Point Processes. It is also important in extreme value theory.

The exponential distribution with intensity parameter $\lambda > 0$ is defined by its desnsity with respect to Lebesgue measure on $[0, \infty)$:

$$x \mapsto \lambda \mathrm{e}^{-\lambda x} .$$

The reciprocal of the intensity parameter is called the scale parameter.

Note that geometric and exponential distributions are connected: if X is exponentially distributed, then $\lceil X \rceil$ is geometrically distributed. For $k \geq 1$:

$$P\Big\{\lceil X \rceil \geq k\Big\} = P\Big\{X > k-1\Big\} = \mathrm{e}^{-\lambda(k-1)} .$$

Exercise 6.4. Check that $x \mapsto \lambda \mathrm{e}^{-\lambda x}$ is a density probability over $[0, \infty)$.

Exercise 6.5. Compute the tail function and the cumulative distribution function of the exponential distribution function with parameter λ.

Exercise 6.6. Let $X_1, \ldots, X_n$ be i.i.d. exponentially distributed. Characterize the distribution of $\min(X_1, \ldots, X_n)$.

If X is exponentially distributed with scale parameter σ, what is the distribution of aX?

6.3 Gamma distribution

Sums of independent exponentially distributed random variables are not exponentially distributed. The family of Gamma distributions encompasses the family of exponential distributions. It is stable under addition and satisfies

Recall Euler's Gamma function:

$$\Gamma(t) = \int_0^\infty x^{t-1} \mathrm{e}^{-x} \mathrm{d}x \qquad \text{for } t > 0 .$$

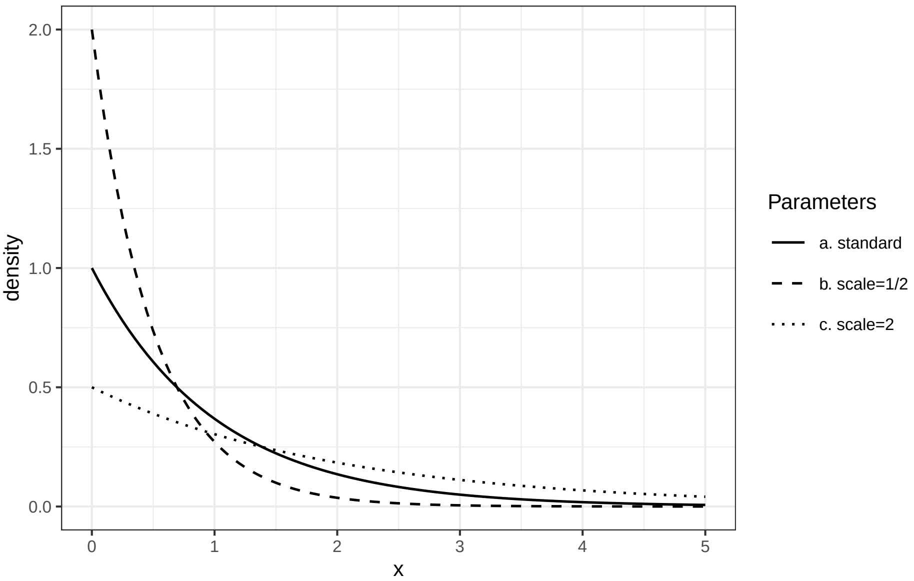

Figure 6.1: Exponential densities with different parameters: scales $1, 2, 1/2$ or equivalently intensities $1, 1/2, 2$. Expectation equals scale,
variance equals squared scale.

The Gamma distribution with shape parameter $p > 0$ and intensity parameter $\lambda > 0$ is defined by its density with respect to Lebesgue measure on $[0, \infty)$:

$$x \mapsto \lambda^p \frac{x^{p-1}}{\Gamma(p)} \mathrm{e}^{-\lambda x} \, .$$

The reciprocal of the intensity parameter is called the scale parameter.

Exercise 6.7. Check that $x \mapsto \lambda^p \frac{x^{p-1}}{\Gamma(p)} \mathrm{e}^{-\lambda x}$ is a density probability over $[0, \infty)$.

Exercise 6.8. If X is Gamma distributed with shape parameter p and scale parameter σ, what is the distribution of aX?

6.4 Univariate Gaussian distributions

Gaussian distributions play a central role in Probability theory, Statistics, Information theory, and Analysis.

The Gaussian or normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2, \sigma > 0$ has density

$$x \mapsto \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } x \in \mathbb{R} \, .$$

The standard Gaussian density is defined by $\mu = 0, \sigma = 1$.

Exercise 6.9. Check that $x \mapsto \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}$ is a probability density over $\mathbb{R}$.

Exercise 6.10. If X is distributed according to a standard Gaussian density, what is the distribution of $\mu + \sigma X$?

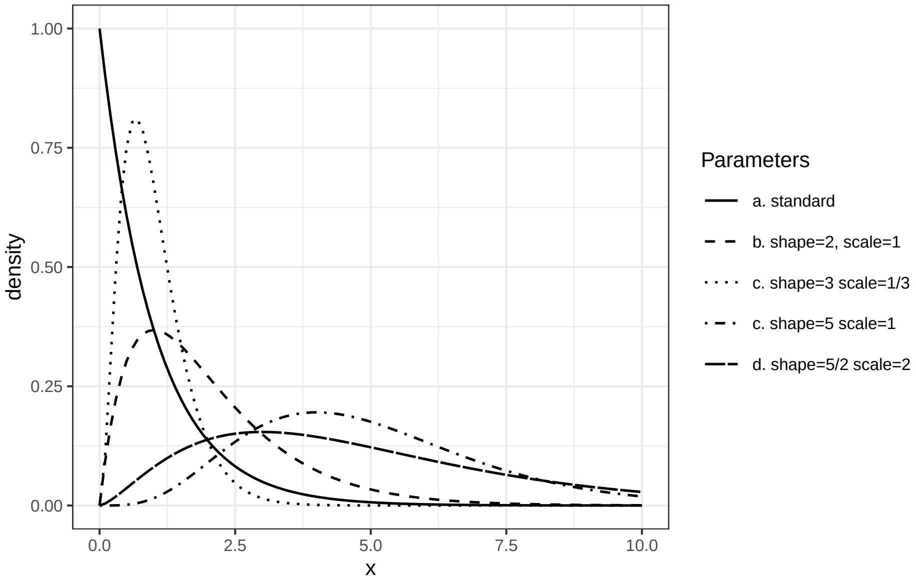

Figure 6.2: Gamma densities with different parameters: scales $1, 1, 1/3, 1, 2$ and shapes $1, 2, 3, 5, 5/2$. Expectation equals shape times scale,
variance equals shape times squared scale.

Exercise 6.11. If X is distributed according to a standard Gaussian density, show that

$$\Pr\{X > t\} \leq \frac{1}{t}\frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}} \qquad \text{for } t > 0\,.$$

6.5 Cumulative distribution functions and absolute continuity

If a cumulative distribution function is defined as the integral of some non-negative Lebesgue integrable function, we know that the corresponding probability distribution is absolutely continuous with respect to Lebesgue measure.

We may ask for a criterion that characterises the cumulative distribution function of absolutely continuous probability distribution. Such a criterion is embodied by the next definition. We overload the expression ***absolutely continuous***.

Definition 6.2 (Absolutely continuous functions). A real valued function F on $[a, b]$ is said to be ***absolutely continuous*** iff for all $\delta > 0$ there exists $\epsilon > 0$ such that for every collection $([a_i, b_i])_{i\leq n}$ for non-overlapping sub-intervals ($[a_i, b_i] \subseteq [a, b]$ for all $i \leq n$ and $\ell([a_i, b_i] \cap [a_j, b_j]) = 0$ for $i \neq j$) with $\sum_{i\leq n} |b_i - a_i| \leq \epsilon$,

$$\sum_{i\leq n} |F(b_i) - F(a_i)| \leq \delta$$

Absolute continuity, differentiability and integration of derivatives are connected by the next Theorem. This Theorem tells us that a cumulative distribution function is absolutely continuous in the sense of Definition Definition 6.2 iff the corresponding probability distribution is absolutely continuous with respect to Lebesgue measure.

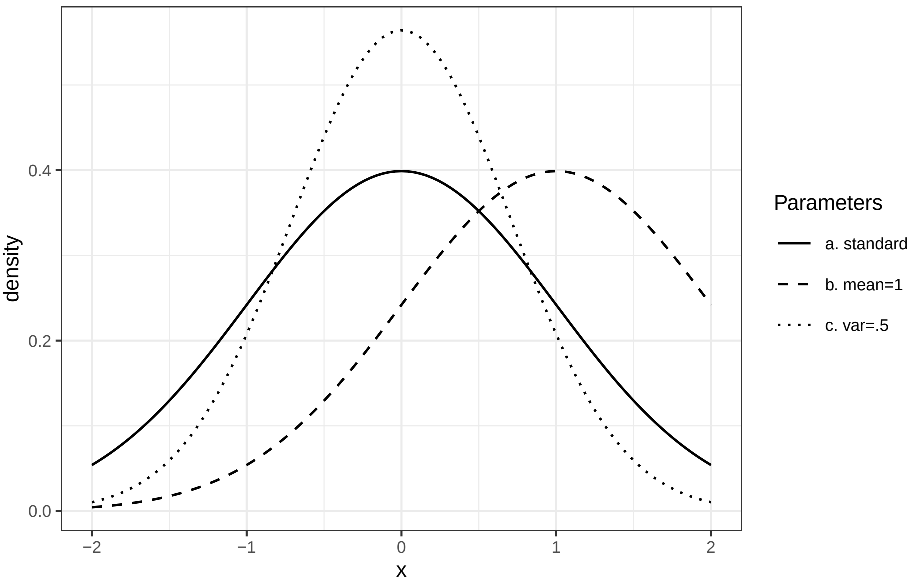

Figure 6.3: Gaussian densities. The location parameter μ coincides with the mean and the median. The scale parameter is the standard deviation. The Inter-Quartile-Range (IQR) is proportional to the standard deviation. If $\Phi^{\leftarrow}$ denotes the quantile function of $\mathcal{N}(0,1)$ then the interquartile range of $\mathcal{N}(\mu,\sigma^2)$ is $\sigma\Big(\Phi^{\leftarrow}(3/4) - \Phi^{\leftarrow}(1/4)\Big) = 2\sigma\Phi^{\leftarrow}(3/4)$.

Theorem 6.2 (Fundamental Theorem of Calculus). ***A real valued function** F **on** $[a,b]$ **is*** absolutely continuous ***iff the next three conditions hold***

1. ***The derivative*** F' ***exists Lebesgue almost everywhere on*** $[a,b]$
2. ***The derivative*** F' ***is Lebesgue integrable***
3. ***For every*** $x \in [a,b]$, $F(x) - F(a) = \int_{[a,b]} \mathbb{I}_{[a,x]}(t)F'(t)\mathrm{d}t$

6.6 Computing the density of an image probability distribution

Recall the change of variable formula in elementary calculus. If ϕ is monotone increasing and differentiable from open A to B and f is Riemann integrable over B, then

$$\int_B f(y)\,\mathrm{d}y = \int_A f(\phi(x))\,\phi'(x)\,\mathrm{d}x$$

Exercise 6.12. Check the elementary change of variable formula.

The goal of this section is state a multi-dimensional generalization of this elementary formula. This is the content of Theorem 6.4). This extension is then used to establish an off-the-shelf formula for computing the density of an image distribution in Theorem 6.5).

Let us start with a uni-dimensional warm-up. When starting from the uniform distribution on $[0,1]$ and applying a monotone differentiable transformation, the density of the image measure is easily computed.

Exercise 6.13. Let ϕ be differentiable and increasing on $[0,1]$, and let P be the uniform distribution on $[0,1]$.

Check that $P \circ \phi^{-1}$ has density $\frac{1}{\phi' \circ \phi^{\leftarrow}}$ on $\phi([0,1])$.

The next proposition extends this observation.

If the real valued random variable X is distributed according to P with density f, and ϕ is monotone increasing and differentiable over $\operatorname{supp}(P)$, then the probability distribution of $Y = \phi(X)$ has density

$$g = \frac{f \circ \phi^{\leftarrow}}{\phi' \circ \phi^{\leftarrow}}$$

over $\phi(\operatorname{supp}(P))$.

Proof. By the fundamental theorem of calculus, the density f is a.e. the derivative of the cumulative distribution function F of P.

The cumulative distribution function of $Y = \phi(X)$ satisfies:

$$\begin{aligned} P\Big\{Y \leq y\Big\} &= P\Big\{\phi(X) \leq y\Big\} \\ &= P\Big\{X \leq \phi^{\leftarrow}(y)\Big\} \\ &= F \circ \phi^{\leftarrow}(y) \end{aligned}$$

Almost everywhere, $F \circ \phi^{\leftarrow}$ is differentiable, and has derivative $\frac{f\circ\phi^{\leftarrow}}{\phi'\circ\phi^{\leftarrow}}$ in $\phi(\operatorname{supp}(P))$, 0 elsewhere. and

$$P\Big\{Y \leq y\Big\} = \int_{(-\infty,y]\cap\phi(\operatorname{supp}(P))} \frac{f \circ \phi^{\leftarrow}(u)}{\phi' \circ \phi^{\leftarrow}(u)} \mathrm{d}u$$

□ □

The next corollary is as useful as simple.

Corollary 6.1. *If the distribution of the real valued random variable X has density f then the distribution of $\sigma X + \mu$ has density $\frac{1}{\sigma} f\Big(\frac{\cdot - \mu}{\sigma}\Big)$, .*

In univariate calculus, it is easy to establish that if a function is continuous and increasing over an open set, it is invertible and its inverse is continuous and increasing. If the function is differentiable with positive derivative, its inverse is also differentiable. Moreover, the differential and the differential of the inverse are related in transparent way.

The Global Inversion Theorem extends the preceding observation to the multivariate setting.

Theorem 6.3 (Global Inversion Theorem). *Let U and V be two non-empty open subsets of $\mathbb{R}^d$. Let ϕ be a continuous bijective from U to V. Assume furthermore that ϕ is continuously differentiable, and that $D\phi_x$ is non-singular at every $x \in U$.*

Then, the inverse function $\phi^{\leftarrow}$ is also continuously differentiable on V and at every $y \in V$:

$$D\phi^{\leftarrow}_y = \Big(D\phi_{\phi^{\leftarrow}(y)}\Big)^{-1} .$$

The Jacobian determinant of ϕ is the determinant of the matrix that represents the differential. It is denoted by J_ϕ. Recall that:

$$J_{\phi^{\leftarrow}}(y) = \Big(J_\phi(\phi^{\leftarrow}(y))\Big)^{-1} .$$

The multidimensional version of the change of variable formula is stated under the same assumptions as the Global Inversion Theorem. We admit the next Theorem.

Theorem 6.4 (Geometric change of variable formula). *Let U and V be two non-empty open subsets of $\mathbb{R}^d$. Let ϕ be a continuous bijective from U to V. Assume furthermore that ϕ is continuously differentiable, and that $D\phi_x$ is non-singular at every $x \in U$.*

Let ℓ denote the Lebesgue measure on $\mathbb{R}^d$.

For any a non-negative Borel-measurable function f:

$$\int_U f(x)\mathrm{d}\ell(x) = \int f(\phi^{\leftarrow}(y))\Big|J_{\phi^{\leftarrow}}(y)\Big|\mathrm{d}\ell(y)\,.$$

Moving from cartesian coordinates to polar/spherical coordinates is easy thanks to an non-trivial application of Theorem 6.4).

The Image density formula is a corollary of the geometric change of variable formula.

Theorem 6.5 (Image density formula). *Let P have density f over open $U \subseteq \mathbb{R}^d$.*

Let ϕ be bijective fron U to $\phi(U)$ and ϕ be continuously differentiable over U with non-singular differential.

The density g of the image distribution $P \circ \phi^{-1}$ over $\phi(U)$ is given by

$$g(y) = f(\phi^{\leftarrow}(y)) \times |J_{\phi^{\leftarrow}}(y)| = f(\phi^{\leftarrow}(y)) \times \Big|J_{\phi}(\phi^{\leftarrow}(y))\Big|^{-1}\,.$$

The proof of Theorem 6.5) from Theorem 6.4) is a routine application of the transfer formula.

Proof. Let B be a Borelian subset of $\phi(U)$. By the transfer formula:

$$\begin{aligned} P\Big\{Y \in B\Big\} &= P\Big\{\phi(X) \in B\Big\} \\ &= \int_U \mathbb{I}_B(\phi(x))f(x)\mathrm{d}\ell(x)\,. \end{aligned}$$

Now, we invoke Theorem 6.4):

$$\begin{aligned} \int_U \mathbb{I}_B(\phi(x))f(x)\mathrm{d}\ell(x) &= \int_{\phi(U)} \mathbb{I}_B(\phi(\phi^{\leftarrow}(y)))f(\phi^{\leftarrow}(y))\Big|J_{\phi^{\leftarrow}}(y)\Big|\mathrm{d}\ell(y) \\ &= \int_{\phi(U)} \mathbb{I}_B(y)f(\phi^{\leftarrow}(y))\Big|J_{\phi^{\leftarrow}}(y)\Big|\mathrm{d}\ell(y)\,. \end{aligned}$$

This suffices to conclude that $f \circ \phi^{\leftarrow}\Big|J_{\phi^{\leftarrow}}\Big|$ is a version of the density of $P \circ \phi^{-1}$ with respect to Lebesgue measure over $\phi(U)$.

□ □

6.7 Application: Gamma-Beta calculus

The image density formula is applied to show a remarkable connexion between Gamma and Beta distributions.

Proposition 6.2. *Let X, Y be independent random variables distributed according to $\Gamma(p, \lambda)$ and $\Gamma(q, \lambda)$ (the intensity parameter are identical). Let $U = X + Y$ and $V = X/(X + Y)$.*

The random variables U and V are independent. Random variable U is distributed according to $\Gamma(p + q, \lambda)$ while V is distributed according to $\operatorname{Beta}(p, q)$.

Proof. The mapping $f :]0,\infty)^2 \to]0,\infty)\times]0,1[$ defined by

$$f(x,y) = \Big(x+y, \frac{x}{x+y}\Big)$$

is one-to-one with inverse $f^{\leftarrow}(u,v) = \Big(uv, u(1-v)\Big)$. The Jacobian matrix of $f^{\leftarrow}$ at (u,v) is

$$\begin{pmatrix} v & u \\ (1-v) & -u \end{pmatrix}$$

with determinant $-uv - u + uv = -u$. The joint image density at $(u,v) \in]0,\infty)\times]0,1[$ is

$$\begin{aligned} &= \lambda^{p+q}\frac{(uv)^{p-1}}{\Gamma(p)}\frac{(u(1-v))^{q-1}}{\Gamma(q)}\mathrm{e}^{-\lambda(uv+u(1-v))}u \\ &= \Big(\lambda^{p+q}\frac{u^{p+q-1}}{\Gamma(p+q)}\mathrm{e}^{\lambda u}\Big) \times \Big(\frac{\Gamma(p+q)}{\Gamma(q)\Gamma(p)}v^{p-1}(1-v)^{q-1}\Big). \end{aligned}$$

The factorization of the joint density proves that the U and V are independent. We recognize that the density of (the distribution of) U is the Gamma density with shape parameter $p+q$, intensity parameter λ. The density of the distribution of V is the Beta density with parameters p and q. □

Exercise 6.14. Assume $X_1, X_2, \ldots, X_n$ form an independent family with each X_i distributed according to $\Gamma(p_i, \lambda)$.

Determine the joint distribution of

$$\sum_{i=1}^n X_i, \frac{X_1}{\sum_{i=1}^n X_i}, \frac{X_2}{\sum_{i=1}^n X_i}, \ldots, \frac{X_{n-1}}{\sum_{i=1}^n X_i}$$

6.8 Bibliographic remarks

Dudley (2002) and Pollard (2002) provide a full development of absolute continuity and self-contained proofs the Radon-Nikodym's Theorem.

Chapter 7

Characterizations of probability distributions

7.1 Motivation

In full generality, a probability distribution is a complex and opaque object. It is a $[0,1]$-valued function defined over a σ-algebra of subsets. A concrete σ-algebra, let alone the abstract notion of σ-algebra, is not easily grasped. Looking for simpler characterizations of probability distributions is a sensible goal. When facing questions like: "are two probability distributions equal¿', we know it suffices to check that the two distributions coincide on generating families of events. This makes Cumulative Distribution Functions (CDFs) precious tools. Cumulative Distribution Functions and their generalized inverse functions (quantile functions) are very convenient when handling maxima, minima, or more generally order statistics of collections of independent random variables, but when it comes to handling sums of independent random variables or branching processes, cumulative distribution functions are of moderate help.

In this lesson, we review three related ways of characterizing probability distributions through functions defined on the real line: Probability Generating Functions (Section 7.2)), Laplace transforms (Section 7.3)) and characteristic functions which extend Fourier transforms to probability distributions (Section 7.4)). The three methods are distinct in scope but they rely on the same idea and share common features.

Indeed, Probability Generating Functions can be seen as special case of Laplace transforms. The latter can be seen as special cases of Fourier transforms. All three methods do characterize probability distributions. They are equipped with inversion formulae.

The three methods provide us with a seamless treatment of sums of independent random variables.

All three methods relate the integrability of probability distributions and the smoothness of transforms.

In the next lessons (Chapter 11), we shall see that the three transforms characterize ***convergence in distribution***.

Probability generating functions, Laplace transforms and characteristic functions deliver an important analytical machinery to Probability Theory. From Analysis, we get off-the-shelf arguments to establish smoothness properties of transforms, and with little more work, we can construct the inversion formulae.

7.2 Probability generating function

In this section, X is an integer-valued random variable, with distribution P, cumulative distribution function F and probability mass function p. Recall that P is completely characterized by the much simpler objects F and p. Now, let Y be another integer-valued random variable living on the same probability space as X, independent from X, with distribution Q, distribution function G and probability mass function q. What can we tell about the distribution of $X + Y$? Is it easy to figure out its cumulative distribution function, its probability mass function?

The probability mass function of (the distribution of) $X + Y$ is the ***convolution*** of p and q

$$\begin{aligned} \mathbb{P}\{X+Y=n\} &= \textstyle\sum_{k=0}^{n} \mathbb{P}\{X+Y=n \wedge X=k\} \\ &= \textstyle\sum_{k=0}^{n} \mathbb{P}\{Y=n-k \wedge X=k\} \\ &= \textstyle\sum_{k=0}^{n} \mathbb{P}\{Y=n-k\} \times \mathbb{P}\{X=k\} \\ &= \textstyle\sum_{k=0}^{n} p(k) \times q(n-k) \\ &= p \star q(n)\,, \end{aligned}$$

where the third equality comes from independence of $\sigma(X)$ and $\sigma(Y)$.

Besides the probability mass function, another function characterizes probability distributions and delivers instantaneous information about the distribution of sums of independent integer-valued random variables and many other things.

Definition 7.1 (Probability Generating Function)**.** The probability generating function (PGF) of a probability distribution over $\mathbb{N}$, defined by its probability mass function (PMF) p is the function $G : [0,1] \to \mathbb{R}$ defined by:

$$G(s) = \sum_{n=0}^{\infty} p(n) s^n\,.$$

Example 7.1. The probability generating function of basic discrete distributions is easily computed. The results are useful and suggestive.

- Bernoulli distribution with parameter p:

$$1 - p + ps = 1 + p(s-1)$$

- Binomial distribution with parameters n and p:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k = (ps + 1 - p)^n = (1 + p(s-1))^n$$

- Poisson distribution with parameter μ:

$$\sum_{n=0}^{\infty} \mathrm{e}^{-\mu} \frac{\mu^n}{n!} s^n = \mathrm{e}^{\mu(s-1)}\,.$$

The next observation follows almost immediately from the definition of probability generating functions.

Proposition 7.1. *A probability generating function G satisfies the following conditions:*

▯ *G is non-negative over $[0,1]$;*
▯ *$G(0) = P\{0\}, \quad G(1) = 1$;*
▯ *G is non-decreasing over $[0,1]$;*
▯ *G is continuous and convex.*

Proof. Properties 1), 2) and 3) are obvious: G is a convex combination of non-negative, non-decreasing, continuous and convex functions.
□ □

Generatingfunctionology lies at the crossing between combinatorics, real analysis, complex analysis, and probability theory. Defining PGF as a *power series* brings within probability theory a collection of theorems that facilitate the identification of probability distributions or that connect integrability properties of the probability distribution with smoothness properties of the PGF.

Keep in mind that a generating function defines a function from the set of complex numbers $\mathbb{C}$ to $\mathbb{C}$:

$$G(z) = \sum_{n=0}^{\infty} p(n) z^n \qquad \text{for all } z \in \mathbb{C} \text{ such that the series converges}.$$

Characterizing the domain of a function defined in that way is crucial. The next proposition is at the core of Power Series theory.

Proposition 7.2. *The* radius of convergence *of the generating function G*

$$G(z) \; = \sum_{n \in \mathbb{N}} p(n) z^n, \qquad z \in \mathbb{C}$$

is the unique $R \in [0, \infty) \cup \{+\infty\}$ such that:

▯ *for every $z \in \mathbb{C}$ with $|z| > R$, the series $\sum_{n \in \mathbb{N}} p(n) z^n$ diverges.*
▯ *for every $z \in \mathbb{C}$ with $|z| < R$, the series $\sum_{n \in \mathbb{N}} p(n) z^n$ is absolutely convergent.*

The open disk $\{z : z \in \mathbb{C}, |z| < R\}$ is called the disk of convergence *of G. The circle $\{z : z \in \mathbb{C}, |z| = R\}$ is called the* circle of convergence *of G.*

The radius of convergence R of the probability generating function $G(z) = \sum_{n \in \mathbb{N}} p(n) z^n$ satisfies

$$\frac{1}{R} = \limsup_n (p(n))^{1/n}.$$

The last statement is called Hadamard's rule for determination of the radius of convergence:

The radius of convergence of a probability generating function is always at least 1.

Proof. Let R be the supremum of all real numbers r such that for every $z \in \mathbb{C}$ with $|z| < r$, the series $\sum_{n \in \mathbb{N}} p(n) z^n$ is absolutely convergent.

As $(p(n))_n$ defines an absolutely convergent series, for every z with $|z| \leq 1$, $\sum_n |p(n) z^n| \leq \sum_n p(n) = 1$. Hence $R \geq 1$.
□ □

Example 7.2. The radius of convergence contains qualitative information about tail behavior:

- For Poisson distributions, the radius of convergence is infinite. This reflects the fast decay of the tail probability of Poisson distributions.

- For geometric distributions, $p(n) = q(1-q)^{n-1}$, the radius of convergence is $1/(1-q)$.
- For power law distributions like $p(n) = n^{-r}/\zeta(r)$ with $r > 1$, the radius of convergence is exactly 1.

Just knowing the radius of convergence of a function defined by a Power Series expansion tells us about the smoothness properties of the function.

Theorem 7.1. *If G is defined as a power series $G(z) = \sum_{n\in\mathbb{N}} a_n z^n$ its (complex) derivative is $G'(z) = \sum_{n\in\mathbb{N}}(n+1)a_{n+1}z^n$. The derivative G' and G have the same radius of convergence.*

This general statement about power series entails a very useful corollary for probability generating functions.

Corollary 7.1 (Inversion formula). *Let f be the probability generating function associated with the probability mass function p. Then f is infinitely many times differentiable over* $[0, 1)$

$$f^{(n)}(s) = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} \times p(k)s^{k-n} ,$$

more specifically:

$$f^{(n)}(0) = n! \times p(n) .$$

A probability distribution over $\mathbb{N}$ is characterized by its probability generating function.

Proof. The property is true for $n = 0$.

Assume it holds for all integers up to n. For $s \in [0, 1)$ and $|h| < 1 - s - \delta$ where δ is a small positive number,

$$\frac{f^{(n)}(s+h)-f^{(n)}(s)}{h} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} \times p(k)\Big(\sum_{j=0}^{k-n-1}(s+h)^{k-n-1-j}s^j\Big)$$

The absolute value of the internal sum is smaller than $(k-n)(1-\delta)^{k-n-1}$. As

$$\sum_{k=n}^{\infty} \frac{k!}{(k-n-1)!} \times p(k) \times (1-\delta)^{k-n-1} < \infty$$

for all $0 < \delta < 1$. By the Dominated Convergence Theorem,

$$\lim_{h\to 0} \frac{f^{(n)}(s+h)-f^{(n)}(s)}{h} = \sum_{k=n+1}^{\infty} \frac{k!}{(k-n-1)!} \times p(k) \times s^{k-n-1} .$$

□ □

The Probability Generating Function of a Poisson distribution with parameter μ equals $\exp(\mu(s-1))$. If we meet a probability distribution with such a PGF, we know it is a Poisson distribution.

Probability Generating Functions allow for easy investigations of sums of independent random variables.

Proposition 7.3. *Let X, Y be independent integer-valued random variable, with probability generating functions G_X and G_Y. The probability generating function G_{X+Y} of $X+Y$ is $G_X \times G_Y$:*

$$G_{X+Y} = G_X \times G_Y .$$

Proof. The proof relies on the fact that non-negative convergent series is commutatively convergent.

$$\begin{aligned}\sum_{n=0}^{\infty}\mathbb{P}\{X+Y=n\}\times s^n &= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}p(k)q(n-k)\right)s^n\\ &= \sum_{k=0}^{\infty}p(k)s^k\sum_{n\geq k}^{\infty}q(n-k)s^{n-k}\\ &= G_X(s)\times G_Y(s)\end{aligned}$$

In measure theoretical language, the proposition is a consequence of the Tonelli-Fubini Theorem:

$$\begin{aligned}G_{X+Y}(s) &= \mathbb{E}\left[s^{X+Y}\right]\\ &= \mathbb{E}\left[s^X\times s^Y\right]\\ &= \int_{\mathbb{R}^2}s^xs^y\mathrm{d}P_X\otimes P_Y(x,y)\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}s^xs^y\mathrm{d}P_X(x)\mathrm{d}P_Y(y)\\ &= \int_{\mathbb{R}}s^y\int_{\mathbb{R}}s^x\mathrm{d}P_X(x)\mathrm{d}P_Y(y)\\ &= \int_{\mathbb{R}}s^yG_X(s)\mathrm{d}P_Y(y)\\ &= G_X(s)\times G_Y(s)\,.\end{aligned}$$

□

Example 7.3. If X and Y are independent Poisson random variables with parameters μ and ν, then $G_{X+Y}(s)=\exp(\mu(s-1))\times\exp(\nu(s-1))=\exp((\mu+\nu)(s-1))$. This is (another) proof that $X+Y$ is Poisson distributed with parameter $\mu+\nu$.

A PGF is infinitely many times differentiable inside the (open) disk of convergence. If the radius of convergence is larger than 1 (as for Poisson distributions), this entails that the PGF is infinitely many times differentiable at 1, If the radius of convergence is exactly 1, the differentiability on the circle of convergence is not prescribed by general theory.

Theorem 7.2 (Integrability and probability generating functions). *Let X be an integer-valued random variable, with probability generating functions f, then*
$\mathbb{E}X^p<\infty$
iff
f is p-times differentiable at 1 and

$$f^{(p)}(1)=\mathbb{E}\left[X(X-1)\dots(X-p+1)\right]\,.$$

Proof. Assume that G is p-times differentiable on the left at 1.

We need to establish that $|X|$ is p-integrable.

Assume that $|X|$ is p-integrable. □

The next question arises quickly: when is a function from $[0,1]$ to $[0,\infty)$ a probability generating function? This question is addressed in a broader perspective in the next section.

7.3 Laplace transform

Laplace transforms characterize probability distributions on $[0,\infty)$.

Definition and elementary properties

Definition 7.2. Let P be a probability distribution function over $[0, \infty]$ with cumulative distribution function F. The Laplace transform of P is the function U from $[0, \infty)$ to $[0, 1]$ defined by

$$U(\lambda) = \mathbb{E}\left[\mathrm{e}^{-\lambda X}\right] = \int_{[0,\infty)} \mathrm{e}^{-\lambda x} \mathrm{d}F(x)$$

where $X \sim P$.

A probability distribution P over $\mathbb{N}$ is also a probability distribution over $[0, \infty)$, as such it has both a probability generating function G and a Laplace transform U. They are connected by

$$U(\lambda) = G(\mathrm{e}^{-\lambda}) .$$

Which properties of Probability Generating Functions are also satisfied by Laplace transforms?

Proposition 7.4. *If $U : [0, \infty) \to [0, 1]$ is the Laplace transform of a probability distribution P over $[0, \infty)$, then*

- $U(0) = 1$*;*
- *U is continuous;*
- *U is non-increasing.*
- *U is convex.*

Exercise 7.1. Check the assertions in the proposition.

Can we recognize Laplace transform of probability distributions over $[0, \infty)$? This is the content of the next Theorem (which proof is beyond the reach of this course).

Theorem 7.3 (Bernstein's Theorem)**.** *A function $U : (0, \infty) \to (0, \infty)$ is the Laplace transform of a probability distribution over $[0, \infty)$ iff*

- *U is infinitely many times differentiable over $(0, \infty)$*
- $U(0) = 1$
- *U is* completely monotonous*: $(-1)^k U^{(k)} \geq 0$ over $(0, \infty)$*

Using the connexion between Probability Generating Functions and Laplace transforms, we are in position to characterize those power series that are Probability Generating Functions.

Corollary 7.2. *A function $G : [0, 1] \to [0, 1]$ is the Probability Generating Function of a probability distribution over $\mathbb{N}$ iff*

- *G is infinitely many times differentiable over $(0, 1)$*
- $G(1) = 1$
- *G is completely monotonous: $(-1)^k G^{(k)} \geq 0$ over $(0, 1)$*

Example 7.4. Let X be Gamma(p, ν)-distributed. The Laplace transform of (the distribution of) X is

$$\begin{aligned} U(\lambda) \quad &= \int_0^\infty \nu \mathrm{e}^{-\lambda x} \mathrm{e}^{-\nu x} \tfrac{(\nu x)^{p-1}}{\Gamma(p)} \mathrm{d}x \\ &= \tfrac{\nu^p}{(\lambda+\nu)^p} \int_0^\infty (\lambda + \nu) \mathrm{e}^{-(\lambda+\nu)x} \tfrac{((\nu+\lambda)x)^{p-1}}{\Gamma(p)} \mathrm{d}x \\ &= \tfrac{\nu^p}{(\lambda+\nu)^p} \, . \end{aligned}$$

Injectivity of Laplace transforms and an inversion formula

Theorem 7.4 (Widder's Theorem)**.** *A probability distribution on $[0,\infty)$ is characterized by its Laplace transform.*

The construction of the inversion formula relies on deviation inequalities for Poisson distribution. The next proposition is easily checked by using Markov's inequality with exponential functions and optimization.

Theorem 7.5 (Tail bounds for Poisson distribution)**.** *Let Z be Poisson distributed. Let $h(x) = \mathrm{e}^x - x - 1$ and $h^*(x) = (x+1)\log(x+1) - x, x \geq -1$ be its convex dual. Then for all $\lambda \in \mathbb{R}$*

$$\log \mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)} = \mathbb{E}Z h(\lambda)\,.$$

For $t \geq 0$

$$\Pr\Big\{Z \geq \mathbb{E}Z + t\Big\} \leq \mathrm{e}^{-\mathbb{E}Z h^*\left(\frac{t}{\mathbb{E}Z}\right)}$$

and for $0 \leq t \leq \mathbb{E}Z$

$$\Pr\Big\{Z \leq \mathbb{E}Z - t\Big\} \leq \mathrm{e}^{-\mathbb{E}Z h^*\left(\frac{-t}{\mathbb{E}Z}\right)}\,.$$

Remark 7.1.

- See Section 3.7) for the notion of convex duality.
- The next bounds on h^* deliver looser but easier tail bounds

$$\begin{array}{lll} h^*(t) & \geq \frac{t^2}{2(1+t/3)} & \text{for } t > 0 \\ h^*(t) & \geq \frac{t^2}{2} & \text{for } t < 0\,. \end{array}$$

Corollary 7.3. *For all positive $x, y, y \neq x$,*

$$\lim_{n\to\infty} \sum_{k=0}^{nx} e^{-ny} \frac{(ny)^k}{k!} = \mathbb{I}_{y<x}\,.$$

We shall check in one of the next lessons that for $x > 0$:

$$\lim_{n\to\infty} \sum_{k=0}^{\lfloor nx \rfloor} e^{-nx} \frac{(nx)^k}{k!} = \frac{1}{2}\,.$$

Proof. Let F be the cumulative distribution function of P and U its Laplace transform. Let $X \sim P$.

It suffices to show that $F(x)$ can be computed from U at any x where F is continuous. Function U is infinitely many times differentiable on $(0,\infty)$. For $k \in \mathbb{N}$,

$$\frac{\mathrm{d}^k U}{\mathrm{d}\lambda^k} = (-1)^k \int_{[0,\infty)} x^k e^{-\lambda x} \mathrm{d}F(x)\,.$$

and U has a power series expansion at every $\lambda \in (0,1)$, for $\lambda' \in (0,1)$:

$$U(\lambda') \quad = \sum_{k=0}^{\infty} \frac{(\lambda'-\lambda)^k}{k!} \frac{\mathrm{d}^k U}{\mathrm{d}\lambda^k}\,.$$

By [Corollary 7.3), for all $0 < y \neq x$, $\lim_{n\to\infty} \sum_{k=0}^{nx} e^{-ny} \frac{(ny)^k}{k!} = \mathbb{I}_{y<x}$.

$$
\begin{aligned}
F(x) \quad & = \int_{\mathbb{R}_+} \mathbb{I}_{y\leq x} \mathrm{d}F(y) \\
& = \int_{\mathbb{R}_+} \mathbb{I}_{y<x} \mathrm{d}F(y) \\
& = \int_{(-\infty,x)} \mathbb{I}_{y<x} \mathrm{d}F(y) + \int_{\{x\}} 1 \mathrm{d}F(y) + \int_{(x,\infty)} \mathbb{I}_{y<x} \mathrm{d}F(y) \\
& = \int_{(-\infty,x)} \mathbb{I}_{y<x} \mathrm{d}F(y) + \int_{\{x\}} 1 \mathrm{d}F(y) + \int_{(x,\infty)} \mathbb{I}_{y<x} \mathrm{d}F(y) \\
& = \int_{(-\infty,x)\cup(x,\infty)} \lim_{n\to\infty} \sum_{k=0}^{nx} e^{-ny} \frac{(ny)^k}{k!} \mathrm{d}F(y) + \int_{\{x\}} 1 \mathrm{d}F(y) \\
& = \lim_{n\to\infty} \sum_{k=0}^{nx} \frac{(-n)^k}{k!} \int_{(-\infty,x)\cup(x,\infty)} e^{-ny} (-y)^k \mathrm{d}F(y) + \int_{\{x\}} 1 \mathrm{d}F(y) \\
& \text{by dominated convergence} \\
& = \lim_{n\to\infty} \sum_{k=0}^{nx} \frac{(-n)^k}{k!} \frac{\mathrm{d}^k U}{\mathrm{d}\lambda^k}_{|\lambda=n} .
\end{aligned}
$$

If F is continuous at x,

$$
F(x) = \lim_{n\to\infty} \sum_{k=0}^{nx} \frac{(-n)^k}{k!} \frac{\mathrm{d}^k U}{\mathrm{d}\lambda^k}_{|\lambda=n} .
$$

If F jumps at x,

$$
F(x) - \frac{P\{X = x\}}{2} = \lim_{n\to\infty} \sum_{k=0}^{nx} \frac{(-n)^k}{k!} \frac{\mathrm{d}^k U}{\mathrm{d}\lambda^k}_{|\lambda=n} .
$$

This process shows that the Laplace transform contains enough information to reconstruct the distribution function which in turn characterizes the probability distribution. □

Laplace transforms of sums of independent non-negative random variables are easily obtained from the Laplace transforms of the summands.

Proposition 7.5. *Let X and Y be two independent $[0, \infty)$-valued random variables, with Laplace transforms U_X and U_Y. The Laplace transform of (the distribution of) $X + Y$ is*

$$
G_{X+Y} = G_X \times G_Y .
$$

Proof.

$$
\begin{aligned}
G_{X+Y}(\lambda) \quad & = \mathbb{E}\left[\mathrm{e}^{\lambda(X+Y)}\right] \\
& = \mathbb{E}\left[\mathrm{e}^{\lambda X} \times \mathrm{e}^{\lambda Y}\right] \\
& = \mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \times \mathbb{E}\left[\mathrm{e}^{\lambda Y}\right] \\
& \text{independence} \\
& = G_X(\lambda) \times G_Y(\lambda) .
\end{aligned}
$$

□

Combining the inversion theorem and the explicit formula for the Laplace transform of Gamma distributions, we recover the fact that sums of independent Gamma-distributed random variables with the same intensity parameter is also Gamma distributed.

If $X \sim \text{Gamma}(p, \lambda)$ is independent from $Y \sim \text{Gamma}(q, \lambda)$ then $X + Y$ has Laplace transform $\left(\frac{\nu}{\lambda+\nu}\right)^{p+q}$ and is $\text{Gamma}(p + q, \lambda)$-distributed.

7.4 Characteristic functions and Fourier transforms

The Laplace transform characterizes probability distributions supported by $[0, \infty)$. Characteristic functions deal with general probability distributions. They extend to multivariate distributions.

Characteristic function

The next transform can be defined for all probability distributions over $\mathbb{R}$. And the definition can be extended to distributions on $\mathbb{R}^k, k \geq 1$.

Let the real-valued random variable X be distributed according to P with cumulative distribution function F, the characteristic function of distribution P is the function from $\mathbb{R}$ to $\mathbb{C}$ defined by

$$\widehat{F}(t) = \mathbb{E}\left[\mathrm{e}^{itX}\right] = \int_{\mathbb{R}} \mathrm{e}^{itx} \mathrm{d}F(x) = \int_{\mathbb{R}} \cos(tx) \mathrm{d}F(x) + i \int_{\mathbb{R}} \sin(tx) \mathrm{d}F(x)\,.$$

If F is absolutely continuous with density f then $\widehat{F}$ is (up to a multiplicative constant) the Fourier transform of f.

Let the real-valued random variable X be distributed according to P with characteristic function $\widehat{F}$.

- $\widehat{F}$ is (uniformly) continuous over $\mathbb{R}$
- $\widehat{F}(0) = 1$
- If X is symmetric, $\widehat{F}$ is real-valued
- The characteristic function of the distribution of $aX + b$ is

$$\mathrm{e}^{itb} \widehat{F}(at)\,.$$

Proof. Let us check the continuity property. The three others are left as exercises.

Trigonometric calculus leads to

$$\begin{aligned} \left|\mathrm{e}^{i(t+\delta)x} - \mathrm{e}^{itx}\right| & = \left|\mathrm{e}^{itx}\right| \times \left|\mathrm{e}^{i\delta x} - 1\right| \\ & \leq \left|\mathrm{e}^{i\delta x} - 1\right| \\ & \leq 2\Big(1 \wedge |\delta x|\Big) \end{aligned}$$

for every $t \in \mathbb{R}, \delta \in \mathbb{R}, x \in \mathbb{R}$. Taking integration with respect to F,

$$\left|\widehat{F}(t+\delta) - \widehat{F}(t)\right| \quad \leq \int 2\Big(1 \wedge |\delta x|\Big) \mathrm{d}F(x)\,.$$

Resorting to the dominated convergence theorem, we conclude

$$\lim_{\delta \to 0} \left|\widehat{F}(t+\delta) - \widehat{F}(t)\right| = 0$$

uniformly in t. □

The next properties are easily checked:

- $|\widehat{F}(t)| \leq 1$ for every $t \in \mathbb{R}$;
-

Compute the characteristic function of:

- The Poisson distribution with parameter $\lambda > 0$;
- The uniform distribution on $[-1, 1]$;
- The triangle distribution on $[-1, 1]$ (density: $1 - |x|$ on $[-1, 1]$);
- The exponential distribution with density $\exp(-x)$ on $[0, +\infty)$;
- The Laplace distribution, density $1/2 \exp(-|x|)$.

Just as Probability Generating Functions and Laplace transforms, Characteristic functions of sums of independent random variables have a simple structure.

Proposition 7.6. *Let X and Y be independent random variables with cumulative distribution functions F_X and F_Y, then*

$$\widehat{F}_{X+Y}(t) = \widehat{F}_X(t) \times \widehat{F}_Y(t)$$

for all $t \in \mathbb{R}$.

Proof. The third equality is a consequence of the independence of X and Y:

$$\begin{aligned}\widehat{F}_{X+Y}(t) &= \mathbb{E}\left[\mathrm{e}^{it(X+Y)}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{itX}\mathrm{e}^{itY}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{itX}\right] \times \mathbb{E}\left[\mathrm{e}^{itY}\right] \\ &= \widehat{F}_X(t) \times \widehat{F}_Y(t)\,.\end{aligned}$$

□

Use a counter-example to prove that

$$\left(\forall t \in \mathbb{R}, \quad \widehat{F}_{X+Y}(t) = \widehat{F}_X(t) \times \widehat{F}_Y(t)\right) \not\Rightarrow X \perp\!\!\!\perp Y.$$

Characteristic function of a univariate Gaussian distribution

It is possible to compute characteristic functions by resorting to Complex Analysis. But we shall refrain from this when computing the most important characteristic function, the characteristic function of the standard Gaussian distribution.

Proposition 7.7. *Let $\widehat{\Phi}$ denote the characteristic function of the standard univariate Gaussian distribution $\mathcal{N}(0, 1)$, the following holds*

$$\widehat{\Phi}(t) = \mathrm{e}^{-\frac{t^2}{2}}\,.$$

Proof. Recall that as the standard Gaussian density is even, the characteristic function is real-valued and even.

Moreover, $\widehat{\Phi}$ is differentiable and the derivative can be computing by interverting expectation and derivation with respect to t.

$$\begin{aligned}\widehat{\Phi}'(t) &= -\mathbb{E}\left[X \sin(tX)\right] \\ &= -\tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x \sin(tx)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x \\ &= \tfrac{1}{\sqrt{2\pi}}\left[\sin(tx)\mathrm{e}^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} - t\tfrac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \cos(tx)\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x \\ &= -t\widehat{\Phi}(t)\,.\end{aligned}$$

Hence, $\widehat{F}$ is a solution of the differential equation: $g'(t) = -tg(t)$ with $g(0) = 1$.

The differential equation is readily solved, and the solution is $g(t) = \mathrm{e}^{-\frac{t^2}{2}}$. □

Why is $\widehat{\Phi}$ differentiable? Why are we allowed to interchange expectation and derivation?

Note that a byproduct of Proposition @ref(prp:proCharFunGauss) is the following integral representation of the Gaussian density.

$$\phi(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{\Phi}(t) \mathrm{e}^{-itx} \mathrm{d}t\,.$$

It does not look interesting, but it is a milestone for the derivation of the general inversion formula below.

Sums of independent random variables and convolutions

The interplay between Characteristic functions/Fourier transforms and summation of independent random variables is one of the most attractive features of this transformation. In order to understand it, we shall need an operation stemming from analysis. Recall that if f and g are two integrable functions, the convolution of f and g is defined as

$$f \star g(x) = \int_{\mathbb{R}} f(x-y) g(y) \mathrm{d}y = \int_{\mathbb{R}} g(x-y) f(y) \mathrm{d}y\,.$$

Note that $f \star g$ is also integrable. It is not too hard to check that if f and g are two probability densities then so is $f \star g$, moreover $f \star g$ is the density of the distribution of $X+Y$ where $x \sim f$ is independent from $Y \sim g$. The next proposition extends this observation.

Proposition 7.8. *Let X, Y be two independent random variables with distributions P_X and P_Y. Assume that P_X is absolutely continuous with density p_X. Then the distribution of $X+Y$ is absolutely continuous and has density*

$$p_x \star P_Y(z) = \int_{\mathbb{R}} p_X(z-y) \mathrm{d}P_Y(y)\,.$$

Proof. Let B be Borel subset of $\mathbb{R}$.

$$\begin{aligned} P\Big\{X+Y \in B\Big\} & = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbb{I}_B(x+y) p_X(x) \mathrm{d}x \Big) \mathrm{d}P_Y(y) \\ & = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbb{I}_B(z) p_X(z-y) \mathrm{d}z \Big) \mathrm{d}P_Y(y) \\ & = \int_{\mathbb{R}} \mathbb{I}_B(z) \Big(\int_{\mathbb{R}} p_X(z-y) \mathrm{d}P_Y(y) \Big) \mathrm{d}z \\ & = \int_{\mathbb{R}} \mathbb{I}_B(z) p_x \star P_Y(z) \mathrm{d}z \end{aligned}$$

where the first equality follows from the Tonelli-Fubini Theorem, the second equality is obtained by change of variable $x \mapsto z = x+y$ for every y, the third equality follows again from the Tonelli-Fubini Theorem. □

Convolution is not tied to Probability theory.

- In Analysis, convolution is known to be a regularizing (smoothing) operation. This also holds in Probability theory: if the distribution of either X or Y has a density and $X \perp\!\!\!\perp Y$, then the distribution of $X+Y$ has a density.
- Convolution with smooth distributions plays an important role in non-parametric statsitics, it is at the root of kernel density estimation.

- Convolution is an important tool in Signal Processing.

Check that if X and Y are independent with densities f_X and f_Y, $f_X \star f_Y$ is a density of the distribution of $X + Y$.

If $Y = 0$ almost surely (its distribution is δ_0), then $p_X \star \delta_0 = p_X$.

What happens in Proposition @ref(prp:convsum), if we consider the distributions of $\sigma X + Y$ and let σ decrease to 0?

Proposition 7.9. *Let X, Y be two independent random variables with distributions P_X and P_Y. Assume that P_X is absolutely continuous with density p_X and that $P_X(-\infty, 0] = \alpha \in (0, 1)$. Then*

$$\lim_{\sigma \downarrow 0} \mathbb{P}\Big\{Y + \sigma X \le a\Big\} = P_Y(-\infty, a) + \alpha P_Y\{a\} \,.$$

Proof.

$$\begin{aligned}\mathbb{P}\Big\{Y + \sigma X \le a\Big\} \quad & = \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{I}_{x \le \frac{a-y}{\sigma}} p_X(x) \mathrm{d}x \mathrm{d}P_Y(y) \\ & = \int_{(-\infty,a)} \int_{\mathbb{R}} \mathbb{I}_{x \le \frac{a-y}{\sigma}} p_X(x) \mathrm{d}x \mathrm{d}P_Y(y) \\ & + \int_{\mathbb{R}} \mathbb{I}_{x \le \frac{a-a}{\sigma}} p_X(x) \mathrm{d}x P_Y\{a\} \\ & + \int_{(a,\infty)} \int_{\mathbb{R}} \mathbb{I}_{x \le \frac{a-y}{\sigma}} p_X(x) \mathrm{d}x \mathrm{d}P_Y(y)\end{aligned}$$

By monotone convergence, the first and third integrals converge respectively to $P_Y(-\infty, a)$ and 0 while the second term equals $\alpha P_Y\{a\}$. □

Injectivity Theorem and inversion formula

The characteristic function maps probability measures to $\mathbb{C}$-valued functions. The main result of this section is that characteristic functions/Fourier transforms define is an injective operator on the set of Probability measures on the real line.

Theorem 7.6. *If two probability distribution P and Q have the same characteristic function, they are equal.*

The injectivity property follows from an explicit inversion recipe. The characteristic function allows us to recover the cumulative distribution function at all its continuity points (just as the Laplace transform did). Again, as continuity points of cumulative distribution functions are dense on $\mathbb{R}$, this is enough.

Proposition 7.10. *Let $X \sim F$ and $Z \sim \mathcal{N}(0, 1)$ be independent. Let $Y = X + \sigma Z$, then:*

▯ *the distribution of Y has characteristic function*

$$\widehat{F}_\sigma(t) = \widehat{\Phi}(t\sigma) \times \widehat{F}(t) = \mathrm{e}^{-\frac{t^2\sigma^2}{2}} \widehat{F}(t)$$

▯ *the distribution of Y is absolutely continuous with respect to Lebesgue measure*

▯ *a version of the density of the distribution of Y is given by*

$$\frac{1}{2\pi} \int_{\mathbb{R}} \mathrm{e}^{-\frac{t^2\sigma^2}{2}} \widehat{F}(t) \mathrm{e}^{-ity} \mathrm{d}t = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{F}_\sigma(t) \mathrm{e}^{-ity} \mathrm{d}t \,.$$

Why can we take for granted the existence of a probability space with two independent random variables X, Z distributed as above?

The proposition states that a density of the distribution of $X + \sigma Z$ can be recovered from the characteristic function of the distribution of $X + \sigma Z$ by the Fourier inversion formula for functions with integrable Fourier transforms.

Proof. The fact that for any $\sigma > 0$, the distribution of $Y = X + \sigma Z$ is absolutely continuous with respect to Lebesgue measure comes from Proposition @ref(prp:convsum).

A density of the distribution of $X + \sigma Z$ is given by

$$\int_{\mathbb{R}} \frac{1}{\sigma} \phi\Big(\frac{y-x}{\sigma}\Big) \mathrm{d}F(x)$$

The characteristic function of Y at t is $\mathrm{e}^{-\frac{t^2\sigma^2}{2}} \widehat{F}(t)$.

$$\begin{aligned} \mathbb{P}\Big\{Y \leq u\Big\} & = \int_{-\infty}^u \int_{\mathbb{R}} \tfrac{1}{\sigma}\phi\Big(\tfrac{y-x}{\sigma}\Big)\mathrm{d}F(x)\mathrm{d}y \\ & = \int_{-\infty}^u \int_{\mathbb{R}} \tfrac{1}{\sigma} \left(\tfrac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{e}^{-it\frac{y-x}{\sigma}} \mathrm{d}t\right) \mathrm{d}F(x)\mathrm{d}y \\ & = \int_{-\infty}^u \left(\int_{\mathbb{R}} \tfrac{1}{\sigma}\tfrac{1}{2\pi}\mathrm{e}^{-\frac{t^2}{2}} \mathrm{e}^{-\frac{ity}{\sigma}} \left(\int_{\mathbb{R}} \mathrm{e}^{\frac{itx}{\sigma}} \mathrm{d}F(x)\right) \mathrm{d}t\right) \mathrm{d}y \\ & = \int_{-\infty}^u \left(\int_{\mathbb{R}} \tfrac{1}{\sigma}\tfrac{1}{2\pi}\mathrm{e}^{-\frac{t^2}{2}} \mathrm{e}^{-\frac{ity}{\sigma}} \widehat{F}(t/\sigma)\mathrm{d}t\right) \mathrm{d}y \\ & = \int_{-\infty}^u \left(\tfrac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\frac{t^2\sigma^2}{2}} \mathrm{e}^{-ity}\widehat{F}(t)\mathrm{d}t\right) \mathrm{d}y\,. \end{aligned}$$

The quantity $\left(\frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\frac{t^2\sigma^2}{2}} \mathrm{e}^{-ity}\widehat{F}(t)\mathrm{d}t\right)$ is a version of the density of the distribution of $Y = X + \sigma Z$ (why?). Note that it is obtained from the same inversion formula that readily worked for the Gaussian density. □

Now we have to show that an inversion formula works for all probability distributions, not only for the smooth probability distributions obtained by adding Gaussian noise. We shall check that we can recover the distribution function from the Fourier transform.

Theorem 7.7. *Let X be distributed according to P, with cumulative distribution function F and characteristic function $\widehat{F}$.*

Then:

$$\lim_{\sigma \downarrow 0} \int_{-\infty}^u \left(\frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-ity} \mathrm{e}^{-\frac{t^2\sigma^2}{2}} \widehat{F}(t)\mathrm{d}t\right) \mathrm{d}y = F(u\_) + \frac{1}{2}P\{u\}$$

where

$$F(u\_) = \lim_{v \uparrow u} F(v) = P(-\infty, u)\,.$$

Proof. The proof consists in combining Propositions @ref(prp:approxident) and @ref(prp:reginversion). □

Note that Theorem 7.7) does not deliver directly the distribution function F. Indeed, if F is not continuous, $u \mapsto \widetilde{F}(u) = F(u\_) + \frac{1}{2}P\{u\}$, is not a distribution function. But the right-continuous modification of $\widetilde{F}$: $u \mapsto \lim_{v\downarrow u} \widetilde{F}(v)$ coincides with F. We have established Theorem 7.6).

When the distribution function is absolutely continuous, Fourier inversion is simpler.

Let X be distributed according to P, with cumulative distribution function F and characteristic function $\widehat{F}$. Assume that $\widehat{F}$ is integrable (with respect to Lebesgue measure). Then:

- P is absolutely continuous with respect to Lebesgue measure;
- $y \mapsto \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{F}(t)\mathrm{e}^{-ity}\mathrm{d}t$ is a uniformly continuous version of the density of P.

Proof. Let X be distributed according to P with cumulative distribution function F and characteristic function $\widehat{F}$. Let Z be independent from X and $\mathcal{N}(0,1)$. Let x be a continuity point of F.

$$\lim_{\sigma\downarrow 0} P\Big\{X+\sigma Z\leq x\Big\} = F(x)$$

$$\begin{aligned}\lim_{\sigma\downarrow 0} P\Big\{X+\sigma Z\leq x\Big\} \quad &= \lim_{\sigma\downarrow 0}\textstyle\int_{-\infty}^{x}\Big(\frac{1}{2\pi}\int_{\mathbb{R}}\mathrm{e}^{-\frac{t^2\sigma^2}{2}}\,\mathrm{e}^{-ity}\widehat{F}(t)\mathrm{d}t\Big)\,\mathrm{d}y\\ &= \textstyle\int_{-\infty}^{x}\frac{1}{2\pi}\int_{\mathbb{R}}\lim_{\sigma\downarrow 0}\mathrm{e}^{-\frac{t^2\sigma^2}{2}}\,\mathrm{e}^{-ity}\widehat{F}(t)\mathrm{d}t\mathrm{d}y\\ &= \textstyle\int_{-\infty}^{x}\frac{1}{2\pi}\int_{\mathbb{R}}\mathrm{e}^{-ity}\widehat{F}(t)\mathrm{d}t\mathrm{d}y\end{aligned}$$

where interversion of limit and integration is justified by dominated convergence. □

We close this section by an alternative inversion formula.

Theorem 7.8 (Inversion formula)**.** *Let P be a probability distribution over $\mathbb{R}$ with cumulative distribution function F, then*

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{\mathrm{e}^{-ita}-\mathrm{e}^{-itb}}{it}\widehat{F}(t)\mathrm{d}t = F(b\_)-F(a)+\frac{1}{2}\Big(P\{b\}+P\{a\}\Big)\,.$$

The proof of Theorem 7.8) can be found in textbooks like (Durrett, 2010) or (Billingsley, 2012).

Let $\widehat{F}$ denote the characteristic function of the probability distribution P, if $\widehat{F}(t)=\mathrm{e}^{-\frac{t^2}{2}}$, then P is the standard univariate Gaussian distribution ($\mathcal{N}(0,1)$).

Let $\widehat{F}$ denote the characteristic function of probability distribution P, if $\widehat{F}(t)=\mathrm{e}^{i\mu t-\frac{\sigma^2t^2}{2}}$, then P is the Gaussian distribution ($\mathcal{N}(\mu,\sigma^2)$).

Another important byproduct of the proof of injectivity of the characteristic function is Stein's identity, an important property of the standard Gaussian distribution.

Theorem 7.9 (Stein's identity)**.** *Let $X\sim\mathcal{N}(0,1)$, and g be a differentiable function such that $\mathbb{E}|g'(X)|<\infty$, then*

$$\mathbb{E}[g'(X)]=\mathbb{E}[Xg(X)]\,.$$

Conversely, if X is a random variable such that

$$\mathbb{E}[g'(X)]=\mathbb{E}[Xg(X)]$$

holds for any differentiable funtion g such that g' is integrable, then $X\sim\mathcal{N}(0,1)$.

Proof. The direct part follows by integration by parts.

To check the converse, note that if X satisfies the identity in the Theorem, then for all $t\in\mathbb{R}$, the functions $t\mapsto\mathbb{E}\cos(tX)$ and $t\mapsto\mathbb{E}\sin(tX)$ satisfy the differential equation $g'(t)=tg(t)$ with conditions $\mathbb{E}\cos(0X)=1$ and $\mathbb{E}\sin(0X)=0$. This entails $\mathbb{E}\mathrm{e}^{itX}=\exp\Big(-\frac{t^2}{2}\Big)$, that is $X\sim\mathcal{N}(0,1)$ □

Differentiability and integrability

Differentiability of the Fourier transform at 0 and integrability are intimately related.

Theorem 7.10. *If X is p-integrable for some $p \in \mathbb{N}$ then the Fourier transform of the distribution of X is p-times differentiable at 0 and the p^{th} derivative equals $i^k \mathbb{E} X^k$.*

Proof. The proof relies on a Taylor expansion with remainder of $x \mapsto \mathrm{e}^{ix}$ at $x = 0$:

$$\mathrm{e}^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} = \frac{i^{n+1}}{n!} \int_0^x (x-s)^n \mathrm{e}^{is} \mathrm{d}s \,.$$

The modulus of the right hand side can be upper-bounded in two different ways.

$$\frac{1}{n+!} \Big| \int_0^x (x-s)^n \mathrm{e}^{is} \mathrm{d}s \Big| \leq \frac{|x|^{n+1}}{(n+1)!}$$

which is good when $|x|$ is small. To handle large values of $|x|$, integration by parts leads to

$$\frac{i^{n+1}}{n!} \int_0^x (x-s)^n \mathrm{e}^{is} \mathrm{d}s = \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1} \left(\mathrm{e}^{is} - 1 \right) \mathrm{d}s \,.$$

The modulus of the right hand side can be upper-bounded by $2|x|^n/n!$.

Applying this Taylor expansion to $x = tX$, using the pointwise upper bounds and taking expectations leads to

$$\begin{aligned} \left| \widehat{F}(t) - \textstyle\sum_{k=0}^n \mathbb{E} \frac{(itX)^k}{k!} \right| & \leq \mathbb{E} \Big[\min \left(\tfrac{|tX|^{n+1}}{(n+1)!}, 2 \tfrac{|tX|^n}{n!} \right) \Big] \\ & = \tfrac{|t|^n}{(n+1)!} \mathbb{E} \Big[\min \Big(|tX|^{n+1}, 2(n+1)|X|^n \Big) \Big] \,. \end{aligned}$$

Note that the right hand side is well defined as soon as $\mathbb{E}|X|^n < \infty$. Now, by dominated convergence,

$$\lim_{t \to 0} \mathbb{E} \Big[\min \Big(|tX|^{n+1}, 2(n+1)|X|^n \Big) \Big] = 0$$

Hence we have established that if $\mathbb{E}|X|^n < \infty$,

$$\widehat{F}(t) = \sum_{k=0}^{n} i^k \mathbb{E} X^k \frac{t^k}{k!} + o(|t|^n) \,.$$

□

In the other direction, the connection is not as simple: differentiability of the Fourier transform does not imply integrability. But the following holds.

Theorem 7.11. *If the Fourier transform $\widehat{F}$ of the distribution of X satisfies*

$$\lim_{h \downarrow 0} \frac{2 - \widehat{F}(h) - \widehat{F}(-h)}{h^2} = \sigma^2 < \infty$$

then $\mathbb{E} X^2 = \sigma^2, .$

Proof. Note that

$$2 - \widehat{F}(h) - \widehat{F}(-h) = 2 \mathbb{E} \Big[1 - \cos(hX) \Big] \,,$$

and using Taylor with remainder formula for cos at 0:

$$1 - \cos x = \int_0^x \cos(s)(x-s)\mathrm{d}s = x^2 \int_0^1 \cos(sx)(1-s)\mathrm{d}s$$

Note that $\int_0^1 \cos(sx)(1-s)\mathrm{d}s \geq 0$ for all $x \in \mathbb{R}$.

$$\begin{aligned} \frac{2\mathbb{E}\left[1-\cos(hX)\right]}{h^2} &= 2\frac{\mathbb{E}\left[h^2X^2\int_0^1 \cos(shX)(1-s)\mathrm{d}s\right]}{h^2} \\ &= 2\mathbb{E}\left[X^2\int_0^1 \cos(shX)(1-s)\mathrm{d}s\right]. \end{aligned}$$

By Fatou's Lemma:

$$\sigma^2 = \lim_{h\downarrow 0} 2\mathbb{E}\left[X^2 \int_0^1 \cos(shX)(1-s)\mathrm{d}s\right] \geq 2\mathbb{E}\left[\liminf_{h\downarrow 0} X^2 \int_0^1 \cos(shX)(1-s)\mathrm{d}s\right]$$

but for all $x \in \mathbb{R}$, by dominated convergence,

$$\liminf_{h\downarrow 0} x^2 \int_0^1 \cos(shx)(1-s)\mathrm{d}s = \frac{x^2}{2}.$$

Hence

$$\sigma^2 \geq \mathbb{E}X^2.$$

The proof is completed by invoking Theorem 7.10). □

Another application: understanding Cauchy distribution

Assume U is uniformly distributed over $]0,1[$, let the real valued random variable X be defined by

$$X = \tan\left(\frac{\pi}{2}(2 \times U - 1)\right)$$

.

As tan is continuously increasing from $-\pi/2$ to $\pi/2$, the cumulative distribution function of the distribution of X is

$$\begin{aligned} \mathbb{P}\{X \leq x\} &= \mathbb{P}\left\{\tan\left(\tfrac{\pi}{2}(2U-1)\right) \leq x\right\} \\ &= \mathbb{P}\left\{U \leq \tfrac{1}{2} + \tfrac{1}{\pi}\arctan(x)\right\} \\ &= \tfrac{1}{2} + \tfrac{1}{\pi}\arctan(x) \end{aligned}$$

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

As arctan has derivative $x \mapsto \frac{1}{1+x^2}$, the cumulative distribution function is absolutely continuous with density:

$$\frac{1}{\pi}\frac{1}{1+x^2}$$

This is the density of the Cauchy distribution.

Note that $\mathbb{E}(X)_+ = \mathbb{E}(X)_- = \mathbb{E}|X| = \infty$. The Cauchy distribution is not integrable.

Now, assume $X_1, X_2, , \ldots, X_n$ are i.i.d. and Cauchy distributed. Let $Z = \sum_{i=1}^n X_i/n$. How is Z distributed? We might compute the convolution power of the Cauchy density. It turns out that starting from the characteristic function is much more simple.

We refrain from computing directly the characteristic function of the Cauchy distribution. We take a roundabout.

Let Y be distributed according to Laplace distribution, that is with density $y \mapsto \frac{1}{2}\exp(-|y|)$ for $y \in \mathbb{R}$. The random variable Y is symmetric ($Y \sim -Y$). Let $\widehat{F}_Y$ denote the characteristic function of (the distribution of) Y.

$$\begin{aligned}
\widehat{F}_Y(t) &= \mathbb{E}\mathrm{e}^{tY} \\
&= \mathbb{E}\cos(tY) \\
&= \int_0^\infty \mathrm{e}^{-y}\cos(ty)\mathrm{d}y \\
&= \left[-\mathrm{e}^{-y}\cos(ty)\right]_0^\infty - t\int_0^\infty \mathrm{e}^{-ty}\sin(ty)\mathrm{d}y \\
&= 1 - t\int_0^\infty \mathrm{e}^{-y}\sin(ty)\mathrm{d}y \\
&= 1 - t\left[-\mathrm{e}^{-y}\sin(ty)\right]_0^\infty - t^2\int_0^\infty \mathrm{e}^{-y}\cos(ty)\mathrm{d}y \\
&= 1 - t^2\widehat{F}_Y(t)
\end{aligned}$$

where we have performed integration by parts twice.

The characteristic function $\widehat{F}_Y$ satisfies

$$\widehat{F}_Y(t) = \frac{1}{1+t^2}\,,$$

up to $\frac{1}{\pi}$, this is the density of the Cauchy distribution.

$$\begin{aligned}
\widehat{F}_X(t) &= \mathbb{E}\mathrm{e}^{itX} \\
&= \int_{-\infty}^\infty \frac{1}{\pi}\frac{1}{1+x^2}\cos(tx)\mathrm{d}x \\
&= \frac{2}{\pi}\int_0^\infty \cos(tx)\widehat{F}_Y(x)\mathrm{d}x \\
&= 2 \times \frac{1}{2\pi}\int_{-\infty}^\infty \mathrm{e}^{-itx}\widehat{F}_Y(x)\mathrm{d}x \\
&= 2 \times \frac{1}{2}\mathrm{e}^{-|t|} = \mathrm{e}^{-|t|}
\end{aligned}$$

where we have used the inversion formula.

Now, the characteristic function of the distribution of Z is

$$\widehat{F}_Z(t) = \left(\mathrm{e}^{-\frac{|t|}{n}}\right)^n = \widehat{F}_X(t)$$

which means $Z \sim X$.

The basic tools of characteristic functions theory allow us to - compute the characteristic function of the Laplace distribution - compute the characteristic function of the Cauchy distribution by inversion - compute the characteristic function of sums of independant Cauchy random variables - show that the Cauchy distribution is 1-stable.

The density of the Laplace distribution is not differentiable at 0, this is reflected in the fact that its Fourier transform (the characteristic function of the Laplace distribution) is not integrable.

Conversely the lack of integrability of the Cauchy distribution is reflected in the non-differentiability of its characteristic function at 0.

7.5 Quantile functions

So far we have seen several characterizations of probability distributions: cumulative distribution functions, Laplace transform for distributions supported on $[0, \infty)$, characteristic functions. The last characterization is praised for its behavior with respect to sums of independent random variables.

For univariate distributions, a companion to the cumulative distribution function is the quantile function. It plays a significant role in simulations, statistics and risk theory.

A cumulative distribution function F is non-negative, $[0,1]$-valued, non-decreasing, right-continuous, with left-limit at any point. The cumulative distribution function of a diffuse probability measure is continuous at any point.

The quantile function $F^{\leftarrow}$ is defined as an extended inverse of the cumulative distribution function F.

Definition 7.3 (Quantile function)**.** The quantile function $F^{\leftarrow}$ of random variable X distributed according to P (with cumulative distribution function F) is defined as

$$\begin{aligned} F^{\leftarrow}(p) \quad &= \inf\Big\{x : P\{X \leq x\} \geq p\Big\} \\ &= \inf\Big\{x : F(x) \geq p\Big\} \qquad \text{for } p \in (0,1). \end{aligned}$$

The quantile function is non-decreasing and left-continuous. The interplay between the quantile and cumulative distribution functions is summarized in the next proposition.

Proposition 7.11. *If F and $F^{\leftarrow}$ are the cumulative distribution function and the quantile function of (the distribution of) X, the following statements hold for $p \in]0,1[$:*

1. $p \leq F(x)$ ***iff*** $F^{\leftarrow}(p) \leq x$.
2. $F \circ F^{\leftarrow}(p) \geq p$.
3. $F^{\leftarrow} \circ F(x) \leq x$.
4. *If F is absolutely continuous, then* $F \circ F^{\leftarrow}(p) = p$

Proof. According to the definition of $F^{\leftarrow}$ if $F(x) \geq p$ then $F^{\leftarrow}(p) \leq x$.

To prove the converse, it suffices to check that $F \circ F^{\leftarrow}(p) \geq p$.

Indeed, if $x \geq F^{\leftarrow}(p)$, as F is non-decreasing $F(x) \geq F \circ F^{\leftarrow}(p)$. Si $y = F^{\leftarrow}(p)$, par definition de $y = F^{\leftarrow}(p)$, il existe une non-increasing sequence $(z_n)_{n\in\mathbb{N}}$ which converges to y such that $F(z_n) \geq p$. Mais as F is right-continuous $\lim_n F(z_n) = F(\lim_n z_n) = F(y)$. Hence $F(y) \geq p$.

We just proved 1. and 2.

3.) is an immediate consequence de 1). Let $p = F(x)$. Hence $p \leq F(x)$, according to 1.) this is equivalent to $F^{\leftarrow}(p) \leq x$, that is $F^{\leftarrow} \circ F(x) \leq x$.

4.) For every p in $]0,1[$, $\{x : p = F(x)\}$ is non-empty (Mean value Theorem). Let $y = \inf\{x : p = F(x)\} = F^{\leftarrow}(p)$. According to 1), $F(y) \geq p$. Now, if $(z_n)_{n\in\mathbb{N}}$ is an increasing sequence converging to y, for every n, $F(z_n) < p$, and, by left-continuity, $F(y) = F(\lim_n z_n) = \lim_n F(z_n) \leq p$. Hence $F(y) = p$, that is $F \circ F^{\leftarrow}(p) = p$. □

Proposition 7.12 (Quantile transformation)**.** *If U is uniformly distributed on $(0,1)$, and F is a cumulative distribution over $\mathbb{R}$, $F^{\leftarrow}(U)$ has cumulative distribution F.*

Proof.

$$\begin{aligned} P\Big\{F^{\leftarrow}(U) \leq x\Big\} \quad &= P\Big\{U \leq F(x)\Big\} \\ &= F(x)\,. \end{aligned}$$

□ □

Remark 7.2. The quantile transformation works whatever the continuity properties of F.

The quantile transformation has many applications. It can be used to show stochastic domination properties.

Example 7.5. In Figures 7.1 up to Figure 7.4, we illustrate quantile functions for discrete (binomial) distributions and for distributions that are neither discrete nor continuous. The quantile function of a discrete distribution is step function that jumps at the cumulative probability of every possible outcome. If a probability distribution is a mixture of a discrete distribution and a continuous distribution, the quantile function jumps at the cumulative probability of every possible outcome of the discrete component.

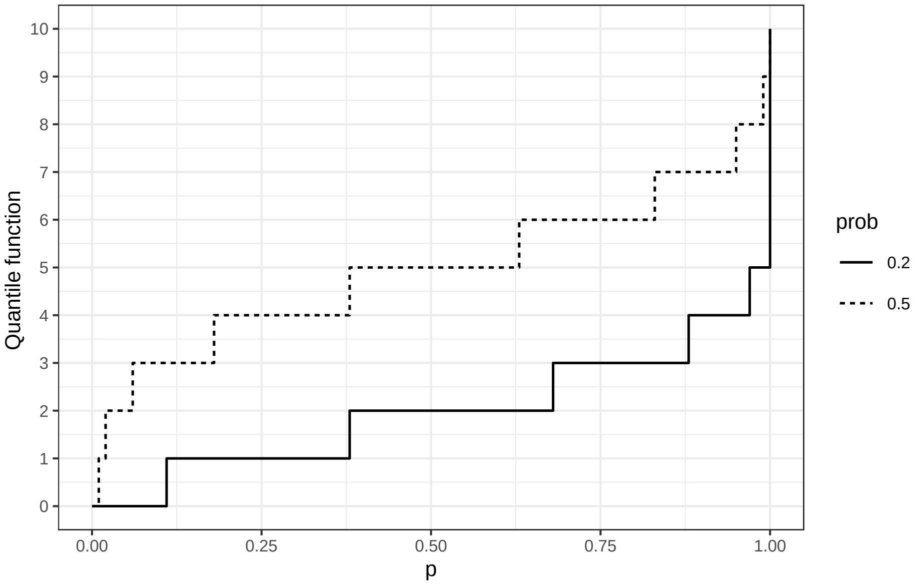

Figure 7.1

Let us conclude this section with an important observation. concerning the behavior of $F(X)$ when $X \sim P$ with cumulative distribution function F.

Corollary 7.4. *If $X \sim P$ with continuous cumulative distribution function F, then $F(X)$ and $1 - F(X)$ are uniformly distributed on* $[0, 1]$.

Exercise 7.2. Prove Corollary 7.4

7.6 Bibliographic remarks

Wilf (2005) explores the interplay between combinatorics, algorithm analysis and generating function theory.

Widder (2015) is a classic reference on Laplace transforms. Laplace transforms play an important role in Point Process Theory, and Extreme Value Theory, to name a few fields of application.

The first part of Chapter 9 from Pollard (2002) describes characteristic functions as Fourier transforms. Properties and applications of characteristic functions are thoroughly discussed in (Durrett, 2010), (Billingsley, 2012).

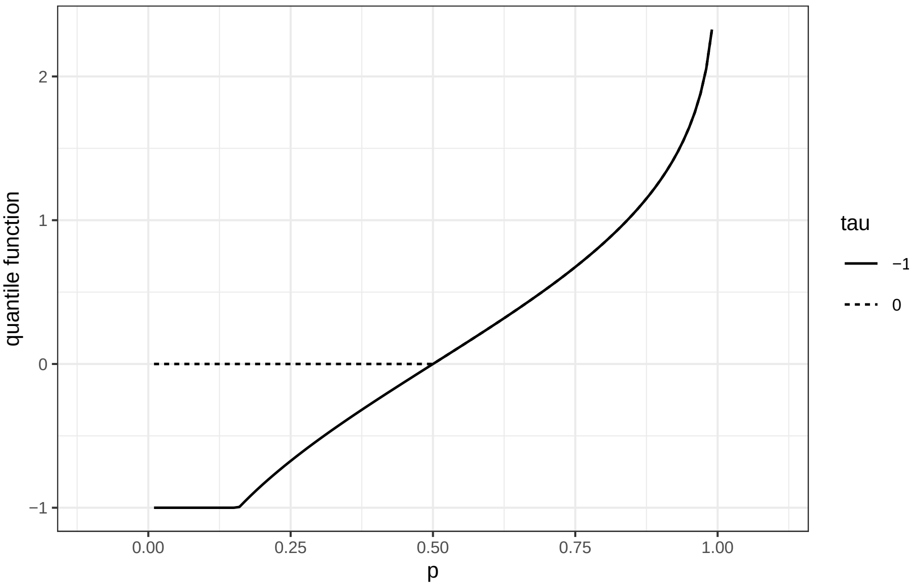

Figure 7.2: Quantile functions $\max(X, \tau)$ where $X \sim \mathcal{N}(0, 1)$ for $\tau \in \{0, 2\}$. Let $\Phi^{\leftarrow}$ denote the quantile function of $\mathcal{N}(0, 1)$. The quantile function of $\max(X, \tau)$ is $\mathbb{I}_{(0,\Phi(\tau)]}(p) \times \tau + \Phi^{\leftarrow}(p) \times \mathbb{I}_{(\Phi(\tau),1)}(p) = \Phi^{\leftarrow}(p \vee \Phi(\tau))$. The two distributions are neither absolutely continuous nor discrete.

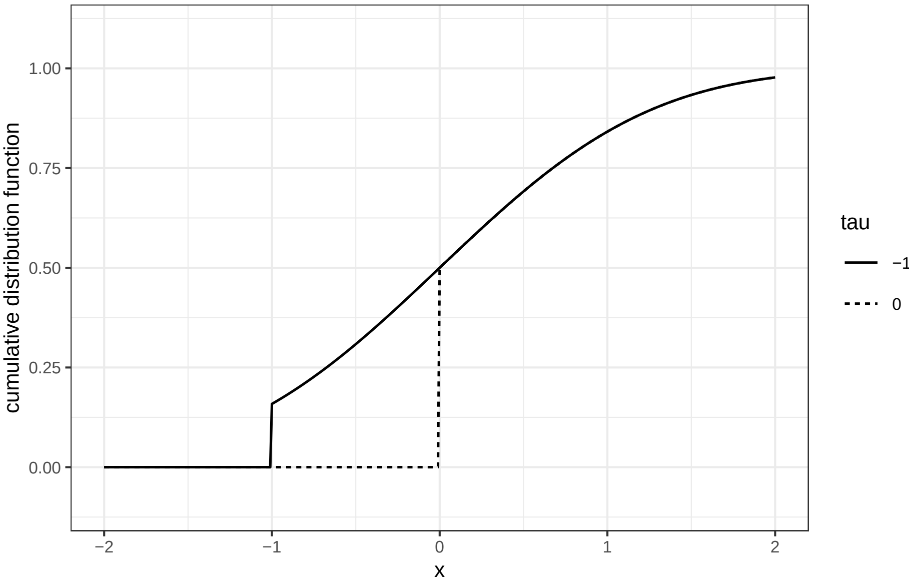

Figure 7.3: Cumulative distribution functions for the probability distributions illustrated in Figure 7.2

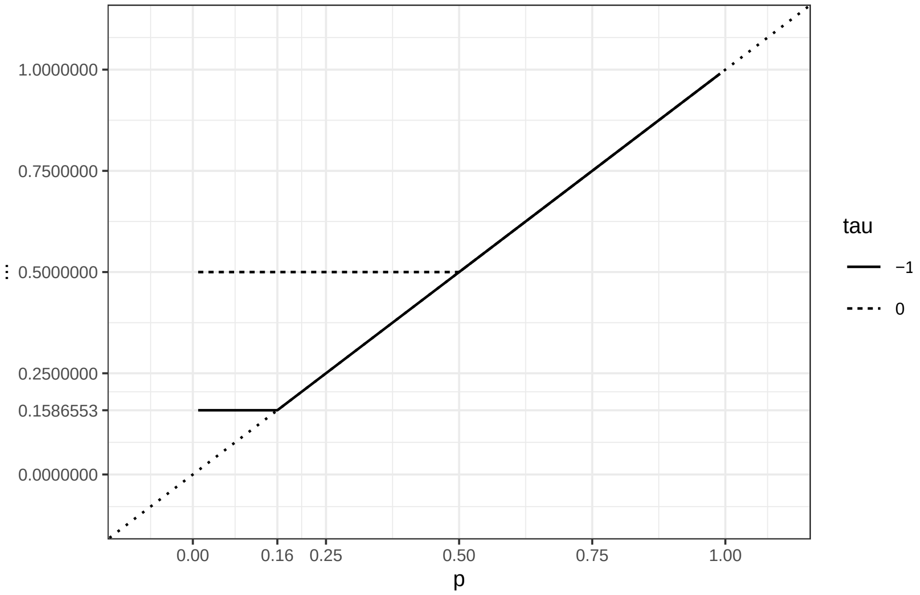

Figure 7.4: Representation of $F \circ F^{\leftarrow}$ for the probability distributions illustrated in Figures 7.2 and Figure 7.3. The function $F \circ F^{\leftarrow}$ always lies above the line $y = x$ (dotted line) as prescribed in Proposition 7.11. Plateaux that lie strictly above the dotted line are in correspondence with jumps of the quantile function.

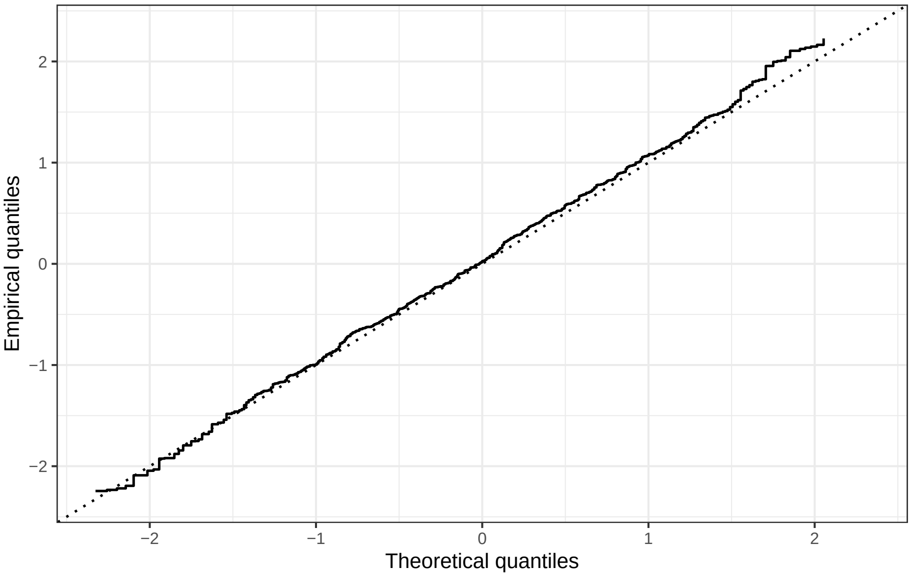

Figure 7.5

Chapter 8

Conditioning

8.1 Defining conditional expectation

In this and the following sections, $(\Omega, \mathcal{F}, P)$ is a probability space, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. The sub-σ-algebra need not be ***atomic*** as in Chapter 5. We cannot define conditional probabilities by conditioning with respect to atoms generating $\mathcal{G}$. Our objective is nervertheless to define conditional expectations with respect to sub-σ-algebra $\mathcal{G}$, while retaining the nice properties surveyed in Chapter 5.

The general definition of conditional expectation starts from the property described in Proposition 5.4.

Definition 8.1 (Conditional expectation)**.** Let $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, then a random variable Y is a version of the conditional expectation of X with respect to $\mathcal{G}$ iff

i. Y is $\mathcal{G}$-measurable.
ii. For every event B in $\mathcal{G}$:

$$\mathbb{E}\left[\mathbb{I}_B X\right] = \mathbb{E}\left[\mathbb{I}_B Y\right] .$$

Leaving aside the question of the existence of a version of conditional expectation of X, we first check that if there exist different versions, they differ only up to a negligible event.

Let $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$, then if Y' and Y are two versions of the conditional expectation of X with respect to $\mathcal{G}$:

$$P\{Y = Y'\} = 1.$$

Proof. As Y and Y' are $\mathcal{G}$-measurable, the event

$$B = \{\omega \,:\, Y(\omega) > Y'(\omega)\}$$

belongs to $\mathcal{G}$. Moreover,

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_B X\right] &= \mathbb{E}\left[\mathbb{I}_B Y\right] \\ &= \mathbb{E}\left[\mathbb{I}_B Y'\right] .\end{aligned}$$

Thus

$$\mathbb{E}\left[\mathbb{I}_B (Y - Y')\right] = 0 .$$

As random variable $\mathbb{I}_B(Y - Y')$ is non-negative, its expectation is zero, it is null with probability 1. Thus

$$P\{Y > Y'\} = 0 \,.$$

We can conclude by proceeding in a similar way for event $\{Y < Y'\}$.
□ □

Still postponing the existence question, let us check now a few properties versions of conditional expectation of X should satisfy.

Let $X_1, X_2 \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$, a_1, a_2 two real numbers, then if Y_1 Y_2 and Z are respectively versions versions of conditional expectation of X_1, X_2 and $a_1X_1 + a_2X_2$ with respect to $\mathcal{G}$, we have

$$P\{a_1Y_1 + a_2Y_2 = Z\} = 1 \,.$$

Proof. Let B be the event of $\mathcal{G}$ defined by

$$\{a_1Y_1 + a_2Y_2 > Z\} \,.$$

We get

$$\begin{array}{rcl} \mathbb{E}[\mathbb{I}_B Z] & = & \mathbb{E}[\mathbb{I}_B(a_1X_1 + a_2X_2)] \\ & = & a_1\mathbb{E}[\mathbb{I}_B X_1] + a_2\mathbb{E}[\mathbb{I}_B X_2] \\ & = & a_1\mathbb{E}[\mathbb{I}_B Y_1] + a_2\mathbb{E}[\mathbb{I}_B Y_2] \\ & = & \mathbb{E}[\mathbb{I}_B(a_1Y_1 + a_2Y_2)] \,, \end{array}$$

and thus

$$\mathbb{E}[\mathbb{I}_B(Z - (a_1Y_1 + a_2Y_2))] = 0 \,.$$

We conclude as in the proceeding proof that $P\{B\} = 0$.

The proof is completed by handling in a similar way the event $\{a_1Y_1 + a_2Y_2 < Z\}$.
□ □

Proposition 8.1. *If $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ algebra of $\mathcal{F}$. If Z is a version the conditional expectation of X with respect to $\mathcal{G}$ and if X is P-a.s. non-negative, then*

$$P\{Z \geq 0\} = 1 \,.$$

The proof reproduces the argument used to established that different versions of the conditional expectation are almost surely equal.

Proof. For $n \in \mathbb{N}$, let B_n denote the event (from $\mathcal{G}$) defined by

$$B_n = \left\{ \mathbb{E}\left[X \mid \mathcal{G}\right] < -\frac{1}{n} \right\} .$$

To prove the proposition, it is enough to check

$$P\{\cup_n B_n\} = 0.$$

As $P\{\cup_n B_n\} = \lim_n P\{B_n\}$, it suffices to check $P\{B_n\} = 0$, for all n, $P\{B_n\} = 0$. For all n,

$$\begin{aligned} 0 &\leq \mathbb{E}[\mathbb{1}_{B_n} X] \\ &= \mathbb{E}\left[\mathbb{1}_{B_n} X\right] \\ &= \mathbb{E}\left[\mathbb{1}_{B_n} \mathbb{E}\left[X \mid \mathcal{G}\right]\right] \\ &\leq -\frac{P\{B_n\}}{n} . \end{aligned}$$

Hence, for all n, $P\{B_n\} = 0$.
□ □

The next corollary is a consequence of Proposition 8.1.

Corollary 8.1. *If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables from $\mathcal{L}_1(\Omega, \mathcal{F}, P)$ satisfying $X_{n+1} \geq X_n$ P-a.s. then there exists an P-a.s. non-decreasing sequence of versions of conditional expectations*

$$\forall n \in \mathbb{N}, \quad \mathbb{E}\left[X_{n+1} \mid \mathcal{F}\right] \geq \mathbb{E}\left[X_n \mid \mathcal{F}\right].$$

Let $\mathcal{E}$ be a π-system generating $\mathcal{G}$ and containing Ω. Check that $\mathbb{E}\left[X \mid \mathcal{G}\right]$ is the unique element from $\mathcal{L}_1\left(\Omega, \mathcal{G}, P\right)$ which satisfies

$$\forall B \in \mathcal{E}, \quad \mathbb{E}\left[\mathbb{1}_B X\right] = \mathbb{E}\left[\mathbb{1}_B \mathbb{E}\left[X \mid \mathcal{G}\right]\right].$$

For nested sub-σ-algebras, conditional expectations satisfy the tower property:

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$ be two nested sub-σ-algebras. Then for every $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$:

$$\mathbb{E}\Big[\mathbb{E}\left[X \mid \mathcal{G}\right] \mid \mathcal{H}\Big] = \mathbb{E}\Big[\mathbb{E}\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\Big] = \mathbb{E}\left[X \mid \mathcal{G}\right] \qquad \text{a.s.}$$

Proof. Almost sure equality $\mathbb{E}\Big[\mathbb{E}\left[X \mid \mathcal{G}\right] \mid \mathcal{H}\Big] = \mathbb{E}\left[X \mid \mathcal{G}\right]$ is trivial: any $\mathcal{G}$-measurable random variable is also $\mathcal{H}$-measurable.

Let us now check the second equality.

For every $B \in \mathcal{G}$,

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_B \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \mid \mathcal{G}\right]\right] &= \mathbb{E}\left[\mathbb{1}_B \mathbb{E}\left[X \mid \mathcal{H}\right]\right] \\ & \quad \text{comme } B \in \mathcal{H} \\ &= \mathbb{E}\left[\mathbb{1}_B X\right]. \end{aligned}$$

□

8.2 Conditional expectation in $\mathcal{L}_2(\Omega, \mathcal{F}, P)$

If we focus on square-integrable random variables, building versions of conditional expectation turn out to be easy. Recall that when the conditioning sub-σ-algebra $\mathcal{G}$ is atomic, according to Proposition 5.5, the condition expectation $\mathbb{E}[X \mid \mathcal{G}]$ defines an optimal predictor of X with respect to quadratic error amongst $\mathcal{G}$-measurable random variables. This characterization remains valid for square integrable random variables even when the conditioning sub-σ-algebra is no more atomic. This is the content of the next theorem.

Theorem 8.1 (Conditional expectation for square integrable random variables)**.** *Let be $X \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$.*

There exists $Y \in \mathcal{L}_2(\Omega, \mathcal{G}, P)$ that minimizes the L_2 distance to X:

$$\exists Y \in \mathcal{L}_2(\Omega, \mathcal{G}, P) \qquad \mathbb{E}(Y - X)^2 = \inf\Big\{\mathbb{E}(Z - X)^2 : Z \in \mathcal{L}_2(\Omega, \mathcal{G}, P)\Big\},$$

that is, Y represents a version of the orthogonal projection *of X on $\mathcal{L}_2(\Omega, \mathcal{G}, P)$.*

A version Y of the orthogonal projection *of X on $\mathcal{L}_2(\Omega, \mathcal{G}, P)$ is also a version of the conditional expectation of X with respect to $\mathcal{G}$:*

$$\forall B \in \mathcal{G}, \quad \mathbb{E}\left[\mathbb{I}_B X\right] = \mathbb{E}\left[\mathbb{I}_B Y\right] .$$

Note that theorem contains two statements: first, there exists a minimizer of $\mathbb{E}(X - Z)^2$ in $\mathcal{L}_2(\omega, \mathcal{F}, P)$, second, such a minimizer is a version of condition expectation defined according to Definition 8.1. Checking the first statement amounts to invoke the right arguments from Hilbert spaces theory.

For the sake of self-reference, we recall basics if Hilbert spaces theory.

Definition 8.2 (Hilbert's space)**.** A real vector space E equipped with a norm $\|\cdot\|$ is a Hilbert space iff $\langle\cdot,\cdot\rangle$ defined by

$$\forall x, y \in E, \langle x, y\rangle = \frac{1}{4}\Big(\|x + y\|^2 + \|x - y\|^2\Big)$$

is an *inner product* and E is complete for the topology induced by the norm.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, then the set $L_2(\Omega, \mathcal{F}, P)$ of equivalence classes of square integrable variables, equipped with $\|X\|^2 = (\mathbb{E}X^2)^{1/2}$ is a Hilbert space.

Remark 8.1. In this context,

$$\langle X, Y\rangle = \mathbb{E}\left[XY\right] .$$

From Hilbert space theory, the essential tool we shall use is the projection Theorem below. Our starting point is the next observation (that follows from results in Chapter 3).

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra, then $L_2(\Omega, \mathcal{G}, P)$ is a closed convex subset (subspace) of $L_2(\Omega, \mathcal{F}, P)$.

We look for the element from $L_2(\Omega, \mathcal{G}, P)$ that is closest (in the L_2 sense) to a random variable from $L_2(\Omega, \mathcal{F}, P)$. The existence and unicity of this closest $\mathcal{G}$-measurable random variable are warranted by the Projection Theorem.

Theorem 8.2 (Projection Theorem)**.** *Let E be a Hilbert space and F a closed convex subset of F. For every $x \in E$, there exists a unique $y \in F$, such that*

$$\|x - y\| = \inf_{z \in F} \|x - z\|.$$

This unique closest point in F is called the (orthogonal) projection of x over F. For any $z \in F$,

$$\langle x - y, z - y\rangle \leq 0\,.$$

If F is a linear subspace of E, the Pythagorean relationship holds:

$$\|x\|^2 = \|y\|^2 + \|x - y\|^2$$

and for any $z \in F$, $\langle x - y, z\rangle = 0$.

Proof. Let $d = \inf_{z \in F} \|x - z\|$. Let $(z_n)_n$ be a sequence of elements from F such that

$$\lim_n \|x - z_n\| = d.$$

According to the parallelogram law,

$$2\left(\|x - z_n\|^2 + \|x - z_m\|^2\right) = \|2x - (z_n + z_m)\|^2 + \|z_n - z_m\|^2.$$

Since F is convex, $(z_n + z_m)/2 \in F$, so

$$\|x - (z_n + z_m)/2\| \geq d\,.$$

Let $\epsilon \in (0,1]$ and n_0 be such that for $n \geq n_0$, $\|x - z_n\| \leq d + \epsilon$. For $n, m \geq n_0$

$$4(d+\epsilon)^2 \geq 4d^2 + \|z_n - z_m\|^2$$

or equivalently

$$\|z_n - z_m\|^2 \leq 4(2d+1)\epsilon\,.$$

Hence, the minimizing sequence $(z_n)_n$ has the Cauchy property. As F is closed, it has a unique limit $y \in F$ and $d = \|x - y\|$.

To verify uniqueness, suppose there exists $y' \in F$, such as $\|x - y'\| = d$. Now, let us build a new sequence $(z'_n)_{n \in \mathbb{N}}$ such that $z'_{2n} = z_n$ and $z'_{2n+1} = y'$. This F-valued sequence satisfies $\lim_n \|z'_n - x\| = d$. By the argument above, it admits a limit y'' in F. The limit y'' coincides with the limit of any sub-sequence, so it equals y and y'.

Fix $z \in F \setminus \{y\}$, for any $u \in (0,1]$, let $z_u = y + u(z - y)$, then $z_u \in F$ and

$$\|x - z_u\|^2 - \|x - y\|^2 = -2u\langle x - y, z - y\rangle + u^2\|z - y\|^2\,.$$

As this quantity is non-negative for $u \in [0,1]$, $\langle x - y, z - y\rangle$ has to be non-positive.

Now suppose that F is a subspace of E.

If there is $y \in F$ such as $\langle x - y, z\rangle = 0$ for any $z \in F$, then y is the orthogonal projection of x on F since for all $z \in F$:

$$\begin{aligned}\|x - z\|^2 &= \|x - y\|^2 - 2\langle x - y, z\rangle + \|z\|^2 \\ &\geq \|x - y\||^2.\end{aligned}$$

Conversely, if y is the orthogonal projection of x on F, for all z of F and all $\lambda \in\in \mathbb{R}$:

$$\begin{aligned}\|x - y\||^2 &\leq \|x - (y + \lambda z)\|^2 \\ &= \|x - y\|^2 - 2\lambda\langle x - y, z\rangle + \lambda^2\|z\|^2,\end{aligned}$$

so $0 \leq 2\lambda\langle x - y, z\rangle + \lambda^2\|z\|^2$. For this polynomial in λ to be of constant sign, it is necessary that $\langle x - y, z\rangle = 0$.

□ □

As $\mathcal{L}_2(\Omega, \mathcal{G}, P)$ is a convex part of $\mathcal{L}_2(\Omega, \mathcal{F}, P)$, the existence and uniqueness of the projection on a closed convex part of a Hilbert space gives the following corollary which matches the first statement in Theorem 8.1).

Given $X \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$, there exists $Y \in \mathcal{L}_2(\Omega, \mathcal{G}, P)$ that minimizes

$$\mathbb{E}\left[(X - Z)^2\right] \qquad \text{for } Z \in \mathcal{L}_2(\Omega, \mathcal{G}, P).$$

Any other minimizer in $\mathcal{L}_2(\Omega, \mathcal{G}, P)$ is P-almost surely equal to~Y.

Proof. Let Ybe a version of the orthogonal projection of X on $L_2(\Omega, \mathcal{G}, P)$ and B an element of $\mathcal{G}$.

The inner product of $\mathbb{I}_B \in \mathcal{L}_2(\Omega, \mathcal{G}, P))$ and $X - Y$is

$$\langle X - Y, \mathbb{I}_B\rangle = \mathbb{E}\left[(X - Y)\mathbb{I}_B\right] .$$

By Theorem 8.2, $\mathbb{E}\left[(X - Y)\mathbb{I}_B\right] = 0$.

□ □

We conclude this section with a Pythagorean theorem for the variance.

Definition 8.3 (Conditional variance)**.** Let $X \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. The ***conditional variance*** of X with respect to $\mathcal{G}$ is defined by

$$\text{Var}\left[X \mid \mathcal{G}\right] = \mathbb{E}\left[(X - \mathbb{E}[X \mid \mathcal{G}])^2 \mid \mathcal{G}\right] .$$

The conditional variance is a ($\mathcal{G}$-measurable) random variable, just as the conditional expectation. It is the conditional expectation of the prediction error that is incurred when trying to predict X using $\mathbb{E}[X \mid \mathcal{G}]$.

Proposition 8.2. *Let $X \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. Then*

$$\text{Var}[X] = \text{Var}\Big[\mathbb{E}\left[X \mid \mathcal{G}\right]\Big] + \mathbb{E}\Big[\text{Var}\left[X \mid \mathcal{G}\right]\Big] .$$

Proof. Recall that $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right]\right] = \mathbb{E}\left[X\right]$.

$$\begin{aligned}
\text{Var}\left[X\right] &= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}\left[X \mid \mathcal{G}\right] + \mathbb{E}\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right])^2\right] \\
&= \mathbb{E}\left[(X - \mathbb{E}\left[X \mid \mathcal{G}\right])^2\right] \\
&\qquad + 2\mathbb{E}\left[(X - \mathbb{E}\left[X \mid \mathcal{G}\right])\left(\mathbb{E}\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)\right] \\
&\qquad + \mathbb{E}\left[(\mathbb{E}\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right])^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(X - \mathbb{E}\left[X \mid \mathcal{G}\right])^2 \mid \mathcal{G}\right]\right] \\
&\qquad + 2\mathbb{E}\Big[\mathbb{E}\left[(X - \mathbb{E}\left[X \mid \mathcal{G}\right]) \mid \mathcal{G}\right]\left(\mathbb{E}\left[X \mid \mathcal{G}\right] - \mathbb{E}\left[X\right]\right)\Big] \\
&\qquad + \text{Var}\left[\mathbb{E}\left[X \mid \mathcal{G}\right]\right] \\
&= \mathbb{E}\left[\text{Var}\left[X \mid \mathcal{G}\right]\right] + \text{Var}\left[\mathbb{E}\left[X \mid \mathcal{G}\right]\right] .
\end{aligned}$$

□ □

8.3 Conditional expectation in $\mathcal{L}_1(\Omega, \mathcal{F}, P)$

To construct the conditional expectation of a random variable, square-integrability is not necessary. This is the meaning of the next theorem.

Theorem 8.3. *If $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$, then there exists an integrable $\mathcal{G}$-measurable random variable, denoted by $\mathbb{E}\left[Y \mid \mathcal{G}\right]$ such that*

$$\forall B \in \mathcal{G}, \mathbb{E}\left[\mathbb{I}_B Y\right] = \mathbb{E}\left[\mathbb{I}_B \mathbb{E}\left[Y \mid \mathcal{G}\right]\right] .$$

In words, conditional expectations according to Definition 8.1 exist for all integrable random variables and all sub-σ-algebras.

Exercise 8.1. Let $\mathcal{G}'$ be a π-system that contains Ω and generates $\mathcal{G}$. If Z is an integrable $\mathcal{G}$-measurable variable that satisfies

$$\forall B \in \mathcal{G}', \mathbb{E}\left[\mathbb{I}_B Y\right] = \mathbb{E}\left[\mathbb{I}_B \mathbb{E}\left[Y \mid \mathcal{G}\right]\right]$$

then $Z = \mathbb{E}\left[Y \mid \mathcal{G}\right]$.

To establish the Theorem 8.3, we use the *usual machinery* of limiting arguments.

Proposition 8.3. *If $(Y_n)_n$ is a non-decreasing sequence of non-negative square-integrable random variables such as $Y_n \uparrow Y$ a.s. then there exists a $\mathcal{G}$-measurable random variable Z such that*

$$\mathbb{E}\left[Y_n \mid \mathcal{G}\right] \uparrow Z \qquad \textit{a.s.}$$

Proof. As $(Y_n)_n$ is non-decreasing, according to Proposition 8.1 $\left(\mathbb{E}\left[Y_n \mid \mathcal{G}\right]\right)_n$ is an (a.s.) non-decreasing sequence of $\mathcal{G}$-measurable random variables, it admits a $\mathcal{G}$-measurable limit (finite or not).

□ □

We now proceed to the proof of Theorem 8.3.

Proof. Without losing in generality, we assume $Y \geq 0$ (if this is not the case, let $Y = (Y)_+ - (Y)_-$ with $(Y)_+ = |Y|\mathbb{I}_{Y \geq 0}$ and $(Y)_- = |Y|\mathbb{I}_{Y<0}$, handle $(Y)_+$ and $(Y)_-$ separately and use the linearity of conditional expectation).

Let

$$Y_n = Y\mathbb{I}_{|Y| \leq n}$$

so that $Y_n \nearrow Y$ everywhere. The random variable Y_n is bounded and thus square-integrable. The random variable

$\mathbb{E}\left[Y_n \mid \mathcal{G}\right]$ is therefore well defined for each n.

The sequence $\mathbb{E}\left[Y_n \mid \mathcal{G}\right]$ is P-a.s. monotonous. It converges monotonously towards a $\mathcal{G}$-measurable random variable Z which takes values in $\mathbb{R}_+ \cup \{\infty\}$. We need to check that this random variable $Z \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$.

By monotonous convergence:

$$\begin{aligned}
\mathbb{E}Y &= \mathbb{E}[\lim_n \uparrow Y_n] \\
&= \lim_n \uparrow \mathbb{E}[Y_n] \\
&= \lim_n \uparrow \mathbb{E}\Big[\mathbb{E}\left[Y_n \mid \mathcal{G}\right]\Big] \\
&= \mathbb{E}\Big[\lim_n \uparrow \mathbb{E}\left[Y_n \mid \mathcal{G}\right]\Big] \\
&= \mathbb{E}Z .
\end{aligned}$$

If $A \in \mathcal{G}$, by monotonous convergence,

$$\lim_n \uparrow \mathbb{E}\left[\mathbb{I}_A Y_n\right] = \mathbb{E}\left[\mathbb{I}_A Y\right]$$

and so

$$\lim_n \uparrow \mathbb{E}\left[\mathbb{I}_A \mathbb{E}\left[Y_n \mid \mathcal{G}\right]\right] = \mathbb{E}\left[\mathbb{I}_A Y\right].$$

By monotonous convergence again:

$$\lim_n \uparrow \mathbb{E}\left[\mathbb{1}_A \lim_n \mathbb{E}\left[Y_n \mid \mathcal{G}\right]\right] = \mathbb{E}\left[\mathbb{1}_A Z\right]$$

□ □

8.4 Properties of (general) conditional expectation

Remark 8.2. In this Section $(\Omega, \mathcal{F}, P)$ is a probability space, $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$. Random variables $(X_n)_n, (Y_n)_n, X, Y, Z$ are meant to be integrables, and a.s. means P-a.s.

The easiest property is:
If $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] .$$

Exercise 8.2. Prove it.

If $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and X is $\mathcal{G}$-measurable then

$$X = \mathbb{E}[X \mid \mathcal{G}] \quad P\text{-a.s.}$$

Exercise 8.3. Prove it.

Using the definition of conditional expectation and monotone approximation by simple functions (see Section 3.2)), we obtain an alternative characterization of conditional expectation.

Let $X \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra, then for every $Y \in \mathcal{L}_1(\Omega, \mathcal{G}, P)$, such that $\mathbb{E}[|XY|] < \infty$

$$\mathbb{E}[XY] = \mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]] .$$

Exercise 8.4. Prove it.

We pocket the next proposition for future and frequent use. We could go ahead with listing many other useful properties of conditional expectation. They are best discovered and established when needed.

If $X, Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and Y is $\mathcal{G}$-measurable then

$$\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}] \quad P\text{-a.s.}$$

Proof. As $Y\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, it suffices to check that for every $B \in \mathcal{G}$,

$$\mathbb{E}[\mathbb{1}_B XY] = \mathbb{E}[\mathbb{1}_B (Y\mathbb{E}[X \mid \mathcal{G}])] .$$

But

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_B XY] &= \mathbb{E}[(\mathbb{1}_B Y)X] \\
&= \mathbb{E}[(\mathbb{1}_B Y)\mathbb{E}[X \mid \mathcal{G}]] \\
&= \mathbb{E}[\mathbb{1}_B (Y\mathbb{E}[X \mid \mathcal{G}])] .
\end{aligned}$$

□ □

8.5 Conditional convergence theorems

Limit theorems from integration theory (monotone convergence theorem, Fatou's Lemma, Dominated convergence theorem) can be adapted to the conditional expectation setting.

Theorem 8.4 (Monotone convergence). *Let the sequence $(X_n)_n$ of non-negative random variables converge monotonously to X ($X_n \uparrow X$ a.s.), with X integrable, then for every sequence of versions of conditional expectations:*

$$\lim_n \uparrow \mathbb{E}\left[X_n \mid \mathcal{G}\right] = \mathbb{E}\left[X \mid \mathcal{G}\right] \textit{ a.s.}$$

Proof. The sequence $X - X_n$ is non-negative and decreases to 0 a.s. It suffices to show that $\lim_n \downarrow \mathbb{E}\left[X - X_n \mid \mathcal{G}\right] = 0$ a.s. Note first that the sequence $\mathbb{E}\left[X - X_n \mid \mathcal{G}\right]$ converges a.s. toward a non-negative limit. We need to check that this limit is a.s. zero.

For $A \in \mathcal{G}$:

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_A \lim_n \mathbb{E}\left[X - X_n \mid \mathcal{G}\right]\right] &= \lim_n \mathbb{E}\left[\mathbb{I}_A \mathbb{E}\left[X - X_n \mid \mathcal{G}\right]\right] \\ &\quad \text{monotone convergence theorem} \\ &= \lim_n \mathbb{E}\left[\mathbb{I}_A \left(X_n - X\right)\right] \\ &\quad \text{monotone convergence theorem} \\ &= 0\,.\end{aligned}$$

□ □

Theorem 8.5 (Conditional Fatou's Lemma). *Let $(X_n)_n$ be a sequence of non-negative random variables, then*

$$\mathbb{E}\left[\liminf_n X_n \mid \mathcal{G}\right] \leq \liminf_n \mathbb{E}\left[X_n \mid \mathcal{G}\right] \quad \textit{a.s.}$$

As for the proof of Fatou's Lemma, the argument boils down to monotone convergence arguments.

Proof. Let $B \in \mathcal{G}$. Let $X = \liminf_n X_n$, X is a non-negative random variable. Let $Y = \liminf_n \mathbb{E}\left[X_n \mid \mathcal{G}\right]$, Y is a $\mathcal{G}$-measurable integrable random variable. The theorem compares $\mathbb{E}\left[X \mid \mathcal{G}\right]$ and Y.

Let $Z_k = \inf_{n \geq k} X_n$. Thus $\lim_k \uparrow Z_k = \liminf_n X_n = X$. According to Theorem 8.4,

$$\mathbb{E}\left[Z_k \mid \mathcal{G}\right] \uparrow_k \mathbb{E}\left[\liminf_n X_n \mid \mathcal{G}\right] \text{ a.s.}$$

For every $n \geq k$, $X_n \geq Z_k$ a.s. Hence by the comparison Theorem (Corollary 8.1)),

$$\forall n \geq k \quad \mathbb{E}\left[Z_k \mid \mathcal{G}\right] \leq \mathbb{E}\left[X_n \mid \mathcal{G}\right] \text{ a.s.}$$

as a countable union of P-negligible events is P-negligible. Hence for every k,

$$\mathbb{E}\left[Z_k \mid \mathcal{G}\right] \leq \liminf_n \mathbb{E}\left[X_n \mid \mathcal{G}\right] \quad \text{a.s.}$$

This entails

$$\lim_k \uparrow \mathbb{E}\left[Z_k \mid \mathcal{G}\right] \leq \liminf_n \mathbb{E}\left[X_n \mid \mathcal{G}\right] \quad \text{a.s.}$$

□ □

Dominated convergence

Let $V \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$. Let sequence $(X_n)_n$ satisfy $|X_n| \leq V$for every n and $X_n \to X$a.s., then for any sequence of versions of conditional expectations of $(X_n)_n$ and X

$$\mathbb{E}\left[X_n \mid \mathcal{G}\right] \to \mathbb{E}\left[X \mid \mathcal{G}\right] \quad \text{a.s.}$$

Proof. Let $Y_n = \inf_{m \geq n} X_m$ and $Z_n = \sup_{m \geq n} X_m$. Hence $-V \leq Y_n \leq Z_n \leq V$. As $Y_n \uparrow X$ and $Z_n \downarrow X$. By the conditional monotone convergence Theorem (Theorem 8.4)), $\mathbb{E}\left[Y_n \mid \mathcal{G}\right] \uparrow \mathbb{E}[X \mid \mathcal{G}]$ and $\mathbb{E}\left[Z_n \mid \mathcal{G}\right] \downarrow \mathbb{E}[X \mid \mathcal{G}]$p.s. Observe that for every n

$$\mathbb{E}\left[Y_n \mid \mathcal{G}\right] \leq \mathbb{E}\left[X_n \mid \mathcal{G}\right] \leq \mathbb{E}\left[Z_n \mid \mathcal{G}\right] \quad \text{a.s.}$$

□ □

Jensen's inequality also has a conditional version. The proof relies again on the variational representation of convex lower semi-comntinuous functions and on the monotonicity property of conditional expectation (Corollary 8.1)).

Jensen's inequality

If g is a lower semi-continuous convex function on $\mathbb{R}$, with $\mathbb{E}\left[|g(X)|\right] < \infty$ then

$$g\left(\mathbb{E}\left[X \mid \mathcal{G}\right]\right) \leq \mathbb{E}\left[g(X) \mid \mathcal{G}\right] \text{ a.s..}$$

Proof. A lower semi-continuous convex function is a countable supremum of affine functions: there exists a countable collection $(a_n, b_n)_n$ such that for every x:

$$g(x) = \sup_n \left[a_n x + b_n\right].$$

$$\begin{aligned} g\left(\mathbb{E}\left[X \mid \mathcal{G}\right]\right) &= \sup_n \left[a_n \mathbb{E}\left[X \mid \mathcal{G}\right] + b_n\right] \\ &= \sup_n \left[\mathbb{E}\left[a_n X + b_n \mid \mathcal{G}\right]\right] \\ &\leq \mathbb{E}\left[\sup_n \left(a_n X + b_n\right) \mid \mathcal{G}\right] P\text{-a.s.} \end{aligned}$$

□ □

Independence

When the conditioning σ-algebra $\mathcal{G}$ is atomic, if the conditioned random variable X is independent from the conditioning σ-algebra, it is obvious that the conditional expectation is an a.s. constant random variable which value equals $\mathbb{E}X$. This remains true in the general framework. It deserves a proof.

Proposition 8.4. *If X is independent from $\mathcal{G}$, then*

$$\mathbb{E}\left[X \mid \mathcal{G}\right] = \mathbb{E}\left[X\right].$$

Proof. Note that $\mathbb{E}[X]$ is $\mathcal{G}$-measurable. Let $B \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}[\mathbb{I}_B X] &= \mathbb{E}[\mathbb{I}_B]\,\mathbb{E}[X] \\ &\qquad \text{by independence} \\ &= \mathbb{E}[\mathbb{I}_B \times \mathbb{E}[X]]\,.\end{aligned}$$

Hence $\mathbb{E}[X] = \mathbb{E}[X \mid \mathcal{G}]$.

□ □

Proposition 8.4 can be generalized to a more general setting.
If sub-σ-algebra $\mathcal{H}$ is independent from $\sigma(\mathcal{G}, \sigma(X))$ then

$$\mathbb{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \mid \mathcal{G}] \quad \text{a.s.}$$

Proof. Recall that conditional expectation with respect to $\sigma(\mathcal{G}, \mathcal{H})$ can be characterized using a π-system containing Ω and generating $\sigma(\mathcal{G}, \mathcal{H})$, for example $\mathcal{G} \times \mathcal{H}$. Let $B \in \mathcal{G}$ and $C \in \mathcal{H}$,

$$\begin{aligned}\mathbb{E}[\mathbb{I}_B \mathbb{I}_C \mathbb{E}[X \mid \mathcal{G}]] &= \mathbb{E}[\mathbb{I}_B \mathbb{E}[X \mid \mathcal{G}]] \times \mathbb{E}[\mathbb{I}_C] \\ &\qquad C \text{ is independent from } \sigma(\mathcal{G}, \sigma(X)) \\ &= \mathbb{E}[\mathbb{I}_B X] \times \mathbb{E}[\mathbb{I}_C] \\ &= \mathbb{E}[\mathbb{I}_C \mathbb{I}_B X] \\ &\qquad C \text{ is independent from } \sigma(\mathcal{G}, \sigma(X))\,.\end{aligned}$$

□ □

8.6 Conditional probability distributions

Easy case: conditioning with respect to a discrete σ-algebra

We come back to the basic setting: $(\Omega, \mathcal{F}, P)$ refers to a probability space while $\mathcal{G} \subseteq \mathcal{F}$ denotes an ***atomic*** sub-σ-algebra generated by a countable partition $(A_n)_n$ of Ω.

Either from conditional expectations with respect to $\mathcal{G}$, or from conditional probabilities knowing the events A_n, we can define a N function of $\Omega \times \mathcal{F}$ per

$$N(\omega, B) = \mathbb{E}_P[\mathbb{I}_B \mid \mathcal{G}](\omega) = P\{B \mid A_n\} \text{ when } \omega \in A_n\,.$$

The N function has two remarkable properties:

i. For every $\omega \in \Omega$, $N(\omega, \cdot)$ defines a probability on $(\Omega, \mathcal{F})$.
ii. For every event $B \in \mathcal{F}$, the function $N(\cdot, B)$ is a $\mathcal{G}$-measurable function.

In this simple atomic setting, we observe that while it is intuitive to define conditional expectation starting from conditional probabilities, we can also proceed the other way around: we can build conditional probabilities starting from conditional expectations.

Impediments

In the general case, we attempt to construct conditional probabilities when the conditioning σ-algebra is not atomic.

For each $B \in \mathcal{F}$, we can rely on the existence of random variable $\sigma(X)$-measurable which is P-a.s. a version of the conditional expectation of $\mathbb{I}_B$ with respect to X. Indeed, for any kind of countable collection of events $(B_n)_n$ of $\mathcal{F}$, we can take for granted that there exists a collection of random variables which, almost surely, form a consistent collection of versions of the expectation of $(\mathbb{I}_{B_n})_n$ with respect to X. If $(B_n)_n$ is non-decreasing tending towards B, by Theorem 8.4), we are confident in the fact that the following holds

$$\lim_n \uparrow \mathbb{E}\left[\mathbb{I}_{B_n} \mid X\right] = \mathbb{E}\left[\mathbb{I}_B \mid X\right] \qquad \text{a.s.}$$

It is therefore tempting to define a conditional probability with respect to $\sigma(X)$ as a function

$$\begin{aligned} \Omega \times \mathcal{F} &\to [0,1] \\ (\omega, B) &\mapsto \mathbb{E}\left[\mathbb{I}_B \mid \sigma(X)\right](\omega) . \end{aligned}$$

Unfortunately, we cannnot guarantee that P-a.s., this object has the properties of a probability distribution $(\Omega, \mathcal{F})$. The problem does not arise from the diffuse nature of the distribution of X but from the size of $\mathcal{F}$. As $\mathcal{F}$ may not be countable, it is possible to build an uncountable non-decreasing sequence of events. Checking the a.s. monotonicity of the sequence of corresponding conditional probabilities looks beyond our reach (an uncountable union of P-negligible events is not necessarily P-negligible).

Fortunately, the situation is not desperate. In most settings envisioned in an introductory course on Probability, we can take the existence of condition probabilities for granted.

In Section 8.6), we first review the easy case, where we can define conditional probabilities that even have a density with respect to a reference measure. In Section 8.6) we shall see that if Ω is not too large, we can rely on the existence of conditional probabilities.

Joint density setting

If $\Omega = \mathbb{R}^k$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^k)$ and the probability distribution P is absolutely continuous with respect to Lebesgue measure (has a density denoted by p), defining conditional density with respect to coordinate projections is almost as simple as conditioning with respect to an atomic σ-algebra.

For the sake of simplicity, we stick to the case $k = 2$. A generic outcome is denoted by $\omega = (x, y)$ and the coordinated projections define two random variables $X(x, y) = x$ and $Y(x, y) = y$. We denote by p_X the ***marginal density*** of the distribution of X:

$$p_X(x) = \int_{\mathbb{R}} p(x, y) \mathrm{d}y.$$

And we agree on $D = \{x : p_X(x) > 0\}$. This is the support of the density p_X (beware, this may be different from the support of distribution $P \circ X^{-1}$).

Exercise 8.5. Check that p_X is the density of $P \circ X^{-1}$.

Having a density allows us to calculate conditional expectation and to define just as easily what we will call a conditional probability of Y knowing X.

Theorem 8.6 (Conditional density). *Let be X, Y be the projection coordinates on $\mathbb{R}^2$. Let P be an absolutely continuous distribution on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ with density $p(.,.)$ with respect to Lebesgue's measure. Let the first marginal density (density of $P \circ X^{-1}$ be denoted by p_X.*

The function

$$
\begin{aligned}
N: \quad & \mathbb{R}^2 \to [0, \infty) \\
& (x, y) \mapsto N(x, y) = \begin{cases} \frac{p(x,y)}{p_X(x)} & \text{if } p_X(x) > 0 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

satisfies the following properties.

i. For each x such that $p_X(x) > 0$, the set function $P_{\cdot \mid X=x}$ defined by

$$
\begin{aligned}
& \mathcal{B}(\mathbb{R}^2) \to [0, 1] \\
& \quad B \mapsto P_{\cdot \mid X=x}\{B\} = \int_{\mathbb{R}} \mathbb{I}_B(x, y) N(x, y) \mathrm{d}y
\end{aligned}
$$

is a probability measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. This probability distribution is supported by $\{x\} \times \mathbb{R}$.

ii. For every $B \in \mathcal{B}(\mathbb{R}^2)$, the function

$$
\omega \mapsto \int_{\mathbb{R}} \mathbb{I}_B(X(\omega), y) N(X(\omega), y) \mathrm{d}y = \mathbb{E}_{P_{\cdot \mid X=X(\omega)}} \mathbb{I}_B
$$

for $\Omega = \mathbb{R}^2$, is $\sigma(X)$-measurable and may be called a version of $\mathbb{E}[\mathbb{I}_B \mid \sigma(X)]$.

iii. For every $B \in \mathcal{B}(\mathbb{R}^2)$

$$
P(B) = \int \left(\int \mathbb{I}_B(s, y) N(s, y) \mathrm{d}y \right) p_X(s) \mathrm{d}s = \int P_{\cdot \mid X=s}(B) p_X(s) \mathrm{d}s.
$$

iv. For any P-integrable function f on $\mathbb{R}^2$, the random variable defined by applying

$$
x \mapsto \int_{\mathbb{R}} f(x, y) N(x, y) \mathrm{d}y
$$

to X is a version of the conditional expectation of $f(X, Y)$ with respect to $\sigma(X)$.

Remark 8.3. For each x such that $p_X(x) > 0$, $P_{\cdot \mid X=x}$ is a probability on $\mathbb{R}^2$. But this probability measure is supported by $\{x\} \times \mathbb{R}$, it is the product of the Dirac mass in $\{x\}$ times the probability distribution on $\mathbb{R}$ defined by the density $N(x, \cdot)$. This is why $N(x, \cdot)$ is often called the conditional density of Y given $X = x$, and the distribution over $\mathbb{R}$ defined by this density is often called the conditional distribution of Y given X.

Exercise 8.6. Is $N(x, y)$ a probability density? If yes, with respect to which σ-finite measure?

The proof of Theorem 8.6) consists of milking the Tonelli-Fubini Theorem.

Proof. Proof of (i). Let us agree on notation:

$$P_x\{B\} = \int_{\mathbb{R}} \mathbb{1}_B(x,y)N(x,y)\mathrm{d}y.$$

The fact that the P_x is $[0,1]$-valued is immediate. Same for the fact that $P_x(\{x\} \times \{\emptyset\}) = 0$ and $P_x(\{x\} \times \{\mathbb{R}\}) = 1$. The same applies to additivity.

It remains to check that if (B_n) is a non-decreasing sequence of Borelians from $\mathbb{R}^2$ that tends to to a limit B then

$$\lim_n \uparrow P_x(B_n) = P_x(B)\,.$$

This is an immediate consequence of the monotonous convergence theorem, for each (x', y') $\lim_n \uparrow \mathbb{1}_{B_n}(x',y')N(x',y') = \mathbb{1}_B(x',y')N(x',y')$.

Proof of ii) As the function $(x,y) \mapsto p(x,y)\mathbb{1}_B(x,y)$ is $\mathcal{B}(\mathbb{R}^2)$-measurable and integrable, by the Tonelli-Fubini Theorem,

$$x \mapsto \int_B p(x,y)\mathbb{1}_B(x,y)\mathrm{d}y$$

is defined almost everywhere and Borel-measurable.

Proof of iii) This is also an immediate consequence of the Tonelli-Fubini Theorem.

Proof of iv), It follows from i.), using the usual approximation by simple functions argument.

□ □

Exercise 8.7. We consider the uniform law on the surface of $\mathbb{R}^2$ defined by $0 \leq x \leq y \leq y \leq 1$. Give the attached density $p()$, the marginal density p_X and the kernel $N(,)$.

Regular conditional probabilities, kernels

We will outline some results that allow us to work within a more general framework. We introduce two new notions.

Conditional probability kernel

Let $(\Omega, \mathcal{F})$ be a measurable space, and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$.

We call ***conditional probability kernel with respect to*** $\mathcal{G}$ a function $N : \Omega \times \mathcal{F} \to \mathbb{R}_+$ that satisfies:

i. For any $\omega \in \Omega$, $N(\omega, \cdot)$ defines a probability on $(\Omega, \mathcal{F})$.
ii. For any $A \in \mathcal{F}$, $N(\cdot, A)$ is $\mathcal{G}$-measurable

If the measurable space is endowed with a probability distribution P, we are interested in conditional probability kernels with respect to $\mathcal{G}$ that are compliant with P. We call them *regular conditional probability kernels*.

Regular conditional probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. A kernel $N : \Omega \times \mathcal{F} \to \mathbb{R}_+$ is a ***regular conditional probability*** with respect $\mathcal{G}$ if and only if

i. For any $B \in \mathcal{F}$, $\omega \mapsto N(\omega, B)$ is a version of the conditional expectation of $\mathbb{I}_B$ knowing $\mathcal{G}$ ($N(\cdot, B)$ is therefore $\mathcal{G}$-measurable):

$$N(\cdot, B) = \mathbb{E}[\mathbb{I}_B \mid \mathcal{G}] \quad P - \text{a.s.}$$

ii. For P-almost all $\omega \in \Omega$, $B \mapsto N(\omega, B)$ defines a probability on $(\Omega, \mathcal{F})$.

A regular conditional probability (whenever it exists) is defined from versions of conditional expectations. Conversely, a regular conditional probability provides us with a way to to compute conditional expectations.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. Let N be a probability kernel on $(\Omega, \mathcal{F})$ with respect to $\mathcal{G}$.

The following properties are equivalent

1. $N(\cdot, \cdot)$ defines a regular conditional probability kernel with respect to $\mathcal{G}$ for $(\Omega, \mathcal{G}, P)$.
2. P-almost surely, for any P-integrable function f on $(\Omega, \mathcal{F})$:

$$\mathbb{E}\left[f \mid \mathcal{G}\right](\omega) = \mathbb{E}_{N(\omega,\cdot)}[f] \, .$$

3. For any P-integrable random variable X on $(\Omega, \mathcal{F})$

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\mathbb{E}_{N(\omega,\cdot)}[X]\right] \, .$$

Remark 8.4. The proof of 1) ⇒ 2) relies on the usual machinery: approximation of positive integrable functions by an increasing sequences of simple functions, monotone convergence of expectation and conditional expectation.

2) ⇒ 3) is trivial.

3) ⇒ 1) is more interesting.

Existence of regular conditional probability distributions when $\Omega = \mathbb{R}$

We shall check the existence of conditional probabilities in at least one non-trivial case.

Let P be a probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\mathcal{G} \subseteq \mathcal{B}(\mathbb{R})$, then there exists a regular conditional probability kernel with respect to $\mathcal{G}$.

We take advantage of the fact that $\mathcal{B}(\mathbb{R})$ is countably generated.

Proof. Let $\mathcal{C}$ be the set formed by half-lines with rational endpoint, the empty set, and $\mathbb{R}$:

$$\mathcal{C} = \left\{(-\infty, q] : q \in \mathbb{Q}\right\} \cup \{\emptyset, \mathbb{R}\} \, .$$

This countable collection of half-lines is a π-system (See Section 2.6)) that generates $\mathcal{B}(\mathbb{R})$.

For $q < q' \in \mathbb{Q}$, we can choose versions of Y_q and $Y_{q'}$ of the conditional expectations of $\mathbb{I}_{(-\infty,q]}$ and $\mathbb{I}_{(-\infty,q']}$ such that

$$Y_q < Y_{q'} \qquad P\text{-a.s.}$$

Observe that $Y_{q'} - Y_q$ is also a version of the conditional expectation of $\mathbb{I}_{(q,q']}$.

A countable union of P-negligible events is P-negligible, so, as $\mathbb{Q}^2$ is countable, we can choose versions $\left(Y_q\right)_{q \in \mathbb{Q}}$ of the conditional expectations of $\mathbb{I}_{(-\infty,q]}$ such that

$$P\text{-a.s.} \qquad \forall q, q' \in \mathbb{Q}, \quad q < q' \Rightarrow Y_q < Y_{q'},$$

Let Ω_0 be the P-almost sure event on which all $Y_q, q \in \mathbb{Q}$ satisfy the good properties.

For each $x \in \mathbb{R}$, we can define Z_x for each $\omega \in \mathbb{R}$ by

$$Z_x(\omega) = \inf\left\{Y_q(\omega) : q \in \mathbb{Q}, x < q\right\}$$

On Ω_0, the function $x \mapsto Z_x(\omega)$ is increasing, it has a limit on the left at each point and it is right-continuous. The function $x \mapsto Z_x(\omega)$ tends to 0 when x tends to $-\infty$, to 1 when x tends towards $+\infty$. In words, on Ω_0, $x \mapsto Z_x(\omega)$ is a cumulative distribution function, it defines so (uniquely) a unique probability measure on $\mathbb{R}$. We will denote it by $\nu(\omega, .)$.

In addition, for each x, Z_x is defined as a countable infimum of $\mathcal{G}$-measurable random variables, Z_x is therefore $\mathcal{G}$-measurable.

It remains to check that for every $B \in \mathcal{F}$, $\omega \mapsto \nu(\omega, B)$ for $\omega \in \Omega_0$, 0 elsewhere, defines a version of the conditional expectation of $\mathbb{1}_B$ with respect to $\mathcal{G}$.

This property is satisfied for $B \in \mathcal{C}$.

Let us call $\mathcal{D}$ the set of all the events for which $\omega \mapsto \nu(\omega, B)$ (on Ω_0, 0 elsewhere) defines a version of the conditional expectation of $\mathbb{1}_B$ with respect to $\mathcal{G}$. We shall show that $\mathcal{D}$ is a λ-system, that is

i. $\mathcal{D}$ contains $\emptyset$ and $\mathbb{R} = \Omega$.
ii. If B, B' belong to $\mathcal{D}$, and $B \subseteq B'$ then $B' \setminus B \in \mathcal{D}$.
iii. If $(B_n)_n$ is a growing sequence of events from $\mathcal{D}$, limit B then $B \in \mathcal{D}$.

Clause i.) is guaranteed by construction.

Clause ii.) If $B' \subseteq B$ belong to $\mathcal{D}$, then by linearity of conditional expectations, if $\mathbb{E}\left[\mathbb{1}_{B' \setminus B} \mid \mathcal{G}\right]$ is a version of the conditional expectation of $\mathbb{1}_{B' \setminus B}$ with respect to $\mathcal{G}$, on an almost-sure event $\Omega_1 \subseteq \Omega_0$:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{B' \setminus B} \mid \mathcal{G}\right] &= \mathbb{E}\left[\mathbb{1}_{B'} - \mathbb{1}_B \mid \mathcal{G}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{B'} \mid \mathcal{G}\right] - \mathbb{E}\left[\mathbb{1}_B \mid \mathcal{G}\right] \\
&= \nu(\omega, B') - \nu(\omega, B) \\
&= \nu(\omega, B' \setminus B)\,.
\end{aligned}$$

Clause iii.). If $(B_n)_n$ is a non-decreasing sequence of events from $\mathcal{D}$, with $B_n \uparrow B$, if $\mathbb{E}\left[\mathbb{1}_B \mid \mathcal{G}\right]$ is a version of the conditional expectation of $\mathbb{1}_B$ with respect to $\mathcal{G}$, on an event $\Omega_1 \subseteq \Omega_0$ with probability 1:

$$\mathbb{E}\left[\mathbb{1}_B \mid \mathcal{G}\right] = \lim_n \mathbb{E}\left[\mathbb{1}_{B_n} \mid \mathcal{G}\right] = \lim_n \nu(\omega, B_n) = \nu(\omega, B)\,.$$

So $B \in \mathcal{D}$.

The Monotone class Theorem (Section 2.6)) tells us that $\mathcal{F} \subseteq \mathcal{D}$.

□ □

Working harder would allow us to show that the existence of regular conditional probabilities is guaranteed as soon as Ω can be endowed with a complete and separable metric space structure and that the σ-algebra $\mathcal{F}$ is the Borelian σ-algebra induced by this metric.

We often define a probability distribution starting from a marginal distribution and a kernel.

Let $(\Omega, \mathcal{F})$ be a measurable space, X a random variable from $(\Omega, \mathcal{F})$, and N a conditional probability kernel with respect to $\sigma(X)$. Let P_X be a probability measure on $(\Omega\sigma(X))$.

Then there exists a unique probability measure P on $(\Omega, \mathcal{F})$ such that $P_X = P \circ X^{-1}$ and N is a regualr conditional probability kernel with respect to $\sigma(X)$, we have for every $B \in \mathcal{F}$:

$$P(B) = \int_{X(\Omega)} N(x, B) \mathrm{d}P_x(x)$$

The following theorem guarantees the existence of a regular conditional probability in all the scenarios we are interested in.

8.7 Efron-Stein-Steele inequalities

In this section, $X_1, \ldots, X_n$ denote independent random variables on some probability space with values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$, and f denote a measurable function from $\mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ to $\mathbb{R}$. Let $Z = f(X_1, \ldots, X_n)$. The random variable Z is a general function of independent random variables. We assume Z is integrable.

If we had $Z = \sum_{i=1}^n X_i$, we could write

$$\operatorname{var}(Z) = \sum_{i=1}^n \operatorname{var}(X_i) = \sum_{i=1}^n \mathbb{E}\Big[\operatorname{var}(Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n)\Big]$$

even though the last expression looks pedantic. The aim of this section is to show that even if f is not as simple as the sum of its arguments, the last expression can still serve as an upper bound on the variance.

It is a natural idea to bound the variance of such a general function by expressing $Z - \mathbb{E}Z$ as a sum of differences and to use the orthogonality of these differences.

More precisely, if we denote by $\mathbb{E}_i$ the conditional expectation operator, conditioned on $(X_1, \ldots, X_i)$, and use the convention $\mathbb{E}_0 = \mathbb{E}$, then we may define

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$$

for every $i = 1, \ldots, n$.

Exercise 8.8. Check that $\mathbb{E}\Delta_i = 0$ and that for $j > i$, $\mathbb{E}_i \Delta_j = 0$ a.s.

Starting from the decomposition

$$Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$$

one has

$$\operatorname{var}(Z) = \mathbb{E}\left[\left(\sum_{i=1}^n \Delta_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}\left[\Delta_i^2\right] + 2\sum_{j>i} \mathbb{E}\left[\Delta_i \Delta_j\right] .$$

Now if $j > i$, $\mathbb{E}_i \Delta_j = 0$ implies that

$$\mathbb{E}_i\left[\Delta_j \Delta_i\right] = \Delta_i \mathbb{E}_i \Delta_j = 0 ,$$

and, a fortiori, $\mathbb{E}\left[\Delta_j \Delta_i\right] = 0$. Thus, we obtain the following analog of the additivity formula of the variance:

$$\operatorname{var}(Z) = \mathbb{E}\left[\left(\sum_{i=1}^n \Delta_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}\left[\Delta_i^2\right] .$$

Observe that up to now, we have not made any use of the fact that Z is a function of independent variables $X_1, \ldots, X_n$.

Independence may be used as in the following argument: for any integrable function $Z = f(X_1, \ldots, X_n)$ one may write, by the Tonelli-Fubini theorem,

$$\mathbb{E}_i Z = \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_i, x_{i+1}, \ldots, x_n)\, d\mu_{i+1}(x_{i+1}) \ldots d\mu_n(x_n) ,$$

where, for every $j = 1, \dots, n$, μ_j denotes the probability distribution of X_j. Also, if we denote by $\mathbb{E}^{(i)}$ the conditional expectation operator conditioned on $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$, we have

$$\mathbb{E}^{(i)} Z = \int_{\mathcal{X}} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n)\, d\mu_i(x_i) \ .$$

Then, again by the Tonelli-Fubini theorem,

$$\mathbb{E}_i \left[\mathbb{E}^{(i)} Z\right] = \mathbb{E}_{i-1} Z.(\#eq: efundind) \tag{8.1}$$

This observation is the key in the proof of the main result of this section which we state next:

Theorem 8.7 (Efron-Stein-Steele's inequalities). *Let $X_1, \dots, X_n$ be independent random variables and let $Z = f(X)$ be a square-integrable function of $X = (X_1, \dots, X_n)$. Then*

$$\operatorname{var}(Z) \le \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] = v \ .$$

Moreover if $X_1', \dots, X_n'$ are independent copies of $X_1, \dots, X_n$ and if we define, for every $i = 1, \dots, n$,

$$Z_i' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n) \ ,$$

then

$$v = \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_+^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_-^2\right]$$

where $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$ denote the positive and negative parts of a real number x. Also,

$$v = \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i)^2\right] \ ,$$

where the infimum is taken over the class of all $X^{(i)}$-measurable and square-integrable variables Z_i, $i = 1, \dots, n$.

Proof. We begin with the proof of the first statement. Note that, using @ref(eq:efundind), we may write

$$\Delta_i = \mathbb{E}_i \left[Z - \mathbb{E}^{(i)} Z\right] .$$

By the conditional Jensen inequality,

$$\Delta_i^2 \le \mathbb{E}_i \left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] \ .$$

Using $\operatorname{var}(Z) = \sum_{i=1}^{n} \mathbb{E}\left[\Delta_i^2\right]$, we obtain the desired inequality. To prove the identities for v, denote by $\operatorname{var}^{(i)}$ the conditional variance operator conditioned on $X^{(i)}$. Then we may write v as

$$v = \sum_{i=1}^{n} \mathbb{E}\left[\operatorname{var}^{(i)}(Z)\right] .$$

Now note that one may simply use (conditionally) the elementary fact that if X and Y are independent and identically distributed real-valued random variables, then $\operatorname{var}(X) = (1/2)\mathbb{E}[(X - Y)^2]$. Since conditionally on $X^{(i)}$, Z_i' is an independent copy of Z, we may write

$$\operatorname{var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}\left[(Z - Z_i')^2\right] = \mathbb{E}^{(i)}\left[(Z - Z_i')_+^2\right] = \mathbb{E}^{(i)}\left[(Z - Z_i')_-^2\right]$$

where we used the fact that the conditional distributions of Z and Z'_i are identical. The last identity is obtained by recalling that, for any real-valued random variable X, $\mathrm{var}(X) = \inf_{a\in\mathbb{R}} \mathbb{E}[(X-a)^2]$. Using this fact conditionally, we have, for every $i = 1, \ldots, n$,

$$\mathrm{var}^{(i)}(Z) = \inf_{Z_i} \mathbb{E}^{(i)}\left[(Z - Z_i)^2\right] .$$

Note that this infimum is achieved whenever $Z_i = \mathbb{E}^{(i)} Z$.

□ □

Observe that in the case when $Z = \sum_{i=1}^n X_i$ is a sum of independent random variables (with finite variance) then the Efron-Stein-Steele inequality becomes an equality. Thus, the bound in the Efron-Stein-Steele inequality is, in a sense, not improvable.

8.8 Bounded differences inequalities

In this section we combine Hoeffding's inequality and conditioning to establish the so-called ***Bounded differences inequality*** (also known as McDiarmid's inequality). This inequality is a first example of the ***concentration of measure phenomenon***. This phenomenon is best portrayed by the following say:

> A function of many independent random variables that does not depend too much on any of them is concentrated around its mean or median value.

Bounded differences inequalities

Let $X_1, \ldots, X_n$ be independent random variables on some probability space with values in $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$. Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n \to \mathbb{R}$ be a measurable function such that there exists non-negative constants $c_1, \ldots, c_n$ satisfying $\forall x_1, \ldots, x_n \in \prod_{i=1}^n \mathcal{X}_i$, $\forall y_1, \ldots, y_n \in \prod_{i=1}^n \mathcal{X}_i$,

$$\Big| f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n) \Big| \leq \sum_{i=1}^n c_i \mathbb{I}_{x_i \neq y_i} .$$

Let $Z = f(X_1, \ldots, X_n)$ and $v = \sum_{i=1}^n \frac{c_i^2}{4}$. Then

$$\mathrm{var}(Z) \leq v ,$$

$$\log \mathbb{E} \mathrm{e}^{\lambda(Z - \mathbb{E}Z)} \leq \frac{\lambda^2 v}{2}$$

and

$$P\Big\{ Z \geq \mathbb{E}Z + t \Big\} \leq \mathrm{e}^{-\frac{t^2}{2v}} .$$

Proof. The variance bound is an immediate consequence of the Efron-Stein-Steele inequalities.

The tail bound follows from the upper bound on the logarithmic moment generating function by Cramer-Chernoff bounding.

Let us now check the upper-bound on the logarithmic moment generating function.

We proceed by inudction on the number of arguments n.

If $n = 1$, the upper-bound on the logarithmic moment generating function is just Hoeffing's Lemma (see Section 9.7)).

Assume the upper-bound is valid up to $n - 1$.

We adopt the same notation as in Section 8.7).

$$\begin{aligned}\mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)} &= \mathbb{E}\left[\mathbb{E}_{n-1}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)}\right] \\ &= \mathbb{E}\left[\mathbb{E}_{n-1}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}_{n-1}Z)}\right] \times \mathrm{e}^{\lambda(\mathbb{E}_{n-1}Z-\mathbb{E}Z)}\right].\end{aligned}$$

Now,

$$\mathbb{E}_{n-1}Z = \int_{\mathcal{X}_n} f(x_1,\ldots,x_{n-1},u)\mathrm{d}P_{X_n}(u) \qquad \text{a.s.}$$

and

$$\begin{aligned}&\mathbb{E}_{n-1}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}_{n-1}Z)}\right] \\ &= \int_{\mathcal{X}_n} \exp\left(\lambda\int_{\mathcal{X}_n} f(x_1,\ldots,x_{n-1},v) - f(x_1,\ldots,x_{n-1},u)\mathrm{d}P_{X_n}(u)\right)\mathrm{d}P_{X_n}(v).\end{aligned}$$

For every $x_1,\ldots,x_{n-1} \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_{n-1}$, for every $v, v' \in \mathcal{X}_n$,

$$\begin{aligned}&\Big|\int_{\mathcal{X}_n} f(x_1,\ldots,x_{n-1},v) - f(x_1,\ldots,x_{n-1},u)\mathrm{d}P_{X_n}(u) \\ &- \int_{\mathcal{X}_n} f(x_1,\ldots,x_{n-1},v') - f(x_1,\ldots,x_{n-1},u)\mathrm{d}P_{X_n}(u)\Big| \leq c_n\end{aligned}$$

hence, by Hoeffding's Lemma

$$\mathbb{E}_{n-1}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}_{n-1}Z)}\right] \leq \mathrm{e}^{\frac{\lambda^2 c_n^2}{8}}.$$

$$\mathbb{E}\mathrm{e}^{\lambda(Z-\mathbb{E}Z)} \leq \mathbb{E}\left[\mathrm{e}^{\lambda(\mathbb{E}_{n-1}Z-\mathbb{E}Z)}\right] \times \mathrm{e}^{\frac{\lambda^2 c_n^2}{8}}.$$

But, if $X_1 = x_1, \ldots X_{n-1} = x_{n-1}$,

$$\mathrm{e}^{\lambda(\mathbb{E}_{n-1}Z-\mathbb{E}Z)} = \int_{\mathcal{X}_n} f(x_1,\ldots,x_{n-1},v)\mathrm{d}P_{X_n}(v) - \mathbb{E}Z,$$

it is a function of $n-1$ independent random variables that satisfies the bounded differences conditions with constants $c_1,\ldots,c_{n-1}$. By the induction hypothesis:

$$\mathbb{E}\left[\mathrm{e}^{\lambda(\mathbb{E}_{n-1}Z-\mathbb{E}Z)}\right] \leq \mathrm{e}^{\frac{\lambda^2}{2}\sum_{i=1}^{n-1}\frac{c_i^2}{4}}.$$

□ □

8.9 Bibliographic remarks

Conditional expectations can be constructed from the Radon-Nikodym Theorem, see (Dudley, 2002).

It is also possible to prove the Radon-Nikodym Theorem starting from the construction of conditional expectation in $\mathcal{L}_2$, see (Williams, 1991).

The Section on Efron-Stein-Steele's inequalities is from (Boucheron, Lugosi, & Massart, 2013)

Bounded difference inequality is due to C. McDiarmid. It became popular in (Theoretical) computer science during the 1990's. See (McDiarmid, 1998)

Chapter 9

Convergences I : almost sure, L_2, L_1, in Probability

9.1 Motivations

We need to put topological structures in the world of random variables living on some probability space. As random variables are (measurable) functions, we shall borrow and adapt the notions used in Analysis: pointwise convergence (Section 9.2)), convergence in $L_p, 1 \leq p < \infty$ (Section 9.3)).

Finally, we define and investigate ***convergence in probability***. This notion weakens both L_p and almost sure (pointwise) convergence. Just as L_p convergences, it can be metrized.

Convergence in probability and almost sure convergence are illustrated by weak and strong law of large numbers (Sections Section 9.5 and Section 9.6). Laws of large numbers assert that empirical means converge towards expectations (under mild conditions), they are the workhorses of statistical learning theory.

In Section 9.7), we look at non-asymptotic counterparts of the weak law of large numbers. We establish exponential tail bounds for sums of independent random variables (under stringent integrability assumptions).

9.2 Almost sure convergence

The notion of almost sure convergence mirrors the notion of pointwise convergence in probabilistic settings.

Recall that a sequence of real-valued functions $(f_n)_n$ mapping some space Ω to $\mathbb{R}$ ***converges pointwise*** to $f : \Omega \to \mathbb{R}$, if for each $\omega \in \Omega$, $f_n(\omega) \to f(\omega)$. There is no uniformity condition.

In the next definition, we assume that random variables are real-valued. The definition is easily extended to multivariate settings.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, a sequence $(X_n)_n$ of random variables converges ***almost surely*** (a.s.) towards a random variable X if the event

$$E = \left\{\omega : \lim_n X_n(\omega) = X(\omega)\right\}$$

has P-probability 1.

Almost sure convergence, is (just) pointwise convergence with probability 1. Almost sure convergence is not tied to integrability. Note that all random variables involved in the above statements live on the same probability space. We may wonder whether we can design a metric for almost-sure convergence? The answer is no, as for pointwise convergence, in general.

9.3 Convergence in L_p

In this section, we consider random variables that satisfy integrability assumptions. The scope of L_p convergences is narrower than the scope of L_p convergences.

We already introduced L_p convergences in Lesson Chapter 3. We recall it for the sake of readibility.

For $p \in [1, \infty)$, L_p is the set of random variables over $(\Omega, \mathcal{F}, P)$ that satisfy $\mathbb{E}|X|^p < \infty$. The p-pseudo-norm is defined by $\|X\|_p = \left(\mathbb{E}|X|^p\right)^{1/p}$. Convergence in L_p means convergence for this pseudo-norm.

Recall that L_p spaces are nested (by Holder's inequality) and complete.

Convergence in $L_q, q \geq 1$ implies convergence in $L_p, 1 \leq p \leq q$.

Almost sure convergence is not tied to integrability. We cannot ask whether almost sure convergence implies L_p convergence. But we can ask whether L_p convergence implies almost sure convergence. The next statement is a by-product of the proof of the completeness of L_p spaces, see Section 3.11).

Convergence in L_p implies almost sure convergence ***along a subsequence***.

A counter-example given in Section 3.11) shows that convergence in L_p does not imply almost-sure convergence.

9.4 Convergence in probability

If we denote by $L_0 = L_0(\Omega, \mathcal{F}, P)$ the set of real-valued random variables, the notion of convergence in probability is relevant to all sequences in L_0 like almost sure convergence. And like convergence in $L_p, p \geq 1$, convergence in probability can be metrized.

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

A sequence $(X_n)_n$ of random variables converges in probability towards a random variable X if for any $\epsilon > 0$

$$\lim_n P\{|X_n - X| \geq \epsilon\} = 0\,.$$

Convergence in $L_p, p \geq 1$ implies convergence in probability.

This is an immediate consequence of Markov's inequality.

The sequence $(X_n)_n$ converges in probability towards X iff

$$\lim_n \mathbb{E}\Big[1 \wedge |X_n - X|\Big] = 0$$

Proof. Assuming convergence in probability,

$$\begin{aligned}\mathbb{E}\Big[1 \wedge |X_n - X|\Big] &\leq \mathbb{E}\Big[(1 \wedge |X_n - X|)\mathbb{I}_{|X-X_n|\geq\epsilon}\Big] + \mathbb{E}\Big[(1 \wedge |X_n - X|)\mathbb{I}_{|X-X_n|<\epsilon}\Big] \\ &\leq P\Big\{|X - X_n| \geq \epsilon\Big\} + \epsilon\end{aligned}$$

the limit of the right-hand side is not larger than ϵ. As we can take ϵ arbitrarily small, this entails that the limit of the left-hand side is zero.

Conversely, for all $0 < \epsilon < 1$

$$P\Big\{|X - X_n| \geq \epsilon\Big\} \quad \leq \tfrac{1}{\epsilon}\mathbb{E}\Big[1 \wedge |X - X_n|\Big]\,.$$

Hence $\lim_n \mathbb{E}\Big[1 \wedge |X_n - X|\Big] = 0$ entails $\lim_n P\{|X - X_n| \geq \epsilon\} = 0$. As this holds for all $\epsilon > 0$, $\lim_n \mathbb{E}\Big[1 \wedge |X_n - X|\Big] = 0$ entails convergence in Probability. □

Almost sure convergence implies convergence in probability.

Proof. Assume $X_n \to X$ a.s., that is $|X_n - X| \to 0$. Then by dominated convergence,

$$\lim_n \mathbb{E}\Big[|X_n - X| \wedge 1\Big] = 0$$

which entails convergence in probability of $(X_n)_n$ towards X. □

Now, we come to a metric which fits perfectly with convergence in probability.

Definition 9.1 (Ky-Fan distance)**.** The Ky-Fan distance is defined as

$$\mathrm{d_{KF}}(X, Y) = \inf_{\epsilon \geq 0} P\Big\{|X - Y| > \epsilon\Big\} \leq \epsilon\,.$$

Note that we have to check that $\mathrm{d_{KF}}$ is indeed a distance. This is the content of Proposition @ref(prp:kyfanprop) below.

In the definition of the Ky-Fan distance, the infimum is attained.

Proof. Let $a > \mathrm{d_{KF}}(X, Y)$ the event $A_a = \Big\{|X - Y| > a\Big\}$ has probability smaller than ϵ. And if $\epsilon < a < b$, $A_b \subseteq A_a$. By monotone converence, $P\Big(\cap_n A_{\epsilon+1/n}\Big) = \lim_n \uparrow P\Big(A_{\epsilon+1/n}\Big) = \epsilon$. □

Proposition 9.1. *Ky-Fan distance satisfies:*

1. $\mathrm{d_{KF}}(X, Y) = 0 \Rightarrow X = Y \qquad$ *a.s.*
2. $\mathrm{d_{KF}}(X, Y) = \mathrm{d_{KF}}(Y, X)$
3. $\mathrm{d_{KF}}(X, Z) \leq \mathrm{d_{KF}}(X, Y) + \mathrm{d_{KF}}(Y, Z)$

Proof. We check that $\mathrm{d_{KF}}$ satisfies the triangle inequality. There exists two events B and C with respective probabilities $\mathrm{d_{KF}}(X, Y)$ and $\mathrm{d_{KF}}(Y, Z)$ such that

$$|X(\omega) - Y(\omega)| \leq \mathrm{d_{KF}}(X, Y) \qquad \text{on } B^c$$

and

$$|Z(\omega) - Y(\omega)| \leq \mathrm{d_{KF}}(Z, Y) \qquad \text{on } C^c\,.$$

On $B^c \cap C^c$, by the triangle inequality on $\mathbb{R}$:

$$|X(\omega) - Z(\omega)| \leq \mathrm{d_{KF}}(X, Y) + \mathrm{d_{KF}}(Y, Z)\,.$$

We conclude by observing

$$\begin{aligned} P\Big(|X(\omega) - Z(\omega)| > \mathrm{d_{KF}}(X, Y) + \mathrm{d_{KF}}(Y, Z)\Big) & \leq P\Big((B^c \cap C^c)^c\Big) \\ & = P(B \cup C) \\ & \leq P(B) + P(C) \\ & = \mathrm{d_{KF}}(X, Y) + \mathrm{d_{KF}}(Y, Z)\,. \end{aligned}$$

□

The two statements are equivalent:

1. $(X_n)_n$ converges in probability towards X
2. $\mathrm{d}_{\mathrm{KF}}(X_n, X)$ tends to 0 as n tends to infinity.

Exercise 9.1. Check the proposition.

We leave the following questions as exercises:

- Is $\mathcal{L}_0(\Omega, \mathcal{F}, P)$ complete under the Ky-Fan metric?
- Does convergence in probability imply almost sure convergence?
- Does convergence in probability imply convergence in $L_p, p \geq 1$?

Finally, we state a more gemeral definition of convergence in probability. The notion can be tailored to random variables that map some universe to some metric space. The connections with almost-sure convergence and L_p convergences remain unchanged.

Definition 9.2 (Convergence in probability, multivariate setting). A sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^k$-valued random variables living on the same probability space $(\Omega, \mathcal{F}, P)$ converges in probability (in $\mathbb{P}$-probability) towards a $\mathbb{R}^k$-valued random variable X iff for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\{\|X_n - X\| > \epsilon\} = 0 \,.$$

9.5 Weak law of large numbers

The weak and the strong law of large numbers are concerned with the convergence of empirical means of independent, identically distributed, integrable random variables towards their common expectation.

Theorem 9.1 (Weak law of large numbers). *If $X_1, \ldots, X_n, \ldots$ are independently, identically distributed, integrable $\mathbb{R}^k$-valued random variables over $(\Omega, \mathcal{F}, P)$ with expectation μ then the sequence $(\overline{X}_n)$ defined by $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges in P-probability towards μ.*

Proof. Assume first that $\mathbb{E}\Big[\Big(X_i - \mu\Big)^2\Big] = \sigma^2 < \infty$. Then, for all $\epsilon > 0$, by the Markov-Chebychev inequality:

$$\begin{aligned} P\Big\{\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i - \mu\Big| > \epsilon\Big\} & \leq \frac{\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mu\Big|^2}{\epsilon^2} \\ & = \frac{\mathbb{E}\Big[\Big(X_i - \mu\Big)^2\Big]}{n\epsilon^2} \\ & = \frac{\sigma^2}{n\epsilon^2} \end{aligned}$$

because the variance of a sum of independent random variables equals the sum of the variances of the summands.

The right-hand side converges to 0 for all $\epsilon > 0$. The WLLN holds for square-integrable random variables.

Let us turn to the general case. Without loss of generality, assume all X_n are centered. Let $\tau > 0$ be a truncation threshold (which value will be tuned later). For each $i \in \mathbb{N}$, X_i is decomposed into a sum:

$$X_i = X_i^\tau + Y_i^\tau$$

with $X_i^\tau = \mathbb{1}_{|X_i|\leq \tau} X_i$ and $Y_i^\tau = \mathbb{1}_{|X_i|>\tau} X_i$. For every $\epsilon > 0$,

$$\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i\Big| > \epsilon\Big\} \subseteq \Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i^\tau\Big| > \frac{\epsilon}{2}\Big\} \cup \Big\{\Big|\frac{1}{n}\sum_{i=1}^n Y_i^\tau\Big| > \frac{\epsilon}{2}\Big\}.$$

Invoking the union bound, Markov's inequality (twice), the boundedness of the variances of the X_i^τ's leads to:

$$\begin{aligned} P\Big\{\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i - \mu\Big| > \epsilon\Big\} & \leq P\Big\{\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i^\tau\Big| > \tfrac{\epsilon}{2}\Big\} + P\Big\{\Big|\tfrac{1}{n}\textstyle\sum_{i=1}^n Y_i^\tau\Big| > \tfrac{\epsilon}{2}\Big\} \\ & \leq 4\frac{\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^n X_i^\tau\Big|^2}{\epsilon^2} + 2\frac{\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^n Y_i^\tau\Big|}{\epsilon} \\ & \leq \tfrac{4\tau^2}{n\epsilon^2} + 2\frac{\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^n Y_i^\tau\Big|}{\epsilon} \\ & \leq \tfrac{4\tau^2}{n\epsilon^2} + 2\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\mathbb{E}\Big|Y_i^\tau\Big|}{\epsilon} \\ & \leq \tfrac{4\tau^2}{n\epsilon^2} + 2\frac{\mathbb{E}\Big|Y_1^\tau\Big|}{\epsilon}. \end{aligned}$$

Taking n to infinity leads to

$$\limsup_n P\Big\{\Big|\frac{1}{n}\sum_{i=1}^n X_i - \mu\Big| > \epsilon\Big\} \leq 2\frac{\mathbb{E}\Big|Y_1^\tau\Big|}{\epsilon}.$$

Now as $\tau \uparrow \infty$ $|Y_1^\tau| \downarrow 0$ while $|Y_1^\tau| \leq |X_1|$, dominated convergence (here a special case of monotone convergence) warrants that $\lim_{\tau\uparrow\infty} \frac{\mathbb{E}\Big|Y_1^\tau\Big|}{\epsilon} = 0$.

This completes the proof of the WLLN. □

9.6 Strong law of large numbers

Infinite product space endowed with cylinders σ-algebra, and infinite product distribution.

Theorem 9.2 (Strong law of large numbers, direct part). *If $X_1, \ldots, X_n, \ldots$ are independently, identically distributed, integrable $\mathbb{R}$-valued random variables over $(\Omega, \mathcal{F}, P)$ with expectation μ then P-a.s.*

$$\lim_{n\to\infty} \overline{X}_n = \mu \qquad \textit{with} \quad \overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i.$$

Recall

Lemma 9.1 (Borel-Cantelli I). *Let $A_1, A_2, \ldots, A_n$ be events from probability space $(\Omega, \mathcal{F}, P)$. If*

$$\sum_n P(A_n) < \infty$$

then:

with probability 1, only finitely many events among $A_1, A_2, \ldots, A_n$ occur:

$$P\Big\{\omega : \sum_n \mathbb{1}_{A_n}(\omega) < \infty\Big\} = 1.$$

Proof. An outcome ω belongs to infinitely many events A_k, iff $\omega \in \cap_n \cup_{k \geq n} A_k$. By monotone convergence,

$$\begin{aligned} P\Big\{\omega : \omega \text{ belongs to infinitely many events } A_k\Big\} &= P\Big\{\cap_n \cup_{k\geq n} A_k\Big\} \\ &= \lim_n \downarrow P\Big\{\cup_{k\geq n} A_k\Big\} \\ &\leq \lim_n \downarrow \textstyle\sum_{k\geq n} P\Big\{A_k\Big\} \\ &= 0\,. \end{aligned}$$

□

Proof of SLLN (direct part). The event $\Big\{\omega : \lim_n \sum_{i=1}^n \frac{X_i}{n} = \mu\Big\}$ belongs to the tail σ-algebra. To check the Strong Law of Large Numbers, it suffices to check that this event has non-zero probability.

Moreover, using the usual decomposition $X = (X)_+ - (X)_-$ where $(X)_+$ and $(X)_-$ are the positive and negative parts of X, we observe that we can assume without loss of generality that X_i's are non-negative.

Recall the definition of truncated variables $X_i^i = \mathbb{I}_{X_i \leq i} X_i$ for $i \in \mathbb{N}$. Let $S_n = \sum_{i=1}^n X_i$ and $T_n = \sum_{i=1}^n X_i^i$.

The difference $S_n - T_n = \sum_{i=1}^n (X_i - X_i^i)$ is a sum of non-negative random variables. As

$$P\{X_i - X_i^i > 0\} = P\{X_i > i\} = P\{X_1 > i\}\,,$$

thanks to $\mathbb{E}X_1 < \infty$,

$$\sum_{i \in \mathbb{N}} P\{X_i - X_i^i > 0\} < \infty\,.$$

By the first Borel-Cantelli Lemma, this implies that almost surely, only finitely many events $\{X_i - X_i^i > 0\}$ are realized. Hence almost surely, T_n and S_n differ by at most a bounded number of summands, and $\lim_n \uparrow (S_n - T_n)$ is finite.

Now

$$\lim_n \uparrow \mathbb{E}\frac{T_n}{n} = \mathbb{E}X_1\,.$$

We shall first check that $T_{n(k)}/n(k)$ converges almost surely towards $\mathbb{E}X_1$ for some (almost) geometrically increasing subsequence $(n(k))_{k\in\mathbb{N}}$.

Fix $\alpha > 1$ and let $n(k) = \lfloor \alpha^k \rfloor$. If for all $\epsilon > 0$, almost surely, only finitely many events

$$\Big\{\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big| \geq n(k) > \epsilon\Big\}$$

occur, then $\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big|/n(k)$ converges almost surely to 0 and thus $T_{n(k)}/n(k)$ converges almost surely to $\mathbb{E}X_1$.

Let

$$\Theta = \sum_{k \in \mathbb{N}} P\Big\{\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big| \geq n(k) > \epsilon\Big\}\,.$$

Thanks to truncation, each $T_{n(k)}$ is square-integrable. By Chebychev's inequality:

$$P\Big\{\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big| \geq n(k) > \epsilon\Big\} \leq \frac{\operatorname{var}(T_{n(k)})}{\epsilon^2 n(k)^2}\,.$$

As X_i^i's are independent,

$$\begin{aligned} \mathrm{var}(T_{n(k)}) &= \textstyle\sum_{i\leq n(k)} \mathrm{var}(X_i^i) \\ &\leq \textstyle\sum_{i\leq n(k)} \mathbb{E}\Big[(X_i^i)^2\Big] \\ &= \textstyle\sum_{i\leq n(k)} \int_0^\infty 2tP\{X_i^i > t\}\mathrm{d}t \\ &\leq \textstyle\sum_{i\leq n(k)} \int_0^i 2tP\{X_1 > t\}\mathrm{d}t\,. \end{aligned}$$

$$\begin{aligned} \Theta \quad &\leq \textstyle\sum_{k\in\mathbb{N}} \frac{1}{\epsilon^2 n(k)^2} \sum_{i\leq n(k)} \int_0^i 2tP\{X_1 > t\}\mathrm{d}t \\ &= \textstyle\frac{1}{\epsilon^2} \sum_{i\in\mathbb{N}} \int_0^i 2tP\{X_1 > t\}\mathrm{d}t \sum_{k:n(k)\geq i} \frac{1}{n(k)^2}\,. \end{aligned}$$

Thanks to the fact that $\alpha^k > 1$ for $k \geq 1$, the following holds:

$$\sum_{k:n(k)\geq i} \frac{1}{n(k)^2} = \sum_{k:\lfloor\alpha^k\rfloor\geq i} \frac{1}{\lfloor\alpha^k\rfloor^2} \leq \frac{4}{i^2}\frac{\alpha^2}{\alpha^2-1}\,.$$

$$\begin{aligned} \Theta \quad &\leq \textstyle\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \sum_{i\in\mathbb{N}} \frac{1}{i^2} \int_0^i 2tP\{X_1 > t\}\mathrm{d}t \\ &\leq \textstyle\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \sum_{i\in\mathbb{N}} \frac{1}{i^2} \sum_{j<i} \int_j^{j+1} 2P\{X_1 > t\}\mathrm{d}t \\ &\leq \textstyle\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \sum_{j=0}^\infty \int_j^{j+1} 2tP\{X_1 > t\}\mathrm{d}t \sum_{i>j} \frac{1}{i^2} \\ &\leq \textstyle\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \sum_{j=0}^\infty \int_j^{j+1} 2tP\{X_1 > t\}\mathrm{d}t \frac{2}{j\vee 1} \\ &\leq \textstyle 8\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \sum_{j=0}^\infty \int_j^{j+1} P\{X_1 > t\}\mathrm{d}t \\ &\leq \textstyle 8\frac{4\alpha^2}{\epsilon^2(\alpha^2-1)} \mathbb{E}X_1 \\ &< \infty\,. \end{aligned}$$

By the first Borell-Cantelli Lemma, with probability 1, only finitely many events

$$\Big\{\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big| \geq n(k) > \epsilon\Big\}$$

occur. As this holds for each $\epsilon > 0$, it holds simultaneously for all $\epsilon = 1/n$, which implies that $\Big|T_{n(k)} - \mathbb{E}T_{n(k)}\Big|/n(k)$ converges almost surely to 0. This also implies that $S_{n(k)}/n(k)$ converges almost surely to $\mathbb{E}X_1$.

To complete the proof, we need to check that this holds for S_n/n.

If $n(k) \leq n < n(k+1)$, as $(S_n)_n$ is non-decreasing,

$$\frac{n(k)}{n(k+1)}\frac{S_{n(k)}}{n(k)} \leq \frac{S_n}{n} \leq \frac{n(k+1)}{n(k)}\frac{S_{n(k+1)}}{n(k+1)}$$

with

$$\frac{1}{\alpha}\Big(1-\frac{1}{\alpha^k}\Big) \leq \frac{n(k+1)}{n(k)} \leq \alpha\left(1+\frac{1}{\lfloor\alpha^k\rfloor}\right)\,.$$

Taking k to infinty, almost surely

$$\frac{1}{\alpha}\mathbb{E}X_1 \leq \liminf_n \frac{S_n}{n} \leq \limsup_n \frac{S_n}{n} \leq \alpha\mathbb{E}X_1\,.$$

Finally, we may choose α arbitrarily close to 1, to establish the desired result.

□ □

Remark 9.1. In the statement of the Theorem, we can replace the independence assumption by a pairwise independence assumption.

Theorem 9.3) shows that, under independence assumption, the conditions in Theorem 9.2) are tight. Before proceeding to the proof of Theorem 9.3), we state and prove the second Borel-Cantelli Lemma.

Lemma 9.2 (Borel-Cantelli II). *Let $A_1, A_2, \ldots, A_n$ be independent events from probability space $(\Omega, \mathcal{F}, P)$.*

If

$$\sum_n P(A_n) = \infty$$

then

with probability 1, infinitely many events among $A_1, A_2, \ldots, A_n$ occur:

$$P\Big\{\omega : \sum_n \mathbb{I}_{A_n}(\omega) = \infty\Big\} = 1\,.$$

Proof. An outcome ω does not belong to infinitely many events A_k, iff $\omega \in \cup_n \cap_{k \geq n} A_k^c$. By monotone convergence,

$$\begin{aligned}
P\Big\{\omega : \omega \text{ does not belong to infinitely many events } A_k\Big\} \quad & = P\Big\{\omega \in \cup_n \cap_{k\geq n} A_k^c\Big\} \\
& = \lim_n \uparrow P\Big\{\cap_{k\geq n} A_k^c\Big\} \\
& = \lim_n \uparrow \lim_{m\uparrow\infty} \downarrow P\Big\{\cap_{k=n}^m A_k^c\Big\} \\
& = \lim_n \uparrow \lim_{m\uparrow\infty} \downarrow \prod_{k=n}^m \Big(1 - P(A_k)\Big\}\Big) \\
& = \lim_n \uparrow \prod_{k=n}^\infty \Big(1 - P(A_k)\Big) \\
& = \lim_n \uparrow \exp\Big(-\sum_{k=n}^\infty P(A_k)\Big) \\
& = \lim_n \uparrow 0 \\
& = 0\,.
\end{aligned}$$

□ □

Theorem 9.3 (Strong law of large numbers, converse part). *Let $X_1, \ldots, X_n, \ldots$ be independently, identically distributed $\mathbb{R}$-valued random variables over some $(\Omega, \mathcal{F}, P)$. If for some finite constant μ,*

$$\lim_{n\to\infty} \sum_{i\leq n} X_i/n = \mu \qquad \textit{almost surely,}$$

then all X_i are integrable and $\mathbb{E}X_i = \mu$.

We may assume that X_i's are non-negative random variables.

Proof. In order to check that the X_i's are integrable, it suffices to show that

$$\sum_{n=0}^\infty P\{X_1 > n\} = \sum_{n=0}^\infty P\{X_n > n\} < \infty.$$

Let $S_n = \sum_{i=1}^n X_i$. Observe that

$$\begin{aligned}
\Big\{\omega : X_{n+1}(\omega) > n+1\Big\} \quad & = \Big\{\omega : S_{n+1}(\omega) - S_n(\omega) > n+1\Big\} \\
& = \Big\{\omega : \tfrac{S_{n+1}(\omega)}{n+1} - \tfrac{S_n(\omega)}{n} > 1 + \tfrac{S_n(\omega)}{n(n+1)}\Big\}\,.
\end{aligned}$$

Assume by contradiction that the X_i's are not integrable. Then by the second Borel-Cantelli Lemma, with probability 1, infinitely many events

$$\left\{\omega : \frac{S_{n+1}}{n+1} - \frac{S_n}{n} > 1 + \frac{S_n}{n(n+1)}\right\}$$

occur. But this cannot happen if S_n/n converges toward a finite limit.
□ □

The law of large numbers is the cornerstone of consistency proofs.

Before shifting to non-exponential inequalities, we point a general result about events that depend on the limiting behavior of sequences of independent random variables.

Definition 9.3 (Tail sigma-algebra). Assume $X_1, \ldots, X_n, \ldots$ are random variables. The tail σ-algebra (or the σ-algebra of tail events) is defined as:

$$\mathcal{T} = \cap_{n=}^{\infty} \sigma\Big(X_n, X_{n+1}, \ldots\Big) .$$

Observe that the event $\sum_{i=1}^{n} X_i/n$ converges towards a finite limit belongs to the tail σ-algebra. The Strong Law of Large Numbers tells us that under integrability and independence assumptions, this tail event has probability 1. This is no accident. The $0-1$-law asserts that under independence, tail events have trivial probabilities.

Theorem 9.4 (0-1-Law). *Assume $X_1, \ldots, X_n, \ldots$ are independent random variables. Any event in the tail σ-algebra $\mathcal{T}$ has probability either 0 or 1.*

Proof. It suffices to check that any event $A \in \mathcal{T}$ satisfies $P(A)^2 = P(A)$, or equivalently that $P(A) = P(A \cap A) = P(A) \times P(A)$, that is A is independent of itself.

For any n, as an event in $\sigma(X_n, X_{n+1}, \ldots)$, A is independent from any event in $\sigma(X_1, \ldots, X_n)$. But this entails that A is independent from any event in $\cup_n \sigma(X_1, \ldots, X_n)$.

Observe that $\cup_n \sigma(X_1, \ldots, X_n)$ is a π-system. Hence, A is independent from any event from the σ-algebra generated by $\cup_n \sigma(X_1, \ldots, X_n)$, which happens to be $\mathcal{F}$. As $A \in \mathcal{T} \subset \mathcal{F}$, A is independent from itself.
□ □

Exercise 9.2. Derive the second Borel-Cantelli Lemma as a special case of the $0-1$-law.

9.7 Exponential inequalities

Laws of large numbers are asymptotic statements. In applications, in Statistics, in Statistical Learning Theory, it is often desirable to have guarantees for fixed n. Exponential inequalities are refinements of Chebychev inequality. Under strong integrability assumptions on the summands, it is possible and relatively easy to derive sharp tail bounds for sums of independent random variables.

Hoeffding's Lemma

Let Y be a random variable taking values in a bounded interval $[a, b]$ and let $\psi_Y(\lambda) = \log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)}$. Then

$$\operatorname{var}(Y) \leq \frac{(b-a)^2}{4} \qquad \text{and} \qquad \psi_Y(\lambda) \leq \frac{1}{2}\frac{(b-a)^2}{4} .$$

Proof. The upper bound on the variance of Y has been established in Section 3.8).

Now let P denote the distribution of Y and let P_λ be the probability distribution with density

$$x \to e^{-\psi_Y(\lambda)} e^{\lambda(x-\mathbb{E}Y)}$$

with respect to P.

Since P_λ is concentrated on $[a,b]$ ($P_\lambda([a,b]) = P([a,b]) = 1$), the variance of a random variable Z with distribution P_λ is bounded by $(b-a)^2/4$. Note that $P_0 = P$.

Dominated convergence arguments allow to compute the derivatives of $\psi_Y(\lambda)$. Namely

$$\psi'_Y(\lambda) = \frac{\mathbb{E}\Big[(Y-\mathbb{E}Y)e^{\lambda(Y-\mathbb{E}Y)}\Big]}{\mathbb{E}e^{\lambda(Y-\mathbb{E}Y)}} = \mathbb{E}_{P_\lambda} Z\,.$$

and

$$\psi''_Y(\lambda) = \frac{\mathbb{E}\Big[(Y-\mathbb{E}Y)^2 e^{\lambda(Y-\mathbb{E}Y)}\Big]}{\mathbb{E}e^{\lambda(Y-\mathbb{E}Y)}} - \left(\frac{\mathbb{E}\Big[(Y-\mathbb{E}Y)e^{\lambda(Y-\mathbb{E}Y)}\Big]}{\mathbb{E}e^{\lambda(Y-\mathbb{E}Y)}}\right)^2 = \mathrm{var}_{P_\lambda}(Z)\,.$$

Hence, thanks to the variance upper bound:

$$\psi''_Y(\lambda) \quad \leq \tfrac{(b-a)^2}{4}\,.$$

Note that $\psi_Y(0) = \psi'_Y(0) = 0$, and by Taylor's theorem, for some $\theta \in [0,\lambda]$,

$$\psi_Y(\lambda) = \psi_Y(0) + \lambda\psi'_Y(0) + \frac{\lambda^2}{2}\psi''_Y(\theta) \leq \frac{\lambda^2(b-a)^2}{8}\,.$$

□ □

The upper bound on the variance is sharp in the special case of a ***Rademacher*** random variable X whose distribution is defined by $P\{X=-1\} = P\{X=1\} = 1/2$. Then one may take $a = -b = 1$ and $\mathrm{var}(X) = 1 = (b-a)^2/4$.

We can now build on Hoeffding's Lemma to derive very practical tail bounds for sums of bounded independent random variables.

Theorem 9.5 (Hoeffding's inequality)**.** ***Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let***

$$S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)\ .$$

Then

$$\mathrm{var}(S) \leq \sum_{i=1}^n \frac{(b_i-a_i)^2}{4}\,.$$

Let v denote the upper bound on the variance. For any $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}e^{\lambda S} \leq \frac{\lambda^2 v}{2}\,.$$

Then for every $t > 0$,

$$P\{S \geq t\} \leq \exp\left(-\frac{t^2}{2v}\right)\ .$$

The proof is based on the so-called Cramer-Chernoff bounding technique and on Hoeffding's Lemma.

Proof. The upper bound on variance follows from $\operatorname{var}(S) = \sum_{i=1}^n \operatorname{var}(X_i)$ and from the first part of Hoeffding's Lemma.

For the upper-bound on $\log \mathbb{E}\mathrm{e}^{\lambda S}$,

$$\begin{aligned}
\log \mathbb{E}\mathrm{e}^{\lambda S} &= \log \mathbb{E}\mathrm{e}^{\sum_{i=1}^n \lambda(X_i - \mathbb{E}X_i)} \\
&= \log \mathbb{E}\Big[\prod_{i=1}^n \mathrm{e}^{\lambda(X_i-\mathbb{E}X_i)}\Big] \\
&= \log\Big(\prod_{i=1}^n \mathbb{E}\Big[\mathrm{e}^{\lambda(X_i-\mathbb{E}X_i)}\Big]\Big) \\
&= \sum_{i=1}^n \log \mathbb{E}\Big[\mathrm{e}^{\lambda(X_i-\mathbb{E}X_i)}\Big] \\
&\le \sum_{i=1}^n \frac{\lambda^2(b_i-a_i)^2}{8} \\
&= \frac{\lambda^2 v}{2}
\end{aligned}$$

where the third equality comes from independence of the X_i's and the inequality follows from invoking Hoeffding's Lemma for each summand.

The Cramer-Chernoff technique consists of using Markov's inequality with exponential moments.

$$\begin{aligned}
P\{S \ge t\} &\le \inf_{\lambda\ge 0} \frac{\mathbb{E}\mathrm{e}^{\lambda S}}{\mathrm{e}^{\lambda t}} \\
&\le \exp\Big(-\sup_{\lambda\ge 0}\big(\lambda t - \log \mathbb{E}\mathrm{e}^{\lambda S}\big)\Big) \\
&\le \exp\Big(-\sup_{\lambda\ge 0}\big(\lambda t - \tfrac{\lambda^2 v}{2}\big)\Big) \\
&= \mathrm{e}^{-\frac{t^2}{2v}}\,.
\end{aligned}$$

□ □

Hoeffding's inequality provides interesting tail bounds for binomial random variables which are sums of independent $[0,1]$-valued random variables. However in some cases, the variance upper bound used in Hoeffding's inequality is excessively conservative. Think for example of binomial random variable with parameters n and μ/n, the variance upper-bound obtained from the boundedness assumption is n while the true variance is μ. This motivates the next two exponential inequalities stated in Theorem 9.6) and Theorem 9.7).

Theorem 9.6 (Bennett's inequality). *Let $X_1, \ldots, X_n$ be independent random variables with finite variance such that $X_i \le b$ for some $b > 0$ almost surely for all $i \le n$. Let*

$$S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$$

and $v = \sum_{i=1}^n \mathbb{E}\left[X_i^2\right]$. Let $\phi(u) = e^u - u - 1$ for $u \in \mathbb{R}$. Then, for all $\lambda > 0$,

$$\log \mathbb{E}e^{\lambda S} \le \frac{v}{b^2}\phi(b\lambda)\,,$$

and for any $t > 0$,

$$P\{S \ge t\} \le \exp\left(-\frac{v}{b^2} h\left(\frac{bt}{v}\right)\right)$$

where $h(u) = \phi^*(u) = (1+u)\log(1+u) - u$ for $u > 0$.

Remark 9.2. Bennett's inequality provides us with improved tail bounds for the binomial random variable with parameters n and μ/n. This binomial random variable is distributed like the sum n independent Bernoulli random variables with parameter μ/n. This fits in the scope of Bennett's inequality, we can choose $b = 1$ and $v = \mu$. The obtained upper bound on the logarithmic moment generating function coincides with logarithmic moment generating function of a centered Poisson random variable with parameter μ, see Theorem 7.5).

Proof. The proof combines the Cramer-Chernoff technique with an ***ad hoc*** upper bound on $\log \mathbb{E}\mathrm{e}^{\lambda(X_i - \mathbb{E}X_i)}$.

By homogeneity, we may assume $b = 1$.

Note that $\phi(\lambda)/\lambda^2$ is non-decreasing over $\mathbb{R}$. For $x \leq 1, \lambda \geq 0$, $\phi(\lambda x) \leq x^2\phi(\lambda)$.

$$\begin{aligned}\log \mathbb{E}\mathrm{e}^{\lambda(X_i - \mathbb{E}X_i)} & = \log \mathbb{E}\mathrm{e}^{\lambda X_i} - \lambda\mathbb{E}X_i \\ & \leq \mathbb{E}\mathrm{e}^{\lambda X_i} - 1 - \lambda\mathbb{E}X_i \\ & = \mathbb{E}\phi(\lambda X_i) \\ & = \mathbb{E}X_i^2\phi(\lambda)\,.\end{aligned}$$

□ □

Whereas Bennett's bound works well for Poisson-like random variables, our last bound is geared towards Gamma-like random variables. It is one of the pillars of statistical learning theory.

Theorem 9.7 (Bernstein's inequality). ***Let*** $X_1, \ldots, X_n$ ***be independent real-valued random variables. Assume that there exist positive numbers*** v ***and*** c ***such that*** $\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq v$ ***and***

$$\sum_{i=1}^n \mathbb{E}\left[(X_i)_+^q\right] \leq \frac{q!}{2} v c^{q-2} \quad \textit{for all integers } q \geq 3\,.$$

Let $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)\,.$

Then for all $\lambda \in (0, 1/c)$***,***

$$\log \mathbb{E}\mathrm{e}^{\lambda(S - \mathbb{E}S)} \leq \frac{v\lambda^2}{2(1 - c\lambda)}\,.$$

For $t > 0$***,***

$$P\{S > t\} \leq \exp\left(-\frac{v}{c^2} h_1\left(\frac{ct}{v}\right)\right)$$

with $h_1(x) = 1 + x - \sqrt{1 + 2x}$***.***

Proof. The proof combines again the Cramer-Chernoff technique with an ***ad hoc*** upper bound on $\log \mathbb{E}\mathrm{e}^{\lambda(S - \mathbb{E}S)}$.

Let again $\phi(u) = e^u - u - 1$ for $u \in \mathbb{R}$.

For $\lambda > 0$,

$$\begin{aligned}\phi(\lambda X_i) & = \sum_{k=2}^\infty \frac{\lambda^k X_i^k}{\lambda^k} \\ & \leq \frac{\lambda^2 X_i^2}{2!} + \sum_{k=3}^\infty \frac{\lambda^k (X_i)_+^k}{\lambda^k}\,.\end{aligned}$$

For $c > \lambda > 0$,

$$\begin{aligned}\log \mathbb{E}\mathrm{e}^{\lambda S} & = \sum_{i=1}^n \log \mathbb{E}\mathrm{e}^{\lambda(X_i - \mathbb{E}X_i)} \\ & \leq \sum_{i=1}^n \mathbb{E}\phi(\lambda X_i) \\ & \leq \frac{\lambda^2 \sum_{i=1}^n \mathbb{E}X_i^2}{2!} + \sum_{k=3}^\infty \frac{\lambda^k \sum_{i=1}^n \mathbb{E}(X_i)_+^k}{k!} \\ & \leq \frac{\lambda^2 v}{2} + \sum_{k=3}^\infty \frac{\lambda^k v c^{k-2}}{2} \\ & = \frac{\lambda^2 v}{2(1 - c\lambda)}\,.\end{aligned}$$

The tail bound follows by maximizing

$$\sup_{\lambda \in [0,1/c)} \lambda t - \frac{\lambda^2 v}{2(1-c\lambda)} = \frac{v}{c^2} \sup_{\eta \in [0,1)} \eta \frac{ct}{v} - \frac{\eta^2}{2(1-\eta)} .$$

□ □

9.8 Bibliographic remarks

(Dudley, 2002) contains a thorough discussion of the various kinds of convergences that can be defined for random variables. In particular, (Dudley, 2002) offers a general perspective on topological issues in probability spaces. (Dudley, 2002) also tackles the problem raised by random variables that take values in (possibly infinite-dimensional) metric spaces.

Laws of large numbers and $0-1$-laws fit in the more general framework of ergodic theorems, see (Dudley, 2002) or (Durrett, 2010). An important example of law of large numbers is the Asymptotic Equipartition Property (AEP) in Information Theory. Note that it holds for a much larger class of sources than the set of memoryless sources (infinite product probability spaces). See (Cover & Thomas, 1991) or [csiszar:korner:1981].

Introduction to exponential inequalities and their applications can be found in (Massart, 2007), (Boucheron et al., 2013).

Chapter 10

Convergence in distribution

10.1 Motivation

Recall Lesson 1. Consider Binomial distributions with parameters $(n, \lambda/n)$ and Poisson distribution with parameter λ. Graphical inspection of probability mass functions suggests that as n grows, Binomial distributions with parameters $(n, \lambda/n)$ look more and more alike Poisson distribution with parameter λ. Comparing probability generating functions is more compelling. The probability generating function of the Binomial distribution with parameters $(n, \lambda/n)$ is $s \mapsto (1 + \lambda(s-1)/n)^n$. As n tends towards infinity, the probability generating functions of the Binomials converge pointwise towards the probability generating function of the Poisson distribution with mean λ: $s \mapsto \exp(\lambda(s-1))$. In Lesson 1, we saw other examples of distributions which tend to look alike some limiting distributions as some parameter moves.

In Lesson 9, we equipped the set $L_0(\Omega, \mathcal{F}, P)$ with topologies (L_p, almost sure convergence, convergence in probability). In this lesson, we consider the set of probability distributions over some measurable space $(\Omega, \mathcal{F})$. This set can be equipped with a variety of topologies. We shall focus on the topology defined by ***convergence in distribution*** also called ***weak convergence***.

In Section 10.2), we introduce weak and vague convergences for sequences of probability distributions. In Section 10.3) Weak convergence induces the definition of convergence in distribution for random variables that possibly live on different probability spaces (just as our occupancy scores in Lesson 1).

Section 10.4) is dedicated to the Portemanteau Theorem. This theorem lists a number of alternative and equivalent characterizations of convergence in distribution. Alternative characterizations are useful in two respects: they may be easier to check than the characterization used in the definition; they may supply a larger range of applications.

In Section 10.5), we state and prove the Lévy continuity theorem. The Levy continuity theorem relates convergence in distribution with pointwise convergence of characteristic functions: characteristic functions not only allow us to identify probability distributions, they are also convergence determining. It could be one more line in the statement of Theorem 10.1). But the Lévy continuity Theorem stands out because it provides us with a concise proof of the Central Limit Theorem for normalized sums of centered i.i.d. random variables. This is the content of Section 10.8).

10.2 Weak convergence, vague convergence

Weak convergence of probability measures assesses the proximity of probability measures by comparing their action on a collection of test functions.

Definition 10.1 (Weak convergence). A sequence of probability distributions $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converges ***weakly*** towards probability distribution P(on $\mathbb{R}^k$)

iff

for any bounded and continuous function f from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $(\mathbb{E}_{P_n}[f])_{n\in\mathbb{N}}$ converges towards $\mathbb{E}_P[f]$.

Remark 10.1. We shall see that the there is some flexibility in the choice of the class of test functions.

But this choice is not unlimited.

If we restrict the collection of test functions to continuous functions with ***compact support*** (which are always bounded), we obtain a different notion of convergence.

Definition 10.2 (Vague convergence). A sequence of probability distributions $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converges vaguely towards measure μ (on $\mathbb{R}^k$) iff for any continuous function f with compact support from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $(\mathbb{E}_{P_n}[f])_{n\in\mathbb{N}}$ converges towards $\mathbb{E}_P[f]$.

Example 10.1. Consider the sequence of probability masses over the integers $(\delta_n)_{n\in\mathbb{N}}$. This sequence converges vaguely towards the null measure. It does not converge weakly.

The next question deserves further thinking.

Exercise 10.1. If a sequence of probability distributions over $\mathbb{R}^k$ converges vaguely towards a probability measure, does it also converge weakly towards this probability measure?

10.3 Convergence in distribution

Definition 10.3 (Convergence in distribution). A sequence $(X_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^k$-valued random variables defined on a sequence of probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ converges in distribution if the sequence $(P_n \circ X_n^{-1})_{n\in\mathbb{N}}$ converges weakly. This is denoted by

$$X_n \rightsquigarrow X \qquad \text{or} \qquad X_n \rightsquigarrow \mathcal{L}$$

($\mathcal{L}$ denotes a probability distribution), the probability spaces are defined implicitly

In order to check or use convergence in distribution, many equivalent characterizations are available. Some of them are listed in the Portemanteau Theorem.

10.4 Portemanteau Theorem

The next list of equivalent characterizations of weak convergence is not exhaustive.

Theorem 10.1 (Portemanteau Theorem). *A sequence of probability distributions $(P_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^k$ converges weakly towards a probability distribution P (on $\mathbb{R}^k$) iff one of the equivalent properties hold:*

1. *For every bounded continuous function f from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $\mathbb{E}_{P_n}[f]$ converges towards $\mathbb{E}_P[f]$.*

2. *For every bounded uniformly continuous function f from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $\mathbb{E}_{P_n}[f]$ converges towards $\mathbb{E}_P[f]$.*
3. *For every bounded Lipschitz function f from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $\mathbb{E}_{P_n}[f]$ converges towards $\mathbb{E}_P[f]$.*
4. *For every P-almost surely bounded and continuous function f from $\mathbb{R}^k$ to $\mathbb{R}$, the sequence $(\mathbb{E}_{P_n}[f])$ converges towards $\mathbb{E}_P[f]$.*
5. *For every closed subset F of $\mathbb{R}^k$, $\limsup P_n(F) \leq P(F)$.*
6. *For every open subset O of $\mathbb{R}^k$, $\liminf P_n(O) \geq P(O)$.*
7. *For every Borelian A such that $P(A^\circ) = P(\overline{A})$ (the boundary of A is P-negligible), $\lim_n P_n(A) = P(A)$.*

In English, as in French, a ***portemanteau*** is a suitcase.

Proof. Implications $1) \Rightarrow 2) \Rightarrow 3)$ are obvious. Lévy's continuity theorem, the major result from Section 10.5) entails that $3) \Rightarrow 1)$.

4).

That $5) \Leftrightarrow 6)$ follows from the fact that the complement of a closed set is an open set.

5) and 6) imply 7):

$$\limsup_n P_n(\overline{A}) \leq P(\overline{A}) = P(A^\circ) \leq \liminf_n P_n(A^\circ)\,.$$

By monotony:

$$\liminf_n P_n(A^\circ) \leq \liminf_n P_n(A) \leq \limsup_n P_n(A) \leq \limsup_n P_n(\bar{(}A))\,.$$

Combining leads to

$$\lim_n P_n(A) = \liminf_n P_n(A) = \limsup_n P_n(A) = P(A^\circ) = P(\overline{A})\,.$$

Let us check that $3) \Rightarrow 5)$. Let F be a closed subset of $\mathbb{R}^k$. For $x \in \mathbb{R}^k$, let $\mathrm{d}(x, F)$ denote the distance from x to F. For $m \in \mathbb{N}$, let $f_m(x) = \big(1 - m\mathrm{d}(x, F)\big)_+$. The function f_m is m-Lipschitz, lower bounded by $\mathbb{I}_F$, and for every $x \in \mathbb{R}^k$ $\lim_m \downarrow f_m(x) = \mathbb{I}_F(x)$.

Weak convergence of P_n to P implies

$$\lim_n \mathbb{E}_{P_n} f_m = \mathbb{E}_P f_m$$

hence for every $m \in \mathbb{N}$

$$\limsup_n \mathbb{E}_{P_n} \mathbb{I}_F \leq \mathbb{E}_P f_m\,.$$

Taking the limit on the right side leads to

$$\limsup_n P_n(F) = \limsup_n \mathbb{E}_{P_n} \mathbb{I}_F \leq \lim_m \downarrow \mathbb{E}_P f_m = \mathbb{E}_P \mathbb{I}_{\mathbb{F}} = P(F)\,.$$

Assume now that 7) holds. Let us show that this entails 1)

Let f be a bounded continuous function. Assume w.l.o.g. that f is non-negative and upper-bounded by 1. Recall that for each σ-finite measure μ

$$\int f \mathrm{d}\mu = \int_{[0,\infty)} \mu\{f > t\} \mathrm{d}t\,.$$

This holds for all P_n and P. Hence

$$\mathbb{E}_{P_n} f = \int_{[0,\infty)} P_n\{f > t\} \mathrm{d}t$$

As $\overline{\{f > t\}} = \{f \geq t\}$, $\overline{\{f > t\}} \setminus \{f > t\}^\circ = \{f = t\}$. The set of values t such that $P\{f = t\} > 0$ is at most countable and thus Lebesgue-negligible. Let E be its complement. For $t \in E$, $\lim_n P_n\{f > t\} = P\{f > t\}$.

$$\begin{aligned}
\lim_n \mathbb{E}_{P_n} f &= \lim_n \int_{[0,1]} P_n\{f > t\} \mathrm{d}t \\
&= \lim_n \int_{[0,1]} P_n\{f > t\} \mathbb{1}_E(t) \mathrm{d}t \\
&= \int_{[0,1]} \lim_n P_n\{f > t\} \mathbb{1}_E(t) \mathrm{d}t \\
&= \int_{[0,1]} P\{f > t\} \mathbb{1}_E(t) \mathrm{d}t \\
&= \int_{[0,1]} P\{f > t\} \mathrm{d}t \\
&= \mathbb{E}_P f .
\end{aligned}$$

□

For probability measures over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, weak convergence is determined by cumulative distribution functions. This is sometimes taken as a definition of weak convergence in elementary books.

Corollary 10.1. *A sequence of probability measures defined by their cumulative distribution functions $(F_n)_n$ converges weakly towards a probability measure defined by cumulative distribution function F iff $\lim_n F_n(x) = F(x)$ at every x which is a continuity point of F.*

For probability measures over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, weak convergence is also determined by quantile functions.

Proposition 10.1. *A sequence of probability measures defined by their quantile functions $(F_n^{\leftarrow})_n$ converges weakly towards a probability measure defined by quantile function $F^{\leftarrow}$ iff $\lim_n F_n^{\leftarrow}(x) = F^{\leftarrow}(x)$ at every x which is a continuity point of $F^{\leftarrow}$.*

Prove Proposition Proposition 10.1.

Proposition 10.2 (Almost sure representation). *If $(X_n)_n$ converges in distribution towards X, then there exists a probability space $(\Omega, \mathcal{F}, P)$ with random variables $(Y_n)_n$ and Y such that $X_n \sim Y_n$ for all n, $X \sim Y$, and*

$$Y_n \to Y \qquad P\text{-a.s.}$$

Remark 10.2. The random variables $(X_n)_n$ and X may live on different probability spaces.

When random variables X_n are real-valued, Proposition 10.2 follows easily from Proposition Proposition 10.1.

Proof. Let $\Omega = [0,1], \mathcal{F} = \mathcal{B}(\mathbb{R})$ and ω be uniformly distributed over $\Omega = [0,1]$. Let $Y_n = F_n^{\leftarrow}(\omega)$ and $Y = F^{\leftarrow}(\omega)$.

Then for each n,

$$P\Big\{Y_n \leq t\Big\} = P\Big\{\omega : F_n^{\leftarrow}(\omega) \leq t\Big\} = P\Big\{\omega : \omega \leq F_n(t)\Big\} = F_n(t)$$

so that $Y_n \sim F_n$. And by the same argument, $Y \sim F$.

As a non-decreasing function has at most countably many discontinuities,

$$P\Big\{\omega : F^{\leftarrow} \text{ is continuous at } \omega\Big\} = 1\,.$$

Now, assume ω is a continuity point of $F^{\leftarrow}$. Then by Proposition Proposition 10.1, $\lim_n F_n^{\leftarrow}(\omega) = F^{\leftarrow}(\omega)$. This translates to

$$P\Big\{\omega : \lim_n Y_n(\omega) = Y(\omega)\Big\} = 1\,.$$

□ □

10.5 L□vy continuity theorem

Theorem 10.2 (Lévy's continuity theorem). ***A sequence** $(P_n)_n$ **of probability distributions over** $\mathbb{R}^d$ **converges weakly towards a probability distribution** P **over** $\mathbb{R}^d$ **iff the sequence of characteristic functions converges pointwise towards the characteristic function of** P.*

Remark 10.3. Theorem 10.2 asserts that weak convergence of probability measures is characterized by a very small subset of bounded continuous functions. To warrant weak convergence of $(P_n)_n$ towards P it is enough to check that $\mathbb{E}_{P_n} f \to \mathbb{E}_P f$ for functions f in family $\{\cos(t\cdot), \sin(t\cdot) : t \in \mathbb{R}\}$. These functions are bounded and infinitely many times differentiable.

Let $(X_n)_n$, X and Z live on the same probability space. If $(X_n)_n$, X and Z are random variables such that for every $\sigma > 0$, $X_n + \sigma Z \rightsquigarrow X + \sigma Z$, then $X_n \rightsquigarrow X$.

Proof. Let h be bounded by 1 and 1-Lipschitz

$$\begin{aligned}\Big|\mathbb{E}h(X_n) - h(X)\Big| \leq & \Big|\mathbb{E}h(X_n) - h(X_n + \sigma Z)\Big| \\ & + \Big|\mathbb{E}h(X_n + \sigma Z) - h(X + \sigma Z)\Big| \\ & + \Big|\mathbb{E}h(X + \sigma Z) - h(X)\Big|\end{aligned}$$

The first and third summand can be handled in the same way.

Let $\epsilon > 0$,

$$\begin{aligned}\Big|\mathbb{E}h(X_n) - h(X_n + \sigma Z)\Big| & \leq \Big|\mathbb{E}(h(X_n) - h(X_n + \sigma Z))\mathbb{1}_{\sigma|Z|>\epsilon}\Big| \\ & \quad + \Big|\mathbb{E}(h(X_n) - h(X_n + \sigma Z))\mathbb{1}_{\sigma|Z|\leq\epsilon}\Big| \\ & \leq 2P\{\sigma|Z| > \epsilon\} + \epsilon\,.\end{aligned}$$

Combining the different bounds leads to

$$\Big|\mathbb{E}h(X_n) - h(X)\Big| \leq 2P\{\sigma|Z| > \epsilon\} + \epsilon + \Big|\mathbb{E}h(X_n + \sigma Z) - h(X + \sigma Z)\Big|$$

The last summand on the right-hand-side tends to 0 as n tends to infinity. The first summand tends to 0 as σ tends to 0.

Hence

$$\limsup_n \left|\mathbb{E}h(X_n) - h(X_n + \sigma Z)\right| \leq \epsilon\,.$$

□ □

Lemma 10.1 (Scheffé's Lemma). *Let $(P_n)_n$ be a sequence of absolutely continuous probability distributions with densities $(f_n)_n$. Assume that densities $(f_n)_n$ converge pointwise towards the density f of some probability distribution P, then $P_n \rightsquigarrow P$.*

Proof.

$$\begin{aligned}\int_{\mathbb{R}} |f_n(x) - f(x)|\mathrm{d}x &= \int_{\mathbb{R}} (f(x) - f_n(x))_+ \mathrm{d}x + \int_{\mathbb{R}} (f(x) - f_n(x))_- \mathrm{d}x \\ &= 2\int_{\mathbb{R}} (f(x) - f_n(x))_+ \mathrm{d}x\,.\end{aligned}$$

Observe $(f - f_n)_+ \leq f$ which belongs to $\mathcal{L}_1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \text{Lebesgue})$. And $(f - f_n)_+$ converges pointwise to 0. Hence, by the dominated convergence theorem, $\lim_n \int_{\mathbb{R}} |f_n - f|\mathrm{d}x = 0$.

For any $A \in \mathcal{B}(\mathbb{R})$,

$$P_n(A) - P(A) = \int_{\mathbb{R}} \mathbb{I}_A (f_n - f) \leq \int_{\mathbb{R}} |f_n - f|\,.$$

We have proved more than weak convergence, namely

$$\lim_n \sup_{A \in \mathcal{B}(\mathbb{R})} |P_n(A) - P(A)| = 0\,.$$

□ □

Proof of continuity theorem. Assume the characteristic functions of $(X_n)_n$ converges pointwise towards the characteristic function of X.

Let Z be a standard Gaussian random variable, independent of all $(X_n)_n$ and of X. For $\sigma > 0$, the distributions of $X_n + \sigma Z$ and $X + \sigma Z$ have densities that are uniquely determined by the characteristic functions of X_n and X. Moreover, a dominated convergence argument shows that the densities of $X_n + \sigma Z$ converge pointwise towards the density of $X + \sigma Z$. By Scheffé's Lemma, this entails that $X_n + \sigma Z \rightsquigarrow X + \sigma Z$.

As this holds for all $\sigma > 0$, this entails that $X_n \rightsquigarrow X$.

□ □

10.6 Refining the continuity theorem

In some situations, we can prove that a sequence of characteristic functions converges pointwise towards some function, but we have no candidate for the limiting distribution. The question arises whether the pointwise limit of characteristic functions is the characteristic function of some probability distribution or something else.

The answer may be negative: if $P_n = \mathcal{N}(0, n)$, the sequence of characteristic functions is $\left(t \mapsto \exp(-nt^2/2)\right)_n$ which converges pointwise to 0 except at 0 where it is equal to 1 all along. The limit is not the characteristic function of any probability measure: it is not continuous at 0.

The next Theorem settles the question.

Theorem 10.3 (Lévy's continuity theorem, second form)**.** *A sequence $(P_n)_n$ of probability distributions over $\mathbb{R}$ converges weakly towards a probability distribution over $\mathbb{R}$ iff the sequence of characteristic functions converges pointwise towards a function that is continuous at 0. The limit function is the characteristic function of some probability distribution.*

Definition 10.4 (Uniform tightness)**.** A sequence of Probability measures $(P_n)_n$ over $\mathbb{R}$ is *uniformly tight* if for every $\epsilon > 0$, there exists some compact $K \subseteq \mathbb{R}$ such that

$$\forall n, \qquad P_n(K) \geq 1 - \epsilon\,.$$

Exercise 10.2. To establish uniform tightness of $(P_n)_n$, it is enough to show that for every $\epsilon > 0$, there exists some $n_0(\epsilon)$, and some compact $K \subseteq \mathbb{R}$ such that

$$\forall n \geq n_(\epsilon), \qquad P_n(K) \geq 1 - \epsilon\,.$$

We admit the (important) next Theorem.

Theorem 10.4 (Prokhorov-Le Cam)**.** *If $(P_n)_n$ is a uniformly tight sequence of probability measures on $\mathbb{R}$, then there exists some probability measure P and some subsequence $(P_{n(k)})_{k\in\mathbb{N}}$ such that*

$$P_{n(k)} \rightsquigarrow P\,.$$

Then

Lemma 10.2 (Uniform tightness Lemma)**.** *Let $(P_n)_n$ be a sequence of probability distributions over $\mathbb{R}$, with characteristic functions $\widehat{F}_n$. If the sequence $(\widehat{F}_n)_n$ converge pointwise towards a function that is continuous at 0 then the sequence $(P_n)_n$ is* uniformly tight.

We shall use the following technical upper bound which is illustrated in Figure 10.1:

$$\forall t \in \mathbb{R} \setminus [-1, 1], \qquad \frac{\sin(t)}{t} \leq \sin(1) \leq \frac{6}{7}\,.$$

Proposition 10.3 (Truncation Lemma)**.** *Let $\widehat{F}$ be the characteristic function of some probability measure P on the real line, then for all $u > 0$:*

$$\frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}(v))\mathrm{d}v \geq \frac{1}{7}P\Big[\frac{-1}{u}, \frac{1}{u}\Big]^c\,.$$

Proof of Truncation Lemma.

$$\begin{aligned}
\frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}_n(v))\mathrm{d}v &= \frac{1}{u}\int_0^u \Big(\int_{\mathbb{R}} (1 - \cos(vw))\mathrm{d}F(w)\Big)\mathrm{d}v \\
&= \int_{\mathbb{R}}\int_0^u \frac{1}{u}\Big((1 - \cos(vw))\mathrm{d}v\Big)\mathrm{d}F(w) \\
&= \int_{\mathbb{R}}\Big(1 - \frac{\sin(uw)}{uw}\Big)\mathrm{d}F(w) \\
&\geq \int_{|uw|\geq 1}\Big(1 - \frac{\sin(uw)}{uw}\Big)\mathrm{d}F_n(w) \\
&\geq (1 - \sin(1))P\Big[\frac{-1}{u}, \frac{1}{u}\Big]^c
\end{aligned}$$

where the two inequalities follow from the bounds on the sinc function.

□ □

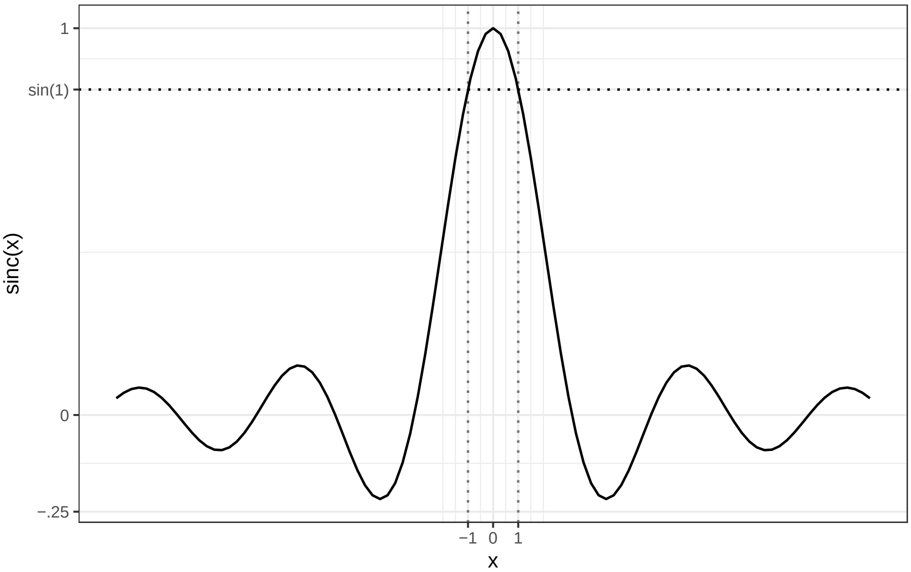

Figure 10.1: The proof of the truncation inequality takes advantage on easy bounds satisfied by the sinc function.

Proof of Uniform tightness Lemma. Assume that the sequence $(\widehat{F}_n)_n$ converge pointwise towards a function $\widehat{F}$ that is continuous at 0.

Note that $\widehat{F}_n(0) = 1$ for all n, hence, trivially, $1 = \lim_n \widehat{F}_n(0) = \widehat{F}(0)$.

As $|\operatorname{Re}\widehat{F}_n(t)| \leq 1$, $|\operatorname{Re}\widehat{F}(t)| \leq 1$ also holds.

Fix $\epsilon > 0$, as $\widehat{F}$ is continuous at 0, for some $u > 0$, for all $v \in [-u, u]$, $0 \geq 1 - \widehat{F}(u) \leq \epsilon/2$. Hence,

$$0 \leq \frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}(v))\mathrm{d}v \leq \epsilon/2\,.$$

By dominated convergence,

$$\lim_n \frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}_n(v))\mathrm{d}v = \frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}(v))\mathrm{d}v \leq \epsilon/2\,.$$

For sufficiently large n, $0 \leq \frac{1}{u}\int_0^u (1 - \operatorname{Re}\widehat{F}_n(v))\mathrm{d}v \leq \epsilon$.

Applying the truncation Lemma, for sufficiently large n, we have

$$P_n\Big[\frac{-1}{u}, \frac{1}{u}\Big]^c \leq 7\epsilon\,.$$

The interval $\left[\frac{-1}{u}, \frac{1}{u}\right]$ is compact.

□ □

Proof of the second form of the continuity theorem. We combine the Uniform Tightness Lemma and the Prokhorov-Le Cam Theorem. Under the assumptions of the second form of the continuity Theorem, there is a probability measure P (with characteristic function $\widehat{F}$), and a subsequence $(P_{n(k)})_{k\in\mathbb{N}}$ such that $P_{n(k)} \rightsquigarrow P$ as $k \to \infty$. This entails $\widehat{F}_{n(k)} \to \widehat{F}$ as $k \to \infty$ pointwise. This also entails that $\widehat{F}_n \to \widehat{F}$ pointwise for the whole sequence. Finally, we are able to invoke Theorem 10.2) to conclude $P_n \rightsquigarrow P$ as $n \to \infty$.

□ □

Remark 10.4. All definitions and results in this section can be extended to the k-dimensional setting for all $k \in \mathbb{N}$.

10.7 Relations between convergences

The alternative characterizations of weak convergence provided by the Portemanteau Theorem (Theorem 10.1)) facilitate the proof of the next Proposition.

Convergence in probability implies convergence in distribution.

Proof. Assume $(X_n)_n$ converges in probability towards X.

Let h be a bounded and Lipschitz function. Without loss of generality, assume that $|f(x)| \leq 1$ for all x and $|f(x) - f(y)| \leq \mathrm{d}(x, y)$.

Let $\epsilon > 0$.

$$\begin{aligned}
\Big|\mathbb{E}h(X_n) - \mathbb{E}h(X)\Big| &= \Big|\mathbb{E}\Big[(h(X_n) - h(X)\mathbb{1}_{\mathrm{d}(X,X_n)>\epsilon}\Big] \\
&\qquad + \mathbb{E}\Big[(h(X_n) - h(X)\mathbb{1}_{\mathrm{d}(X,X_n)\leq\epsilon}\Big]\Big| \\
&\leq \mathbb{E}\Big[2\mathbb{1}_{\mathrm{d}(X,X_n)>\epsilon}\Big] \\
&\qquad + \mathbb{E}\Big[|h(X_n) - h(X)|\mathbb{1}_{\mathrm{d}(X,X_n)\leq\epsilon}\Big] \\
&\leq 2P\{\mathrm{d}(X, X_n) > \epsilon\} + \epsilon\,.
\end{aligned}$$

Convergence in probability entails that

$$\limsup_n \Big|\mathbb{E}h(X_n) - \mathbb{E}h(X)\Big| \leq \epsilon.$$

As this holds for every $\epsilon > 0$, for every bounded Lipschitz function h, $\lim_n \Big|\mathbb{E}h(X_n) - \mathbb{E}h(X)\Big| = 0$. This is sufficient to establish convergence in distribution of $(X_n)_n$.

□ □

10.8 Central limit theorem

The Lévy Continuity Theorem (Theorem 10.2)) is the conerstone of a very concise proof the simplest version of the Central Limit Theorem (CLT). Under square-integrability assumption, the CLT refines the Laws of Large Numbers. It states that as n tends to infinity, the fluctuations of the empirical mean $\sum_{i=1}^n X_i/n$ around its expectation tends to be of order $1/\sqrt{n}$ and, once rescaled, to be normally distributed.

Theorem 10.5. *Let $X_1, \ldots, X_n, \ldots$ be i.i.d. with finite variance σ^2 and expectation μ. Let $S_n = \sum_{i=1}^n X_i$.*

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \rightsquigarrow \mathcal{N}(0, \sigma^2)\,.$$

Proof. Let $\widehat{F}$ denote the characteristic function of the (common) distribution of the random variables $((X_i - \mu)/\sigma)_i$. Recall fron Lesson 7, that the centering and square integrability assumptions imply that

$$\widehat{F}(t) = \widehat{F}(0) + \widehat{F}'(0)t + \frac{\widehat{F}'(0)}{2}t^2 + t^2R(t) = 1 - \frac{t^2}{2} + t^2R(t)$$

where $\lim_{t\to 0} R(t) = 0$. Let $\widehat{F}_n$ denote the characteristic function of $\sqrt{n}\left(\frac{S_n}{n} - \mu\right)/\sigma$. Fix $t \in \mathbb{R}$,

$$\widehat{F}_n(t) = \left(\widehat{F}(t/\sqrt{n})\right)^n = \left(1 - \frac{t^2}{2n} + \frac{t^2}{n} R(t/\sqrt{n})\right)^n .$$

As $n \to \infty$,

$$\lim_n \left(1 - \frac{t^2}{2n} + \frac{t^2}{n} R(t/\sqrt{n})\right)^n = \mathrm{e}^{-\frac{t^2}{2}} .$$

On the right-hand-side, we recognize the characteristic function of $\mathcal{N}(0, 1)$.

□ □

Remark 10.5. The conditions in the Theorem statement allows for a short proof. They are by no mean necessary. The summands need not be identically distributed. The summands need not be independent. A version of the Lindeberg-Feller Theorem states that under mild assumptions, centered and normalized sums of independent square-integrable random variables converge in distribution towards a Gaussian distribution.

10.9 Cramer- Wold device

So far, we have discussed characteristic functions for real valued random variables. But characteristic functions can be defined for vector-valued random variables. If X is a $\mathbb{R}^k$-valued random variable, its characteristic function maps $\mathbb{R}^k$ to $\mathbb{C}$:

$$\begin{aligned} \mathbb{R}^k &\to \mathbb{C} \\ t &\mapsto \mathbb{E}\mathrm{e}^{i\langle t, X\rangle} . \end{aligned}$$

The importance of multivariate characteristic functions is reflected in the next device which proof is left to the reader. It consists in the adapting the proof of Theorem 7.6).

Theorem 10.6 (Cramer-Wold). *The distribution of a $\mathbb{R}^k$-valued random vector $X = (X_1, \ldots, X_k)^T$ is completely determined by the collection of distributions of univariate random variables $\langle t, X\rangle = \sum_{i=1}^n t_i X_i$ where $(t_1, \ldots, t_n)^T$ belongs to $\mathbb{R}^n$.*

Theorem 11.1) provides a short path to the Multivariate Central Limit Theorem.

Theorem 10.7. *Let $X_1, \ldots, X_n, \ldots$ be i.i.d. vector valued random variables with finite covariance Γ and expectation μ. Let $S_n = \sum_{i=1}^n X_i$.*

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \rightsquigarrow \mathcal{N}(0, \Gamma) .s$$

10.10 Weak convergence and transforms

In Lesson Chapter 7, we introduced different characterizations of probability distributions: probability generating functions, Laplace transforms, Fourier transforms (characteristic functions), cumulative distribution functions, quantiles functions. Within their scope, all those transforms are *convergence determining*: if a sequence of probability distributions converges weakly, so does (pointwise) the corresponding sequence of transforms, at least at the continuity points of the limiting transform.

In the next two theorems, each random variable is assumed to live on some (implicit) probability space.

A sequence of ***non-negative*** random variables $(X_n)_n$ converges in distribution towards the ***non negative*** random variable X iff the sequence of Laplace transforms converges pointwise towards the Laplace transform of the probability distribution of X.

The proof parallels the derivation of Theorem 10.2).

As probability generating functions allows us to recover Laplace transforms, the next theorem is a special case of the statement concerning Laplace transforms.

A sequence of integer-valued random variables $(X_n)_n$ converges in distribution towards the integer-valued random variable X iff the sequence of Laplace transforms converges pointwise towards the Laplace transform of the probability distribution of X.

10.11 Bibliographic remarks

Dudley (2002) discusses convergence in distributions in two chapters: the first one is dedicated to distributions on $\mathbb{R}^d$ and the central limit theorem; the second chapter addresses more general universes. In the first chapter, the central limit theorem is extended to triangular arrays that is to sequences of not necessarily identically distributed random variables (Lindeberg's Theorem).

Dudley (2002) investigates convergence in distributions as ***convergence of laws on separable metric spaces***, that is in a much broader context than we do in these notes. The reader will find there a complete proof of the Prokhorov-Le Cam Theorem and an in-depth discussion of its corollaries. In (Dudley, 2002), a great deal of effort is dedicated to the metrization of the weak convergence topology. The reader will also find in this book a full picture of almost sure representation arguments.

The proof of the Lévy Continuity Theorem given here is taken from (Pollard, 2002).

Using metrizations for weak convergence allows us to investigate rate of convergence in limit theorems. This goes back at least to the Berry-Esseen's Theorem (1942). Quantitative approaches to weak convergence have acquired a new momentum with the popularization of Stein's method. This methods is geared towards, but exclusively focused on, general yet quantitative versions of the Central Limit Theorem (Chen, Goldstein, & Shao, 2011) . A thorough yet readable introduction to Stein's method is (Nathan Ross, 2011).

Chapter 11

Refinments and extensions of thr Central Limit Theorem

11.1 Motivation

In Section 11.4), we associate convergence in distribution with metrics. By metrizing convergence in distribution, we can quantify rates of convergence.

In Section 11.5) we turn a relatively recent approach to weak convergence: Stein's method. This is best illustrated with variations on the Central Limit Theorem. Recall that a byproduct of the proof that characteristic functions identify probability distributions was Stein's identity.

11.2 Central limit theorem

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. with finite variance σ^2 and expectation μ. Let $S_n = \sum_{i=1}^n X_i$.

$$\sqrt{n}\left(\frac{S_n}{n} - \mu\right) \rightsquigarrow \mathcal{N}(0, \sigma^2)\,.$$

Proof. TODO

□ □

11.3 Cramer-Wold device

The next theorem is a corollary of **?@thm-thmLevyInj**).

Theorem 11.1 (Cramer-Wold). *The distribution of a $\mathbb{R}^n$-valued random vector $X = (X_1, \ldots, X_n)^T$ is completely determined by the collection of distributions of univariate random variables $\langle t, X\rangle = \sum_{i=1}^n t_i X_i$ where $(t_1, \ldots, t_n)^T$ belongs to $\mathbb{R}^n$.*

11.4 Metrizations of weak convergence

When we talk about laws on complete separable metric spaces, convergence in distribution can be defined in relation to a distance. In fact, several distances are possible and you can

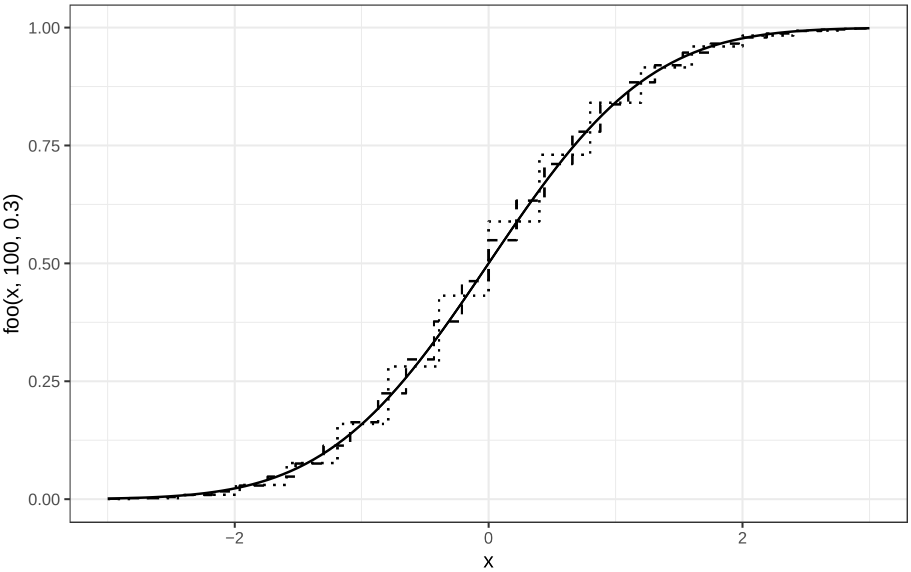

Figure II.1: Pointwise convergence of cumulative distribution functions towards the Gaussian cumulative distribution functions of $\mathcal{N}(0,1)$ (plain line) illustrates the Central Limit Theorem. The dotted and the dashed lines represent the cumulative distribution function of $(X_n - np)/\sqrt{n(p*(1-p))}$ where $X_n \sim$ Binomial(n,p) for $p = .3$ and $n = 30(dotted), 100(dashed)$.

choose any of them according to your needs. The distances we are interested in are of the form

$$\mathrm{d}_{\mathcal{H}}(P,Q) = \sup_{h\in\mathcal{H}} \int h\mathrm{d}P - \int h\mathrm{d}Q$$

where $\mathcal{H}$ is a well-chosen collection of measurable functions.

The Kolmogorov-Lévy metric is used implicitly in the formulation of Glivenko-Cantelli's theorem, it is controlled in the inequality of Dvoretzky-Kiefer-Wolfowitz.

Definition II.1 (Kolmogorov-Lévy distance). If W and Z are two real random variables, the Kolmogorov-Lévy distance between their distributions is defined by

$$\mathrm{d}_K(\mathcal{L}(W),\mathcal{L}(Z)) = \sup_{x\in\mathbb{R}} |\mathbb{P}\{W \leq x\} - \mathbb{P}\{Z \leq x\}| \ .$$

The fact that the Kolmogorov-Lévy distance metrizes the weak convergence towards a limiting distribution with a continuous cumulative distribution function is checked with the argument used in the classical proofs of Glivenko-Cantelli's theorem.

If $\lim_n x_n x_n = x$, the mass sequence δ_{x_n} converges weakly towards the mass δ_x, however, if $x_n \neq x$, $\mathrm{d}_K(\delta_{x_n}, \delta_x) = 1$.

To metrize weak convergence in general settings, we can use the Levy-Prokhorov distance:

$$\inf\Big\{\epsilon : \epsilon > 0, \qquad P(-\infty, x-\epsilon] - \epsilon \leq Q(-\infty, x] \leq P(-\infty, x-\epsilon] + \epsilon\Big\} \, .$$

The Kolmogorov-Lévy distance represents a relevant relaxation of the total variation distance (which does not metrize convergence in distribution). The fact that the test functions which define the Kolmogorov-Lévy distance are not absolutely continuous does not make life any easier. That is why we often work with a distance defined by more user-friendly functions. This distance belongs to the family of distances known as Monge-Wasserstein distances, or transportation distance.

Wasserstein distance

Let W and Z be $\mathbb{R}^k$-valued random variables with distributions $\mathcal{L}(W)$,

$$\mathrm{d}_M(\mathcal{L}(W), \mathcal{L}(Z)) = \sup_{h:1-\text{Lipschitz}} |\mathbb{E}h(W) - \mathbb{E}h(Z)| \ .$$

When we want to evaluate the distance with respect to an absolutely continuous distribution like the Gaussian distribution, the use of one metric over another is not very important.

Si P et Q sont deux lois sur $\mathbb{R}$, et si Q possède une densité majorée par C vis-à-vis de la mesure de Lebesgue, alors

$$\mathrm{d}_K(P, Q) \leq \sqrt{2C\mathrm{d}_M(P, Q)} \, .$$

Proof. Pour $x \in \mathbb{R}$, on note $h_x = \mathbb{I}_{(-\infty,x]}$ et pour $\epsilon > 0$, on définit $h_{x,\epsilon}$ par

$$h_{x,\epsilon}(w) = \begin{cases} 1 & \text{si } x \geq w \\ 0 & \text{si } w > x + \epsilon \\ 1 - \frac{w-x}{\epsilon} & \text{si } x \leq w \leq x + \epsilon \end{cases}$$

On note q la densité de Q par rapport à la mesure de Lebesgue. La fonction $h_{x,\epsilon}$ est $1/\epsilon$-Lipschitz et elle approche h_x. Si $W \sim P$ et $Z \sim Q$,

$$\mathbb{E}h_x(W) - \mathbb{E}h_x(Z) = \underbrace{\mathbb{E}h_x(W) - \mathbb{E}h_{x,\epsilon}(Z)}_{\text{(i)}} + \underbrace{\mathbb{E}h_{x,\epsilon}(Z) - \mathbb{E}h_x(Z)}_{\text{(II)}} \ .$$

On majore ensuite (simplement) les deux expressions (i) et (ii).

$$\begin{aligned} \text{(i)} &\leq \mathbb{E}h_{x,\epsilon}(W) - \mathbb{E}h_{x,\epsilon}(Z) \\ &\leq \frac{1}{\epsilon}\mathrm{d}_M(P, Q) \end{aligned}$$

et

$$\begin{aligned} \text{(ii)} &\leq \int_x^{x+\epsilon} \left(1 - \frac{w-x}{\epsilon}\right) q(z)\mathrm{d}z \\ &\leq C\frac{\epsilon}{2} \, . \end{aligned}$$

En choisissant $\epsilon = \sqrt{2\mathrm{d}_M(P, Q)/C}$, et en optimisant le choix de x, on obtient le résultat désiré.

□ □

II.5 Stein's method

In Stein's approach, we are interested in the approximation of a target law (normal law, Fish, χ^2_k, ...) by the law of a random variable whose mode of manufacturing (sum of independent random variables, order statistics, logarithm of likelihood ratio,...). The target law is characterized by an identity. In the case of of normal law, it's Stein's identity.

Stein's Identity

Let the real random variable Z be distributed according to Q. if for every absolutely continuous function f such that $f'(Z)$ is Q-integrable, we have

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$$

then

$$Q = \mathcal{N}(0,1)\,.$$

The converse is true.

Proof. See Lemma 12.1 and **?@lem-steinslemma-bis**. □

This identity leads us to define an operator $\mathcal{A}$ that acts on absolutely continuous functions: $\mathcal{A}f(x) = f'(x) - xf(x)$. This operator allows us to characterize $\mathcal{N}(0,1)$:

$$Q = \mathcal{N}(0,1) \iff \forall f \text{a.c} \quad 0 = \int \mathcal{A}f(z)Q(\mathrm{d}z)\,.$$

Now, for function $h \in \mathcal{H}$, si on peut définir f_h tel que

$$\mathcal{A}f_h(w) = h(w) - \int h(z)\phi(z)\mathrm{d}z\,,$$

on a

$$\mathbb{E}_Q h(W) - \mathbb{E}_{\mathcal{N}(0,1)}h(Z) = \mathbb{E}_Q \mathcal{A}f_h(W)\,.$$

In order to upper bound $d_{\mathcal{H}}(Q, \mathcal{N}(0,1))$, it suffices to upper bound the following collection of expectations:

$$\mathbb{E}_Q\mathcal{A}f_h(W) \qquad \text{for} h \in \mathcal{H}, \text{ and } \mathcal{A}f_h(w) = h(w) - \int h(z)\phi(z)\mathrm{d}z\,.$$

This last goal led to the development of creative methods for approximating integrals. In order to understand what is at stake, we have to investigate the smoothness and boundedness properties of functions $\mathcal{A}f_h, h \in \mathcal{H}$, when $\mathcal{H}$ itself is defined by smoothness constraints.

In the sequel, we abbreviate $\int_{\mathbb{R}} h(z)\phi(z)\mathrm{d}z$ to $\Phi(h)$.

Lemma 11.1. *Let f_h be the solution of the differential equation*

$$f_h'(w) - wf_h(w) = h(w) - \Phi(h)\,.$$

given by

$$f_h(w) = \begin{cases} \mathrm{e}^{w^2/2}\int_w^\infty \mathrm{e}^{-t^2/2}(\Phi(h) - h(t))\mathrm{d}t & \text{if } w \geq 0 \\ \mathrm{e}^{w^2/2}\int_{-\infty}^w \mathrm{e}^{-t^2/2}\left(h(t) - \Phi(h)\right)\mathrm{d}t & \text{if } w \leq 0\,. \end{cases}$$

If h is absolutely continuous then f_h is the only bounded solution of the differential equation and

$$\|f_h\|_\infty \leq 2\|h'\|_\infty, \qquad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}}\|h'\|_\infty, \qquad \|f_h''\|_\infty \leq 2\|h'\|_\infty\,.$$

Proof. We first check that f_h is the unique bounded solution of the differential equation,

Note that adding a constant to h leaves $h - \Phi(h)$ invariant. So we assume $h(0) = 0$. We have $|h(w)| \leq \|h'\|_\infty |w|$.

For $w > 0$, observe

$$e^{w^2/2} \int_w^\infty t\, e^{-t^2/2} dt = 1 \qquad \Phi(|h|) \leq \|h'\|_\infty \sqrt{\frac{2}{\pi}}\,.$$

Function $w \mapsto e^{w^2/2}$ satisfies the differential equation. Now look for a bounded solution like

$$f(w) = g(w) e^{w^2/2}\,.$$

The differential equation reads

$$g'(w) = e^{-w^2/2} \left(h(w) - \Phi(h)\right)\,.$$

A solution is provied by function

$$g(w) = \begin{cases} \int_w^\infty e^{-t^2/2} \left(\Phi(h) - h(t)\right) dt & \text{if } w \geq 0 \\ \int_{-\infty}^w e^{-t^2/2} \left(h(t) - \Phi(h)\right) dt & \text{if } w \leq 0\,. \end{cases}$$

(integrability assumptions about h warrant that g is well-defined), hence f_h is a solution of the differential equation stated in Lemma Lemma II.1

Besides, for $w > 0$

$$f_h(w) = g(w) e^{w^2/2} \leq \|h'\|_\infty \left(\sqrt{\frac{2}{\pi}} \frac{\overline{\Phi}(w)}{\phi(w)} + 1 \right)\,,$$

the ratio $\frac{\overline{\Phi}(w)}{\phi(w)}$ is non-increasing over $[0,\infty)$, it is always smaller than $\sqrt{\frac{\pi}{2}}$. The boundedness of f_h follows:

$$\|f_h\|_\infty \leq 2\|h'\|_\infty\,.$$

The other solutions of the differential equation are $f_h + c e^{w^2/2}$. They are not bounded.

Les us now bound the first and second derivatives of f_h.

We first check

$$h(w) - \Phi(h) = \int_{-\infty}^w h'(t)\Phi(t) dt - \int_w^\infty h'(t)\overline{\Phi}(t) dt\,.$$

Plugging into the definition de f_h, we get :

$$\begin{aligned} f_h(w) = & -\sqrt{2\pi} e^{w^2/2} (1 - \Phi(w)) \int_{-\infty}^w h'(t)\Phi(t) dt \\ & - \sqrt{2\pi} e^{w^2/2} \Phi(w) \int_w^\infty h'(t)(1 - \Phi(t)) dt \end{aligned}$$

$$\begin{aligned} f_h'(w) & = w f_h(w) + h(w) - \Phi(h) \\ & = (1 - \sqrt{2\pi} e^{w^2/2} (1 - \Phi(w))) \int_{-\infty}^w h'(t)\Phi(t) dt \\ & \quad - (1 + \sqrt{2\pi} e^{w^2/2} \Phi(w)) \int_w^\infty h'(t)(1 - \Phi(t)) dt \end{aligned}$$

hence

$$
\begin{aligned}
|f_h'(w)| \leq \|h'\|_\infty \sup_{w\in\mathbb{R}} \Big(& |1-\sqrt{2\pi}\mathrm{e}^{w^2/2}(1-\Phi(w))| \int_\infty^w \Phi(t)\mathrm{d}t \\
& |1+\sqrt{2\pi}\mathrm{e}^{w^2/2}\Phi(w)| \int_w^\infty (1-\Phi(t)) \Big)
\end{aligned}
$$

To complete the upper bound on $\|f_h'\|_\infty$, it suffices to bound the supremum as w varies.

To establish existence and boundedness of the second derivative, we differentiate $f_h'(w) = wf_h(w) + h(w) - \Phi(h)$ to get

$$
\begin{aligned}
f_h''(w) &= f_h(w) + wf_h'(w) + h'(w) \\
&= (1+w^2)f_h(w) + ww(h(w) - \Phi(h)) + h'(w) \,.
\end{aligned}
$$

On réutilise les calculs esquissés précédemment pour majorer $\|f_h''\|_\infty$.

□ □

Le fait que si h est 1-Lipschitz, f_h soit deux fois dérivable et de dérivée seconde bounded is convenient lorsqu'il faut majorer $\mathbb{E}\mathcal{A}f_h(W)$. La façon dont on peut tirer avantage de cette régularité dépend de ce qu'on sait sur la structure de W. Le theorem suivant qui peut être vu comme une combinaison des theorems de Berry-Esseen et de Lindeberg-Feller, montre ce qui peut etre obtenu lorsque W est une somme de variables aléatoires indépendantes pas nécessairement identique distribuées, centrées et suffisamment intégrables.

II.6 A Berry-Esseen bound for a Lindeberg-Feller-like Central Limit Theorem

Le theorem suivant peut etre lu comme un résultat intermédiaire entre le theorem de Berry-Esseen **?@thm-tcl-berry** (il donne une distance à la loi normale) et le theorem central limite de Lindeberg-Feller **?@thm-tcl-lindeberg** (il traite de sommes de variables aléatoires indépendantes mais pas nécessairement identiquement distribuées). Les conditions utilisées ici sont plus fortes que celles décrites dans les deux theorems précédents. Dans le theorem de Lindeberg-Feller, la condition d'intégrabilité uniforme est minimale. Dans l'énoncé du theorem de Berry-Esseen, on se contente de postuler que les summands have a finite third moment. Ici on suppose un peu plus: les summands ont un kurtosis bounded.

Rappelons que le kurtosis d'une loi est défini par

$$
\frac{\mathbb{E}[(X-\mathbb{E}X)^4]}{(\mathbb{E}[(X-\mathbb{E}X)^2])^2} \,.
$$

The kurtosis of Gausian est toujours égal à 3. Pour les lois Gamma, le kurtosis ne dépend que du paramètre de forme p (il vaut $3+6/p$).

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes centrées, de kurtosis majoré par κ. On définit $\sigma^2 = \sum_{i=1}^n \operatorname{var}(X_i)$ et $W = \sum_{i=1}^n X_i/\sigma$.

$$
\mathrm{d}_M(\mathcal{L}(W), \mathcal{N}(0,1)) \leq \sqrt{\frac{2}{\pi}} \left(\frac{1}{\sigma^3} \sum_{i=1}^n \mathbb{E}|X_i|^3 + \sqrt{\kappa \sum_{i=1}^n \frac{\operatorname{var}(X_i)^2}{\left(\sum_{i=1}^n \operatorname{var}(X_i)\right)^2}} \right) \,.
$$

Proof of ?@thm-lindeberg-feller-stein. Dans la preuve h est une fonction 1-Lipschitzienne, f la solution bounded $\mathcal{A}f = h - \Phi(h)$. Pour majorer $\mathbb{E}h(W) - \Phi(h)$, on va majorer $|\mathbb{E}[f_h'(W) - Wf(W)]|$.

Dans la suite, pour $i \in 1, \ldots, n$, $W_i = \sum_{j \neq i} X_j/\sigma$.

Observe first that as W_i and X_i are independent and centered

$$\mathbb{E}[X_i f(W_i)] = 0\,.$$

$$\begin{aligned}\mathbb{E}[Wf(W)] &= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i f(W)\right] \\ &= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n (X_i f(W) - X_i f(W_i))\right] \\ &= \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i (f(W) - f(W_i))\right]\,.\end{aligned}$$

In the sequel, we will rely on

$$\begin{aligned}\mathbb{E}[Wf(W)] = \mathbb{E}&\left[\frac{1}{\sigma}\sum_{i=1}^n X_i(f(W) - f(W_i) - (W-W_i)f'(W))\right] \\ &+ \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i((W-W_i)f'(W))\right]\,.\end{aligned}$$

$$\begin{aligned}|\mathbb{E}[f'(W) - Wf(W)]| \leq &\underbrace{\left|\mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n X_i(f(W) - f(W_i) - (W-W_i)f'(W))\right]\right|}_{\text{(i)}} \\ &+ \underbrace{\left|\mathbb{E}\left[\left(1 - \frac{1}{\sigma}\sum_{i=1}^n X_i(W-W_i)\right)f'(W)\right]\right|}_{\text{(i)}}\,.\end{aligned}$$

To upper bound (i), l'inégalité des accroissements finis suffit

$$\begin{aligned}\text{(i)} &\leq \mathbb{E}\left[\frac{1}{\sigma}\sum_{i=1}^n |X_i|\,|f(W) - f(W_i) - (W-W_i)f'(W)|\right] \\ &\leq \frac{1}{\sigma}\sum_{i=1}^n \mathbb{E}\left[|X_i|\frac{|X_i|^2}{\sigma^2}\|f''\|_\infty\right] \\ &\leq \|f''\|_\infty \frac{\sum_{i=1}^n \mathbb{E}|X_i|^3}{\sigma^3}\,.\end{aligned}$$

Pour majorer (ii), on utilise Cauchy-Schwarz et l'hypothèse de kurtosis.

$$
\begin{aligned}
\text{(ii)} &\leq \mathbb{E}\left[\left|\left(1-\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2\right)\right| |f'(W)|\right] \\
&\leq \|f'\|_\infty \mathbb{E}\left[\left|\left(1-\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2\right)\right|\right] \\
&\leq \|f'\|_\infty \frac{1}{\sigma^2}\left(\sum_{i=1}^n \operatorname{var}(X_i^2)\right)^{1/2} \\
&\leq \|f'\|_\infty \frac{1}{\sigma^2}\left(\kappa\sum_{i=1}^n \operatorname{var}(X_i)^2\right)^{1/2} \\
&\leq \|f'\|_\infty \left(\kappa\sum_{i=1}^n \left(\frac{\operatorname{var}(X_i)}{\sum_{i=1}^n \operatorname{var}(X_i)}\right)^2\right)^{1/2}
\end{aligned}
$$

□ □

En utilisant le fait que pour X centrée, de kurtosis κ

$$
\mathbb{E}|X_i|^3 \leq (\mathbb{E}|X_i|^4)^{3/4} \leq \kappa^{3/4}(\mathbb{E}|X_i|^2)^{3/2} ,
$$

le membre droit du majorant peut se majorer lui même en

$$
\kappa^{3/4}\|f''\|_\infty \left(\sum_{i=1}^n \left(\frac{\operatorname{var}(X_i)}{\sum_{j=1}^n \operatorname{var}(X_j)}\right)^{3/2}\right) + \kappa^{1/2}\|f'\|_\infty \left(\sum_{i=1}^n \left(\frac{\operatorname{var}(X_i)}{\sum_{i=1}^n \operatorname{var}(X_i)}\right)^2\right)^{1/2}
$$

Si les X_i sont identiquement distribuées, alors le majorant du theorem s'écrit

$$
\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}}\left(\kappa^{3/4}+\kappa^{1/2}\right) .
$$

11.7 Bibliographic remarks

La possibilité de métriser la convergence en distribution (et la convergence en probabilité) est traitée avec beaucoup de rigueur et de clarté dans (Dudley, 2002).

Stein's method pour établir des versions précises et générales du theorem central limite est décrite dans (N. Ross, 2011).

A thorough yet readable treatment of Stein's method is (Chen et al., 2011).

[Section]) follows the first pages of (N. Ross, 2011)

La démonstration complète du **?@lem-regu-stein**} se trouve dans (Chen et al., 2011).

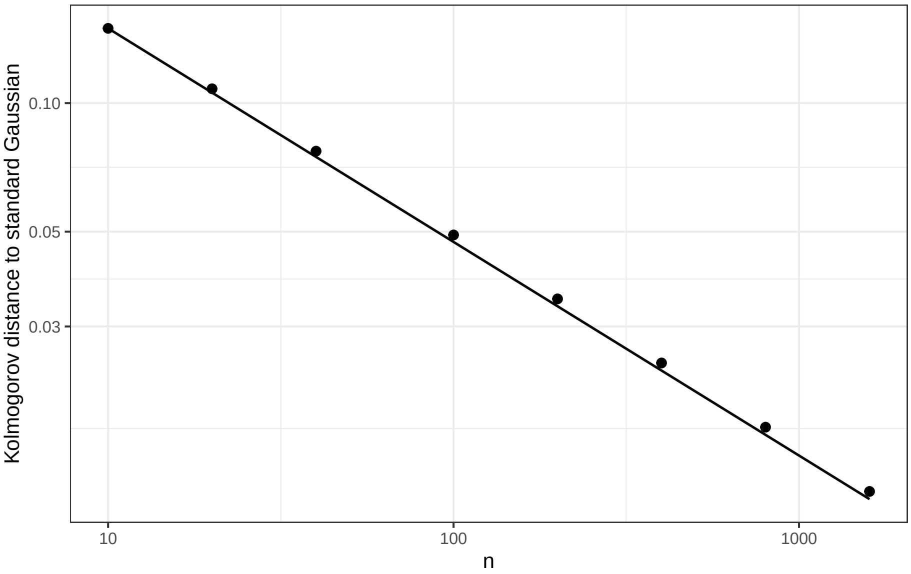

Figure 11.2: Kolmogorov distance between standard Gaussian and distribution of $(X_n - np)/\sqrt{np(1-p}$ with X_n binomially distributed with parameters n and $p = .3$. The rate of convergence in the central limit theorem is asymptotically of order $1/\sqrt{n}$.

Chapter 12

Gaussian vectors

12.1 Univariate Gaussian distribution

The standard Gaussian density is denoted by ϕ:

$$\phi(x) = \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \, .$$

The corresponding cumulative distribution function is denoted by Φ. The survival function $1 - \Phi$ is denoted by $\overline{\Phi}$.

$$\Phi(x) = \int_{-\infty}^{x} \phi(u) \mathrm{d}u \, .$$

Let is denote by $\mathcal{N}(0,1)$ (expectation 0, variance 1) the standard Gaussian probability distribution, that is the probability distribution defined by ϕ.

Any affine transform of a standard Gaussian random variable is distributed according to a univariate Gaussian distribution. If $X \sim \mathcal{N}(0,1)$ then $\sigma X + \mu \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with density $\frac{1}{\sigma}\phi\left(\frac{\cdot-\mu}{\sigma}\right)$, cumulative distribution function $\Phi\left(\frac{\cdot-\mu}{\sigma}\right)$.

The standard Gaussian distribution is characterized by the next identity.

Lemma 12.1 (Stein's Lemma)**.** ***Let** $X \sim \mathcal{N}(0,1)$**, let** g **be an absolutely continuous function with derivative** g' **such that** $\mathbb{E}[|Xg(X)|] < \infty$**, then** $g'(X)$ **is integrable and***

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)] \, .$$

Proof. The proof relies on integration by parts. First note that replacing g by $g - g(0)$ changes neither g', nor $\mathbb{E}[Xg(X)]$. We may assume that $g(0) = 0$.

$$
\begin{aligned}
\mathbb{E}[Xg(X)] &= \int_{\mathbb{R}} xg(x)\phi(x)\mathrm{d}x \\
&= \int_0^\infty xg(x)\phi(x)\mathrm{d}x + \int_{-\infty}^0 xg(x)\phi(x)\mathrm{d}x \\
&= \int_0^\infty x \int_0^\infty g'(y)\mathbb{I}_{y\leq x}\mathrm{d}y\phi(x)\mathrm{d}x - \int_{-\infty}^0 x \int_{-\infty}^0 g'(y)\mathbb{I}_{y\geq x}\mathrm{d}y\phi(x)\mathrm{d}x \\
&= \int_0^\infty g'(y) \int_0^\infty \mathbb{I}_{y\leq x} x\phi(x)\mathrm{d}x\mathrm{d}y - \int_{-\infty}^0 g'(y)\int_{-\infty}^0 x\phi(x)\mathbb{I}_{y\geq x}\mathrm{d}x\mathrm{d}y \\
&= \int_0^\infty g'(y) \int_y^\infty x\phi(x)\mathrm{d}x\mathrm{d}y - \int_{-\infty}^0 g'(y) \int_{-\infty}^y x\phi(x)\mathrm{d}x\mathrm{d}y \\
&= \int_0^\infty g'(y)\phi(y)\mathrm{d}y - \int_{-\infty}^0 -g'(y)\phi(y)\mathrm{d}y \\
&= \int_{-\infty}^\infty g'(y)\phi(y)\mathrm{d}y\,.
\end{aligned}
$$

The last inequality is justified by Tonelli-Fubini's Theorem. Then, we rely on $\phi'(x) = -x\phi(x)$. □

The characteristic function is a very efficient tool when handling Gaussian distributions.

Proposition 12.1. *The characteristic function of* $\mathcal{N}(\mu, \sigma^2)$ *is*

$$
\widehat{\Phi}(t) := \mathbb{E}\left[\mathrm{e}^{\imath tX}\right] = \mathrm{e}^{\imath t\mu - \frac{t^2\sigma^2}{2}}\,.
$$

Proof. It is enough to check the proposition for $\mathcal{N}(0,1)$. As ϕ is even,

$$
\begin{aligned}
\widehat{\Phi}(t) &= \int_{-\infty}^\infty \mathrm{e}^{\imath tx}\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x \\
&= \int_{-\infty}^\infty \cos(tx)\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x\,.
\end{aligned}
$$

Derivation with respect to t, interchanging derivation and expectation (why can we do that?)

$$
\widehat{\Phi}'(t) = \int_{-\infty}^\infty -x\sin(tx)\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x\,.
$$

Now relying on Stein's Identity with $g(x) = -\sin(tx)$ and $g'(x) = -t\cos(tx)$

$$
\begin{aligned}
\widehat{\Phi}'(t) &= -t\int_{-\infty}^\infty \cos(tx)\phi(x)\mathrm{d}x \\
&= -t\widehat{\Phi}(t)\,.
\end{aligned}
$$

We immediately get $\widehat{\Phi}(0) = 1$, and solving the differential equation leads to

$$
\log\widehat{\Phi}(t) = -\frac{t^2}{2}\,.
$$

□

The fact that the characteristic function completely defines the probability distribution provides us with a converse of Lemma 12.1.

Lemma 12.2 (Stein's Lemma (bis)). *Let X be a real-valued random variable on some probability space. If, for any differentialle function g such that g' and $x \mapsto xg(x)$ are integrable, the following holds*

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)]$$

then the distribution of X is standard Gaussian

Proof. Consider the real $\widehat{F}$ and the imaginary part $\widehat{G}$ of the characteristic function of the distribution pf X, the identity entails that $\widehat{F}'(t) = -t\widehat{F}(t)$ and $\widehat{G}'(t) = -t\widehat{G}(t)$ with $\widehat{F}(0) = 1$ and $\widehat{G}(0) = 0$. Solving the two differential equations leads to $\widehat{F}(t) = \mathrm{e}^{-t^2/2}$ and $\widehat{G}(t) = 0$. We just checked that the characteristic function of the distribution of X is the characteristic function of $\mathcal{N}(0,1)$. □

It is now easy to check that the distribution of the sum of two independent Gaussian random variables is a Gaussian random variable.

If X and Y are two independent random variables distributed according to $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\mu', \sigma'^2)$ then $X + Y$ is distributed according to $\mathcal{N}(\mu + \mu', \sigma^2 + \sigma'^2)$.

Check and justify.

The moment generating function of a Gaussian random variable is given by

$$s \mapsto \mathbb{E}\left[\mathrm{e}^{sX}\right] = \mathrm{e}^{\frac{s^2}{2}} .$$

From Markov's inequality, we obtain interesting upper bounds on the Gaussian tail function. Some calculus allows us to refine the tail bounds

Proposition 12.2 (Tail probabilities for Gaussian distribution). ***For** $x \geq 0$,*

$$\frac{\phi(x)}{x}\left(1 - \frac{1}{x^2}\right) \leq \overline{\Phi}(x) \leq \min\left(\mathrm{e}^{-\frac{x^2}{2}}, \frac{\phi(x)}{x}\right) .$$

Proof. The proof boils down to repeated integration by parts.

$$\begin{aligned}
\overline{\Phi}(x) &= \int_x^\infty \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u \\
&= \left[-\frac{1}{\sqrt{2\pi}u}\mathrm{e}^{-\frac{u^2}{2}}\right]_x^\infty - \int_x^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{u^2}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u.
\end{aligned}$$

As the second term is non-positive,

$$\overline{\Phi}(x) \leq \left[-\frac{1}{\sqrt{2\pi}u}\mathrm{e}^{-\frac{u^2}{2}}\right]_x^\infty = \frac{\phi(x)}{x}.$$

This is the first part of the right-hand inequality, the other part comes from Markov's inequality. For the left-hand inequality, we have to upper bound $\int_x^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{u^2}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u$.

$$\begin{aligned}
\int_x^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{u^2}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u &= \left[\frac{-1}{\sqrt{2\pi}}\frac{1}{u^3}\mathrm{e}^{-\frac{u^2}{2}}\right]_x^\infty - \int_x^\infty \frac{1}{\sqrt{2\pi}}\frac{3}{u^4}\mathrm{e}^{-\frac{u^2}{2}}\mathrm{d}u \\
&\leq \frac{1}{\sqrt{2\pi}}\frac{1}{x^3}\mathrm{e}^{-\frac{x^2}{2}}.
\end{aligned}$$

□

Proposition 12.3 (Moments). *For a standard Gaussian random variable,*

$$\mathbb{E}\left[X^k\right] == \begin{cases} 0 & \textit{if } k \textit{ is odd} \\ \frac{k!}{2^{k/2}(k/2)!} = \frac{\Gamma(k+1)}{2^{k/2}\Gamma(k/2+1)} & \textit{if } k \textit{ is even.} \end{cases}$$

Proof. Thanks to distributional symmetry, $\mathbb{E}\left[X^k\right] = 0$ for all odd$k$. We handle even powers using integration by parts:

$$\mathbb{E}\left[X^{k+2}\right] \quad = \quad (k+1)\mathbb{E}\left[X^k\right] \, .$$

Induction on k leads to,

$$\mathbb{E}\left[X^{2k}\right] \quad = \quad \prod_{j=1}^{k}(2j-1) = \frac{(2k)!}{2^k k!} \, .$$

□

Note that $(2k)!/(2^k k!)$ is also the number of partitions of $\{1, \ldots, 2k\}$ into subsets of cardinality 2.

The *skewness* is null, the kurtosis (ratio of fourth centred moment over squared variance equals 3:

$$\mathbb{E}[X^4] = 3 \times \mathbb{E}[X^2]^2 \, .$$

12.2 Gaussian vectors

A Gaussian vector is a collection of univariate Gaussian random variables that satisfies a very stringent property:

Definition 12.1 (Gaussian Vector). A random vector $(X_1, \ldots, X_n)^T$ is a ***Gaussian vector*** iff for any real vector $(\lambda_1, \lambda_2, \ldots, \lambda_n)^T$, the distribution of the univariate random variable $\sum_{i=1}^n \lambda_i X_i$ is Gaussian.

Not every collection of Gaussian random variables forms a Gaussian vector.

The random vector $(X, \epsilon X)$ with $X \sim \mathcal{N}(0.1)$, independent of ϵ which is worth ± 1 with probability $1/2$, is not a Gaussian vector although both X and ϵX are univariate Gaussian random variables.

Check that ϵX is a Gaussian random variable.

Yet there are Gaussian vectors! A simple way to obtain a Gaussian vector is provided by the next proposition (checked by a characteristic function argument).

If $X_1, \ldots, X_n$ is a sequence of independent Gaussian random variables, then $(X_1, \ldots, X_n)^t$ is a Gaussian vector.

In the sequel, a ***standard Gaussian vector*** is a random vector with independent coordinates with each coordinate distributed according to $\mathcal{N}(0, 1)$.

We will see how to construct general Gaussian vectors. Before this, let us check that the joint distribution of a Gaussian random vector is completely characterized by its covariance matrix and its expectation vector.

Recall that the ***covariance*** of random vector $X = (X_1, \ldots, X_n)^T$ is the matrix K with dimension $n \times n$ with coefficients

$$K[i,j] = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

Without loss of generality, we may assume that random vector X is centered For every $\lambda = (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{R}^n$, we have:

$$\mathrm{var}(\langle \lambda, X\rangle) = \lambda^t K \lambda = \mathrm{trace}(K\lambda\lambda^t)$$

(this is does not depend on any Gaussianity assumption).

Indeed,

$$\begin{array}{rcl} \mathrm{var}(\langle \lambda, X\rangle) & = & \mathbb{E}\left[\left(\sum_{i=1}^n \lambda_i X_i\right)^2\right] \\ & = & \sum_{i,j=1}^n \mathbb{E}\left[\lambda_i\lambda_j X_i X_j\right] \\ & = & \sum_{i,j=1}^n \lambda_i \lambda_j K[i,j] \\ & = & \lambda^t K \lambda\,. \end{array}$$

The characteristic function of a Gaussian vector X with expectation vector μ and covariance K satisfies

$$\mathbb{E}\mathrm{e}^{\imath\langle\lambda,X\rangle} = \mathrm{e}^{\imath\langle\lambda,\mu\rangle - \frac{\lambda^t K\lambda}{2}}\ .$$

A linear transform of a Gaussian vector is a Gaussian vector.

If $Y = (Y_1, \ldots, Y_n)^T$ is a Gaussian vector with covariance K and A a real matrix with dimensions $p \times n$, then $A \times Y$ is Gaussian vector with expectation $A \times \mathbb{E}Y$ and covariance matrix

$$AKA^T.$$

Proof. Without loss of generality, we assume Y is centred.

For any $\lambda \in \mathbb{R}^p$, $\langle \lambda, AY\rangle = \langle A^T\lambda, Y\rangle$, thus $A \times Y$ is Gaussian with variance

$$\lambda^T AKA^T\lambda\,.$$

The covariance of $A \times Y$ is determined by this observation. □

To manufacture Gaussian vectors with general covariance matrices, we rely on an important notion from matrix analysis.

Definition 12.2 (Semi-Definite Positive matrices)**.** A symmetric matrix M with dimensions $k \times k$ is Definite Positive (respectively Semi-Definite Positive) iff, for any non-null vector $v \in \mathbb{R}^k$,

$$v^T M v > 0 \qquad (\text{resp.} \qquad v^T M v \geq 0)\,.$$

We denote by $\mathsf{dp}(k)$ (resp. $\mathsf{sdp}(k)$), the cones of Definite Positive (resp. Semi-Definite Positive) matrices.

If K is the covariance matrix of a random vector, K is symmetric, Semi-Definite Positive.

Proof. If X is a $\mathbb{R}^k$-valued random vector, with covariance K, for any vector $\lambda \in \mathbb{R}^n$,

$$\lambda^T K \lambda = \sum_{i,j \leq k} K_{i,j} \lambda_i \lambda_j = \operatorname{cov}(\langle \lambda, X \rangle, \langle \lambda, X \rangle)$$

soit $\lambda^T K \lambda = \operatorname{var}(\langle \lambda, X \rangle)$. The variance of a univariate random variable is always non-negative. □

The next observation is the key to the construction to general Gaussian vectors.

Proposition 12.4 (Cholesky's factorization). *If A is a Semi-definite Positive symmetric matrix then there exists (at least) a real matrix B such that $A = B^T B$.*

We do not check this proposition. This is a basic Theorem from matrix analysis. It can be established from the ***spectral decomposition theorem*** for symmetric matrices. It can also be established by a simple constructive approach: a positive definite matrix K admits a ***Cholesky decomposition***, in other words, there exists a triangular matrix lower than L such that $K = L \times L^T$.

The next proposition is a corollary of the general formula for image densities.

If A is a symmetric positive definite matrix ($A \in \mathtt{dp}(n)$), then the distribution of the centred Gaussian vector with covariance matrix A is absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}^n$:

$$\frac{1}{(2\pi)^{n/2} \det(A)^{1/2}} \exp\left(- \frac{x^t A^{-1} x}{2} \right) .$$

Proof. The density formula is trivially correct for standard Gaussian vectors. For the general case, it is enough to invoke the image density formula to the image of the standard Gaussian vector by the bijective linear transformation defined by the Cholesky factorization of A. The determinant of the Cholesky factor is the square root of the determinant of A. □

Is the distribution of a Gaussian vector X with ***singular*** covariance matrix absolutely continuous with respect to Lebesgue measure?

Definition 12.3 (Gaussian space). If $X = (X_1, \ldots, X_n)^T$ is a centered Gaussian vector with covariance matrix K , the set $\left\{ \sum_{i=1}^n \lambda_i X_i = \langle \lambda, X \rangle ; \lambda \in \mathbb{R}^n \right\}$ is the Gaussian space generated by $X = (X_1, \ldots, X_n)^T)$.

The Gaussian space is a real vector space. If $(\Omega, \mathcal{F}, P)$ denotes the probability space, X lives on, the Gaussian space is a subspace of $L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$. It inherits the inner product structure from $L^2_{\mathbb{R}}(\Omega, \mathcal{F}, P)$.

This inner-product is completely defined by the covariance matrix K.

$$\begin{aligned}
\left\langle \sum_{i=1}^n \lambda_i X_i, \sum_{i=1}^n \lambda'_i X_i \right\rangle &\equiv \mathbb{E}_P \left[\left(\sum_{i=1}^n \lambda_i X_i \right) \left(\sum_{i=1}^n \lambda'_i X_i \right) \right] \\
&= \sum_{i,i'=1}^n \lambda_i \lambda'_{i'} K[i, i'] \\
&= (\lambda_1, \ldots, \lambda_n) K \begin{pmatrix} \lambda'_1 \\ \vdots \\ \lambda'_n \end{pmatrix} \\
&= \operatorname{trace} \left(K \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \begin{pmatrix} \lambda'_1 & \ldots & \lambda'_n \end{pmatrix} \right) .
\end{aligned}$$

Different Gaussian vectors may generate the same Gaussian space. Explain how and why.

Gaussian spaces enjoy remarkable properties. Independence of random variables belonging to the same Gaussian space may checked very easily.

Proposition 12.5. *Two random variables Z and Y, belonging to the same Gaussian space, are independent iff they are orthogonal (or decorrelated), that is iff*

$$\operatorname{Cov}_P[Y, Z] = \mathbb{E}_P[YZ] = 0.$$

Without loss of generality, we assume covariance matrix K is positive definite.

Proof. Independence always implies orthogonality.

Without loss of generality, we assume that the Gaussian space is generated by a standard Gaussian vector, let $Z = \sum_{i=1}^n \lambda_i X_i$ and $Y = \sum_{i=1}^n \lambda'_i X_i$.

If Z and Y are orthogonal (or non-correlated)

$$\mathbb{E}[ZY] = \sum_{i=1}^n \lambda_i \lambda'_i = 0.$$

To show that Z and Y are independent, it is enough to check that for all μ and μ' in $\mathbb{R}$

$$\mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\mathrm{e}^{\imath\mu' Y}\right] = \mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\right] \times \mathbb{E}\left[\mathrm{e}^{\imath\mu' Y}\right].$$

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\mathrm{e}^{\imath\mu' Y}\right] &= \mathbb{E}\left[\mathrm{e}^{\imath\mu \sum_i \lambda_i X_i}\mathrm{e}^{\imath\mu' \sum_i \lambda'_i X_i}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n \mathrm{e}^{\imath(\mu\lambda_i + \mu'\lambda'_i)X_i}\right] \\
&X_i \text{ are independent} \ldots \\
&= \prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{\imath(\mu\lambda_i + \mu'\lambda'_i)X_i}\right] \\
&= \prod_{i=1}^n \mathrm{e}^{-(\mu\lambda_i + \mu'\lambda'_i)^2/2} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n \mu^2\lambda_i^2 + 2\mu\mu'\lambda_i\lambda'_i + \mu'^2\lambda_i'^2\right) \\
&\text{orthogonality} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^n \mu^2\lambda_i^2 + \mu'^2\lambda_i'^2\right) \\
&\ldots \\
&= \mathbb{E}\left[\mathrm{e}^{\imath\mu Z}\right] \times \mathbb{E}\left[\mathrm{e}^{\imath\mu' Y}\right].
\end{aligned}$$

□

The next proposition is a direct consequence.

If E and E' are two linear sub-spaces of the Gaussian space generated by the Gaussian vector with independent coordinates $X_1, \ldots, X_n$, the (Gaussian) random variables belonging to subspace E and the random (Gaussian) variables belonging to the E' space are independent if and only these two subspaces are orthogonal.

12.3 Convergence of Gaussian vectors

Recall the Lévy continuity theorem (Theorem 10.2)), it relates weak convergence for probability measures and simple convergence for characteristic functions.

A sequence of probability distributions $(P_n)_{n\in\mathbb{N}}$ over $\mathbb{R}^k$ converges weakly towards P iff for every $\vec{s}\in\mathbb{R}^k$:

$$\mathbb{E}_{P_n}\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right] \to \mathbb{E}_{P}\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right]\,.$$

For every $\vec{s}\in\mathbb{R}^k$, functions $\vec{x}\mapsto\cos(\langle\vec{s},\vec{x}\rangle)$ and $\vec{x}\mapsto\sin(\langle\vec{s},\vec{x}\rangle)$ are continuous and bounded, They are also infinitely many times differentiable.

It is remarkable and useful that weak convergence can be checked on this small set of functions.

L▯vy-Continuity Theorem (bis)

A sequence of probability distributions $(P_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^k$ converges weakly towards a probability distribution iff there exists a function f over $\mathbb{R}^k$, continuous at $\vec{0}$, such that for all $\vec{s}\in\mathbb{R}^k$:

$$\mathbb{E}_{P_n}\left[\mathrm{e}^{\imath\langle\vec{s},\vec{X}\rangle}\right] \to f(\vec{s})\,.$$

Then, function f is the characteristic function of some probability distribution P.

The continuity condition at 0 is necessary. The characteristic function of a probability distribution is always continuous at 0. Continuity at 0 warrants the tightness of the sequence of probability distributions.

If a sequence of k-dimensional Gaussian vectors (X_n) is defined by a $\mathbb{R}^k$-valued sequence $(\vec{\mu}_n)_n$ and a $\mathsf{SDP}(k)$-valued sequence $(K_n)_n$ and

$$\begin{aligned}\lim_n \vec{\mu}_n &= \mu\in\mathbb{R}^k\\ \lim_n K_n &= K\in\mathsf{SDP}(k)\end{aligned}$$

then the sequence $(X_n)_n$ converges in distribution towards $\mathcal{N}(\vec{\mu},K)$ (if $K=0$, the limit distribution is δ_μ).

12.4 Gaussian conditioning

Let $(X_1,\ldots,X_n)^T$ be a Gaussian vector with distribution $\mathcal{N}(\mu,K)$ where $K\in\mathsf{DP}(n)$. The covariance matrix K is partitioned into blocks

$$K=\left[\begin{array}{cc} A & B^t\\ B & W\end{array}\right]$$

where $A\in\mathsf{DP}(k)$, $1\le k<n$, and $W\in\mathsf{DP}(n-k)$.

We are interested in the conditional expectation of $(X_1,\ldots,X_k)^T$ with repsect to $\sigma(X_{k+1},\ldots,X_n)$ and in the conditional distribution of $(X_1,\ldots,X_k)^T$ with respect to $\sigma(X_{k+1},\ldots,X_n)$.

The Schur complement of A in K is defined as

$$W-BA^{-1}B^T\,.$$

This definition makes sense for symmetric matrices when A is non-singular.

If $K \in \mathsf{DP}(n)$ then the Schur complement of A in K also belongs to $\mathsf{DP}(n-k)$

In the statement of the next theorems, $A^{-1/2}$ denotes the Cholesky factor of A^{-1}: $A^{-1} = A^{-1/2} \times (A^{-1/2})^T$.

L'espérance conditionnelle de $(X_{k+1}, \ldots, X_n)^t$ sachant $(X_1, \ldots, X_k)^t$ est une transformation affine de $(X_1, \ldots, X_k)^t$:

$$\mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \begin{matrix} X_1 \\ \vdots \\ X_k \end{matrix}\right] = \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + (BA^{-1}) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \right) .$$

The conditional distribution of $(X_{k+1}, \ldots, X_n)^T$ with respect to $\sigma(X_1, \ldots, X_k)$ is a Gaussian distribution whose expectation is the conditional expectation $(X_{k+1}, \ldots, X_n)^T$ with respect to $\sigma(X_1, \ldots, X_k)$ and whose variance is the Schur complement of the covariance of $(X_1, \ldots, X_k)^T$ in the covariance matrix of $(X_1, \ldots, X_n)^T$.

We will first study the conditional density, and, with a minimum amount of calculation, establish that it is Gaussian. Conditional expectation will be calculated as expectation under conditional distribution.

To characterize conditional density, we rely on a distributional representation argument (any Gaussian vector is distributed as the image of a standard Gaussian vector by an affine transformation) and a matrix analysis result that is at the core of the Cholesky factorization of positive semi-definite matrices.

$(X_1, \ldots, X_n)^T$ is distributed as the image of standard Gaussian vector by a block triangular matrix

, et utiliser des propriétés des lois conditionnelles pour établir à la fois les deux résultats.

Let K be a symmetric definite positive matrix with dimensions $n \times n$

$$K = \begin{bmatrix} A & B^t \\ B & W \end{bmatrix}$$

where A has dimensions $k \times k$, $1 \leq k < n$.

Then, the Schur-complement of A with respect to K

$$W - BA^{-1}B^t$$

is positive definite. Sub-matrices A and $W - BA^{-1}B^t$ both have a Cholesky decomposition \$A = L_1 L_1^t ,W - B A^{-1} B^t = L_2 L_2^t\$ where L_1, L_2 are lower triangular, and K's factorization reads like:

$$K = \begin{bmatrix} L_1 & 0 \\ B(L_1^t)^{-1} & L_2 \end{bmatrix} \times \begin{bmatrix} L_1^t & L_1^{-1}B^t \\ 0 & L_2^t \end{bmatrix} .$$

Proof. Without loss of generality, we check the statement on centered vectors. The Cholesky factorization of K allows us to write

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \begin{bmatrix} L_1 & 0 \\ B(L_1^t)^{-1} & L_2 \end{bmatrix} \times \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

where $(Y_1, \ldots, Y_n)^t$ is a centered standard Gaussian vector.

In the sequel, we assume $(X_1, \ldots, X_n)^T$ and $(Y_1, \ldots, Y_n)^T$ live on the same probability space. As L_1 is invertible, the σ-algebras generated by $(X_1, \ldots, X_k)^T$ and $(Y_1, \ldots, Y_k)^T$ are equal. We agree on $\mathcal{G} = \sigma(X_1, \ldots, X_k)$. The conditional expectations and conditional distributions also coincide .

$$\begin{aligned}
\mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \mathcal{G}\right] &= \mathbb{E}\left[B(L_1^t)^{-1}\begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} \mid \mathcal{G}\right] + \mathbb{E}\left[L_2\begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix} \mid \mathcal{G}\right] \\
&= B(L_1^t)^{-1}L_1^{-1}\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} = BA^{-1}\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} ,
\end{aligned}$$

car $(Y_{k+1}, \ldots, Y_n)^t$ is centered and independent from $\mathcal{G}$.

Note that residuals

$$\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} - \mathbb{E}\left[\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} \mid \mathcal{G}\right] = L_2\begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix}$$

are independent from $\mathcal{G}$. This is a Gaussian property. For general square integrable random variables, we may only assert that residuals are orthogonal to $\mathcal{G}$-measurable random variables.

The conditional distribution of $(X_{k+1}, \ldots, X_n)^T$ with respect to $(X_1, \ldots, X_k)^T$ coincides with the conditional distribution of

$$B(L_1^t)^{-1} \times \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} + L_2 \times \begin{pmatrix} Y_{k+1} \\ \vdots \\ Y_n \end{pmatrix}$$

conditionally on $(Y_1, \ldots, Y_k)^T$. As $(Y_1, \ldots, Y_k)^t = L_1^{-1}(X_1, \ldots, X_k)^T$, the conditional distribution we are looking for is Gaussian with expectation

$$BA^{-1} \times \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$$

(the conditional expectation) and variance $L_2 \times L_2^t = W - BA^{-1}B^t$. □

If $(X, Y)^T$ is a centered Gaussian vector with $\text{var}(X) = \sigma_x^2$, $\text{var}(Y) = \sigma_y^2$ and $\text{cov}(X, Y) = \rho\sigma_x\sigma_y$, the conditional distribution of Y with respect to X is

$$\mathcal{N}\left(\rho\sigma_y/\sigma_x X, \sigma_y^2(1-\rho^2)\right) .$$

The quantity ρ is called the ***linear correlation coefficient*** between X and Y. By Cauchy-Schwarz's inequality, $\rho \in [-1, 1]$.

These two theorems are usually addressed in the order in which they are stated. Conditional expectation is characterized by adopting the L^2 (predictive) viewpoint: the conditional expectation of the random vector Y knowing X is defined as the best X-measurable predictor of the vector Y with respect to quadratic error (the random vector Z, X-measurable that minimizes $\mathbb{E}\left[\|Y - Z\|^2\right]$).

In order to characterize conditional expectation, we first compute the optimal affine predictor of $(X_{k+1}, \ldots, X_n)^T$ based on $(X_1, \ldots, X_k)^T$. This optimal affine predictor is

$$\begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + (BA^{-1}) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \right) ,$$

(if Gaussian vectors are centred, this amounts to determine the matrix P with dimensions $(n-k)\times k$ which minimizes trace$(PAP^t - 2BP^t)$). The optimal affine predictor is a Gaussian vector, one can check that the residual vector

$$\begin{pmatrix} X_{k+1} \\ \vdots \\ X_n \end{pmatrix} - \left\{ \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix} + (BA^{-1}) \times \left(\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \right) \right\}$$

is also Gaussian and orthogonal to the affine predictor. The residual vector is independent from the affine predictor.

This is enough to establish that the affine predictor is the orthogonal projection of $(X_{k+1}, \ldots, X_n)^T$ on the closed linear subspace of square-integrable $(X_1, \ldots, X_k)^T$-measurable random vectors.

This proves that the affine predictor is the conditional expectation.

In the notes, we deal with a special case of linear conditioning.

To fugure out general linear conditioning, consider $X \sim \mathcal{N}(0, K)$ (we assume centering to alleviate notation and computations, translating does not change the relevant σ-algebras and thus conditioning), where $K \in \mathsf{DP}(n)$, and a linear transformation defined by matrix H with dimensions $m \times n$. H is assumed to have rank m. Agree on $Y = HX$. Considering the Gaussian vector $[X^T : Y^T]$ with covariance matrix

$$\begin{bmatrix} K & KH^t \\ HK & HKH^t \end{bmatrix}$$

and adapting the previous computations (the covariance matrix is not positive definite any more), we may check that the conditional distribution of X with respect to Y is Gaussian with expectation

$$KH^T(HKH^T)^{-1}$$

and variance

$$K - KH^t(HKH^T)^{-1}HK .$$

The linearity of conditional expectation is a property of Gaussian vectors and linear conditioning. If you condition with respect to the norm $\|X\|_2$, the conditional distribution is not Gaussian anymore.

12.5 About Gamma distributions

Investigating the norm of Gaussian vectors will prompt us to introduce χ^2 distributions, a sub-family of Gamma distributions.

Gamma distributions

A Gamma distribution with parameters $(p, \lambda)\}$ ($\lambda \in \mathbb{R}_+$ and $p \in \mathbb{R}_+$), is a distribution on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ with density

$$g_{p,\lambda}(x) \equiv \frac{\lambda^p}{\Gamma(p)} \mathbf{1}_{x\geq 0} x^{p-1} e^{-\lambda x}$$

where $\Gamma(p) \equiv \int_0^\infty t^{p-1}e^{-t}\mathrm{d}t$.

Parameter p is called the ***shape*** parameter, λ is called the ***rate*** or ***intensity*** parameter, $1/\lambda$ is called the ***scale*** parameter.

If $X \sim \text{Gamma}(p, 1)$ then $\sigma X \sim \text{Gamma}(p, 1/\sigma)$ for $\sigma > 0$.

Euler's $\Gamma()$ function interpolates the factorial. For every positive real p, $\Gamma(p+1) = p\Gamma(p)$. If p is integer, $\Gamma(p+1) = p!$

Check that $\Gamma(1/2) = \sqrt{\pi}$.

If $X \sim \text{Gamma}(p, \lambda)$ $\mathbb{E}X = \frac{p}{\lambda}$ and $\text{var}(X) = \frac{p}{\lambda^2}$.

The next proposition is a cornerstone of Gamma-calculus. The sum of two independent Gamma-distributed random variables is Gamma distributed if they have the same intensity (or scale) parameter.

If X and Y are independent Gamma-distributed random variables with the same intensity parameter λ $X \sim \text{Gamma}(p, \lambda), Y \sim \text{Gamma}(q, \lambda)$ then $X + Y \sim \text{Gamma}(p + q, \lambda)$.

Proof. The density of the distribution of $X + Y$ is the convolution of the densities $g_{p,\lambda}$ et $g_{q,\lambda}$.

$$\begin{aligned}
g_{p,\lambda} * g_{q,\lambda}(x) &= \int_{\mathbb{R}} g_{p,\lambda}(z)g_{_{q,\lambda}}(x-z)\mathrm{d}z \\
&= \int_0^x g_{p,\lambda}(z)g_{_{q,\lambda}}(x-z)\mathrm{d}z \\
&= \int_0^x \frac{\lambda^p}{\Gamma(p)}z^{p-1}\mathrm{e}^{-\lambda z}\frac{\lambda^q}{\Gamma(q)}(x-z)^{q-1}\mathrm{e}^{-\lambda(x-z)}\mathrm{d}z \\
&= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)}\mathrm{e}^{-\lambda x}\int_0^x z^{p-1}(x-z)^{q-1}\mathrm{d}z \\
&\quad \text{changement de variable } z = xu \\
&= \frac{\lambda^{p+q}}{\Gamma(p)\Gamma(q)}\mathrm{e}^{-\lambda x}x^{p+q-1}\int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u \\
&= g_{p+q,\lambda}(x)\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u\,.
\end{aligned}$$

We may pocket the next observation:

$$B(p,q) := \int_0^1 u^{p-1}(1-u)^{q-1}\mathrm{d}u$$

satisfies $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$. □

Gamma distributions with parameters $(k/2, 1/2)$ for $k \in \mathbb{N}$ deserve to be named: they are χ^2 distributions with k degrees of freedom.

Proposition 12.6 (Chi-square distributions). *The χ^2 distribution with k degrees of freedom (denoted by χ^2_k) has density over* $[0, \infty)$*,*

$$\frac{x^{\frac{1}{2}(k-2)}e^{-\frac{x}{2}}}{2^{k/2}\Gamma(k/2)}.$$

The sum of k independent squared standard Gaussian random variables is distributed according to the chi-square distributions with k degrees of freedom χ^2_k.

Proof. According to proposition **?@prp-densite-gamma**), it suffices to establish the proposition $k = 1$.

Let $X \sim \mathcal{N}(0, 1)$, for $t \geq 0$,

$$\begin{aligned} \mathbb{P}\left\{X^2 \leq t\right\} = & \qquad \Phi(\sqrt{t}) - \Phi(-\sqrt{t}) \\ = & \qquad 2\Phi(\sqrt{t}) - 1\,. \end{aligned}$$

Now, differentiating with respect to t, applying the chain rule provides us with a formula for the density:

$$2\frac{1}{2\sqrt{t}}\phi(\sqrt{t}) = \frac{1}{\sqrt{2\pi t}}\mathrm{e}^{-\frac{t}{2}} = \left(\frac{1}{2}\right)^{1/2} \frac{t^{-1/2}}{\Gamma(1/2)}\mathrm{e}^{-\frac{t}{2}}\,.$$

□

12.6 Norms of centred Gaussian vectors

The distribution of the squared Euclidean norm of a centered Gaussian vector only depends on the spectrum of its covariance matrix.

If $X := (X_1, X_2, \ldots, X_n)^T \sim \mathcal{N}(0, A)$ with $A = LL^T$ (L lower triangular), if $M \in \text{SDP}(n)$, then $X^T M X$ is distributed like $\sum_{i=1}^n \lambda_i Z_i$ where $(\lambda_i)_{i\in\{1,\ldots,n\}}$ denote the eigenvalues of $L^T \times M \times L$ and where Z_i are independent χ^2_1-distributed random variables.

This is a corollary of an important property of standard Gaussian vectors: rotational invariance. The standard Gaussian distribution is invariant under orthogonal transform (a matrix O is orthogonal iff $OO^T = \text{Id}$).

Proof. Matrix A may be factorized as $A = LL^t$ (Cholesky), and X is distributed like LY where Y is standard Gaussian. The quadratic form $X^T M X$ is thus distributed like $Y^T L^T M L Y$. There exist an orthogonal transform O such that $L^T M L = O^t \text{ diag}(\lambda_i) O$. Random vector OY is distributed like $\mathcal{N}(0, I_n)$. □

12.7 Norm of non-centred Gaussian vectors

The distribution of the squared norm of a Gaussian vector with covariance matrix σ^2 Id depends on the norm of the expectation but does not depend on its direction. In addition, this distribution stochastically can be compared with the distribution of the squared norm of a centred Gaussian vector with the same covariance.

rdering

In a probability space endowed with distribution $\mathbb{P}$, a real random variable X is stochastically smaller than random variable Y, if

$$\mathbb{P}\{X \leq Y\} = 1\,.$$

The distribution of Y is said to stochastically dominate the distribution of X.

If X is stochastichally less than Y and if F and G denote the cumulative distribution functions of X and Y, then for all $x \in \mathbb{R}$, $F(x) \geq G(x)$. Quantile functions $F^{\leftarrow}, G^{\leftarrow}$ satisfy $F^{\leftarrow}(p) \leq G^{\leftarrow}(p)$ for $p \in (0, 1)$.

Conversely.

If F and G are two cumulative distribution functions that satisfy $\forall x \in \mathbb{R}\; F(x) \geq G(x)$ then there exists a probability space equipped with a probability distribution $\mathbb{P}$ and two random variables X and Y with cumulative distribution functions F, G that satisfy:

$$\mathbb{P}\{X \leq Y\} = 1\,.$$

The proof proceeds by a ***quantile coupling*** argument.

Proof. It is enough to endow $([0,1], \mathcal{B}([0,1])$ with the uniform distribution. Let $X(\omega) = F^{\leftarrow}(\omega), Y(\omega) = G^{\leftarrow}(\omega)$. Then the distribution of X (resp. Y) has cumulative distribution function F (resp. G) and the following holds:

$$\mathbb{P}\{X \leq Y\} = \mathbb{P}\{F^{\leftarrow}(U) \leq G^{\leftarrow}(U)\} = 1\,.$$

□

If $X \sim \mathcal{N}\left(0, \sigma^2 \operatorname{Id}\right)$ and $Y \sim \mathcal{N}\left(\theta, \sigma^2 \operatorname{Id}\right)$ with $\theta \in \mathbb{R}^d$ then

$$\|Y\|^2 \sim \left((Z_1 + \|\theta\|_2)^2 + \sum_{i=1}^{d} Z_i^2\right)$$

where Z_i are i.i.d. according to $\mathcal{N}(0, \sigma^2)$.

For every $x \geq 0$,

$$\mathbb{P}\left\{\|Y\| \leq x\right\} \leq \mathbb{P}\left\{\|X\| \leq x\right\}\,.$$

The distribution of $\|Y\|^2/\sigma^2$ (non-centred χ^2 with parameter $\|\theta\|_2/\sigma$) ***stochastichally dominates*** the distribution of $\|X\|^2/\sigma^2$ (centred χ^2 with the same number of degrees of freedom).

Proof. The Gaussian vector Y is distributed like $\theta + X$. There exists an orthogonal transform O such that

$$O\theta = \begin{pmatrix} \|\theta\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\,.$$

Vectors OY and OX respectively have the same norms as X and Y.

The squared norm of Y is distributed as the squared norm of OY, that is like $(Z_1 + \|\theta\|_2)^2 + \sum_{i=2}^d Z_i^2$. This proves the first part of the theorem.

To establish the second part of the theorem, it suffices to check that for every $x \geq 0$,

$$\mathbb{P}\left\{(Z_1 + \|\theta\|_2)^2 \leq x\right\} \leq \mathbb{P}\left\{X_1^2 \leq x\right\}\,,$$

that is

$$\mathbb{P}\left\{|Z_1 + \|\theta\|_2| \leq \sqrt{x}\right\} \leq \mathbb{P}\left\{|X_1| \leq \sqrt{x}\right\}\,,$$

or

$$\Phi(\sqrt{x} - \|\theta\|_2) - \Phi(-\sqrt{x} - \|\theta\|_2) \leq \Phi(\sqrt{x}) - \Phi(-\sqrt{x})\,.$$

For $y > 0$, the function mapping $[0, \infty)$ to $\mathbb{R}$, defined by $a \mapsto \Phi(y-a) - \Phi(-y-a)$ is non-increasing with respect to a: it derivative with respect to a equals $-\phi(y-a) + \phi(-y-a) = \phi(y+a) - \phi(y-a) \leq 0$. The conclusion follows □

The last step of the proof reads as

$$\mathbb{P}\left\{X \in \theta + C\right\} \leq \mathbb{P}\left\{X \in C\right\}$$

where $X \sim \mathcal{N}(0, \operatorname{Id}_1)$, $\theta \in \mathbb{R}$ and $C = [-\sqrt{x}, \sqrt{x}]$. This inequality holds in dimension $d \geq 1$ if C is compact, convex, symmetric. This (subtle) result is called Anderson's Lemma.

12.8 Cochran Theorem and consequences

Cochran

Let $X \sim \mathcal{N}(0, \mathrm{I}_n)$ and $\mathbb{R}^n = \oplus_{j=1}^k E_j$ where E_j are pairwise orthogonal linear subspaces of $\mathbb{R}^n$. Denote by π_{E_j} the orthogonal projection on E_j.

The collection of Gaussian vectors $\left(\pi_{E_j} X\right)_{j \leq k}$ is independent and for each~j

$$\|\pi_{E_j} X\|_2^2 \sim \chi^2_{\dim(E_j)} \, .$$

Proof. The covariance matrix of $\pi_{E_j} X$ is $\pi_{E_j} \pi_{E_j}^t = \pi_{E_j}$. The eigenvalues of π_{E_j} are 1 with multiplicity $\dim(E_j)$ and 0. The statement about the distribution of $\|\pi_{E_j} X\|_2^2$ is a corollary of **?@prp-normgausssstand** and **?@prp-normespectre**.

To prove stochastic independence, let us consider $\mathcal{I}, \mathcal{J} \subset \{1, \ldots, k\}$ with $\mathcal{I} \cap \mathcal{J} = \emptyset$. It is enough to check that for all $(\alpha)_{j \in \mathcal{I}}, (\beta_j)_{j \in \mathcal{J}}$, the characteristic functions of

$$\left(\sum_{j \in \mathcal{I}} \langle \alpha_j, \pi_{E_j} X \rangle, \sum_{j \in \mathcal{J}} \langle \beta_j, \pi_{E_j} X \rangle \right)$$

can be factorized. It suffices to check that the two Gaussians are orthogonal.

$$\mathbb{E}\left[\left(\sum_{j \in \mathcal{I}} \langle \alpha_j, \pi_{E_j} X \rangle \right) \times \left(\sum_{j' \in \mathcal{J}} \langle \beta_{j'}, \pi_{E_{j'}} X \rangle \right) \right] \quad = \quad \sum_{j \in \mathcal{I}, j' \in \mathcal{J}} \alpha_j^t \pi_{E_j} \pi_{E_{j'}} \beta_{j'} = 0 \, .$$

□

The next result is a cornerstone of statistical inference in Gaussian models. It is a corollary of Cochran's Theorem.

Student

Of $(X_1, \ldots, X_n)$ are i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$, if $\overline{X}_n := \sum_{i=1}^n X_i / n$ et $V := \sum_{i=1}^n (X_i - \overline{X}_n)^2$, then

1. $\overline{X}_n$ is distributed according to $\mathcal{N}(\mu, \sigma^2/n)$,
2. V is independent from $\overline{X}_n$
3. V/σ^2 is distributed according to χ^2_{n-1}.

Proof. Without loss of generality, we may assume that $\mu = 0$ et $\sigma = 1$.

As

$$\begin{pmatrix} \overline{X}_n \\ \vdots \\ \overline{X}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \times \begin{pmatrix} 1 & \ldots & 1 \end{pmatrix} X$$

the vector $(\overline{X}_n, \ldots, \overline{X}_n)^t$ is the orthogonal projection of the standard Gaussian vector X on the line generated by $(1, \ldots, 1)^t$.

Vector $(X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)^t$ is the orthogonal projection fo Gaussian vector X on the hyperplane which is orthogonal to $(1, \ldots, 1)^t$.

According to Cochran's Theorem (Section 12.8), random vectors $(\overline{X}_n, \ldots, \overline{X}_n)^t$, and $(X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)^t$ are independent.

The distribution of $\overline{X}_n$ is trivially Gaussian.

The distribution of V is characterized using Cochran's Theorem. □

Distribution

If $X \sim \mathcal{N}(0,1), Y \sim \chi^2_p$ and if X and Y are independent, then $Z = X/\sqrt{Y/p}$ is distributed according to a (centred) Student distribution with p degrees of freedom.

12.9 Gaussian concentration

The very definition of Gaussian vectors characterizes he distribution of any affine function of a standard Gaussian vector. If the linear part of the affine function is defined by a vector λ, we know that the variance will be $\|\lambda\|_2^2$. What happens if we are interested in fairly regular functions of a standard Gaussian vector? for example if we consider L-lipschitzian functions? These are generalizations of affine functions. We cannot therefore expect a general increase in the variance of the L-Lipschitzian functions of a standard Gaussian vector better than L^2 (in the linear case the Lipschitz constant is the Euclidean norm of λ). It is remarkable that the bound provided for linear functions extends to Lipschitzian functions. It is even more remarkable that this bound does not involve the dimension of the ambient space.

Let $X \sim \mathcal{N}(0, \mathrm{Id}_d)$.

1. if f is differentiable on $\mathbb{R}^d$,
$$\operatorname{var}(f(X)) \leq \mathbb{E}\|\nabla f\|^2 \qquad \text{(Poincaré's inequality)}$$

2. if f is L-Lipschitz on $\mathbb{R}^d$,
$$\operatorname{var}(f(X)) \leq L^2$$
and for $\lambda > 0$
$$\log \mathbb{E}\mathrm{e}^{\lambda(f(X)-\mathbb{E}f)} \leq \frac{\lambda^2 L^2}{2}\,.$$
For every $t \geq 0$,
$$\mathbb{P}\left\{f(X) - \mathbb{E}f(X) \geq t\right\} \leq \mathrm{e}^{-\frac{t^2}{2L^2}}\,.$$

The proof relies on the next identity.

Covariance identity

Let X, Y be two independent $\mathbb{R}^d$-valued standard Gaussian vectors, let f, g be two differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$.

$$\operatorname{cov}(f(X), g(X)) = \int_0^1 \mathbb{E}\left\langle \nabla f(X), \nabla g\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{d}\alpha$$

We start by checking this proposition on functions $x \mapsto \mathrm{e}^{\imath\langle\lambda,x\rangle}, x \in \mathbb{R}^d$.

Proof. Let us first check Poincaré's inequality.

We choose $f = g$. Starting from the covariance identity, thanks to Cauchy-Schwarz's inequality:

$$\begin{aligned}
\operatorname{var}(f(X)) &= \operatorname{cov}(f(X), f(X)) \\
&= \int_0^1 \mathbb{E}\left\langle \nabla f(X), \nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{d}\alpha \\
&\leq \int_0^1 \left(\mathbb{E}\|\nabla f(X)\|^2\right)^{1/2} \times \left(\mathbb{E}\|\nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\|^2\right)^{1/2} \mathrm{d}\alpha\,.
\end{aligned}$$

The desired results follows by noticing that X and $\alpha X + \sqrt{1-\alpha^2}Y$ are both $\mathcal{N}(0, \mathrm{Id})$-distributed.

To obtain the exponential inequality, choose f differentiable and 1-Lipschitz, and $g = \exp(\lambda f)$ pour $\lambda \geq 0$. Without loss of generality, assume $\mathbb{E}f(X) = 0$. The covariance identity and the chain rule imply

$$\begin{array}{rcl}\operatorname{cov}\left(f(X), \mathrm{e}^{\lambda f(X)}\right) & = & \lambda \int_0^1 \mathbb{E}\left[\left\langle \nabla f(X), \nabla f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)\right\rangle \mathrm{e}^{\lambda f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)}\right] \mathrm{d}\alpha \\ & \leq & \lambda L^2 \int_0^1 \mathbb{E}\left[\mathrm{e}^{\lambda f\left(\alpha X + \sqrt{1-\alpha^2}Y\right)}\right] \mathrm{d}\alpha \\ & = & \lambda L^2 \mathbb{E}\left[\mathrm{e}^{\lambda f(X)}\right]\end{array}$$

Define $F(\lambda) := \mathbb{E}\left[\mathrm{e}^{\lambda f(X)}\right]$. Note that we have just established a differential inequality for F, checking $\operatorname{cov}(f, \mathrm{e}^{\lambda f}) = F'(\lambda)$ since f is centred:

$$F'(\lambda) \leq \lambda L^2 F(\lambda)\,.$$

Solving this differential inequality under $F(0) = 1$, for $\lambda \geq 0$

$$F(\lambda) \leq \mathrm{e}^{\frac{\lambda^2 L^2}{2}}\,.$$

The same approach works for $\lambda < 0$. It is enough to invoke Markov's exponential inequality and to optimize with respect to $\lambda = t/L^2$. □

Concentration inequalities describe the behavior of the norm of high-dimensional Gaussian vectors.

If X is a standard d-dimensional Gaussian vector, then

$$\operatorname{var}(\|X\|_2) \leq 1$$

and

$$\sqrt{d-1} \leq \mathbb{E}\|X\|_2 \leq \sqrt{d}\,.$$

Proof. The Euclidean norm is 1-Lipschitz (triangle inequality). The first inequality follows fron Poincaré's inequality.

The upper bound on expectation follows from Jensen's inequality.

The lower bound on expectation follows from $(\mathbb{E}\|X\|_2)^2 = \mathbb{E}\|X\|_2^2 - \operatorname{var}(\|X\|_2) = d - \operatorname{var}(\|X\|_2)$ and from the variance upper bound. □

Let $X \sim \mathcal{N}(0, K)$ where K is in $\mathsf{DP}(d)$ and $Z = \max_{i \leq d} X_i$.

Show

$$\operatorname{Var}(Z) \leq \max_{i \leq d} K_{i,i} := \max_{i \leq d} \operatorname{Var}(X_i).$$

Let $X, Y \sim \mathcal{N}(0, \mathrm{Id}_n)$ with X, Y indépendent.

Show

$$\sqrt{2n-1} \leq \mathbb{E}[\|X-Y\|] \leq \sqrt{2n}$$

and

$$\mathbb{P}\left\{\|X-Y\| - \mathbb{E}[\|X-Y\|] \geq t\right\} \leq \mathrm{e}^{-t^2}\,.$$

12.10 Bibliographic remarks

Gaussian literature is very abundant, see for example (**?**). Much of this literature is relevant to statistics.

The lemmas **?@lem-stein** and **?@lem-steinbis** that characterize the Gaussian standard are the starting point of Stein's (Charles) method to demonstrate the central limit theorem (and many other results). This relatively recent development is described in (N. Ross, 2011).

Matrix analysis and algorithmics play an important role in Gaussian analysis and statistics. The books (Horn & Johnson, 1990), and if we wish to go further (Bhatia, 1997), provide an introduction to the concepts and techniques of matrix factorization and elements of perturbation theory.

There is a multi-dimensional version of the laws of χ^2 that appear when determining the law of variance empirical. These are the laws of Wishart. They were the subject of intensive studies in random matrix theory, see for example (Anderson, Guionnet, & Zeitouni, 2010)

Gaussian concentration plays an important role in non-parametric statistics and is a source of inspiration in statistical learning. M. Ledoux's book (Ledoux, 2001) provides an elegant perspective on this issue.

Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). ***An introduction to random matrices*** (Vol. 118). Cambridge: Cambridge University Press.

Bhatia, R. (1997). ***Matrix analysis***. Springer-Verlag.

Billingsley, P. (2012). ***Probability and measure***. John Wiley & Sons, Inc., Hoboken, NJ.

Boucheron, S., Lugosi, G., & Massart, P. (2013). ***Concentration inequalities***. Oxford University Press.

Chen, L. H. Y., Goldstein, L., & Shao, Q.-M. (2011). ***Normal approximation by Stein's method*** (p. xii+405). Springer, Heidelberg. https://doi.org/10.1007/978-3-642-15007-4

Cover, T., & Thomas, J. (1991). ***Elements of information theory***. John Wiley & sons.

Dudley, R. M. (2002). ***Real analysis and probability*** (Vol. 74, p. x+555). Cambridge: Cambridge University Press.

Durrett, R. (2010). ***Probability: Theory and examples***. Cambridge University Press.

Hiriart-Urruty, J.-B., & Lemaréchal, C. (1993). ***Convex analysis and minimization algorithms. I*** (Vol. 305, p. xviii+417). Springer-Verlag, Berlin.

Horn, R. A., & Johnson, C. R. (1990). ***Matrix analysis***. Cambridge University Press.

Ledoux, M. (2001). ***The concentration of measure phenomenon***. AMS.

Massart, P. (2007). ***Concentration inequalities and model selection. Ecole d'eté de probabilité de saint-flour* XXXIV** (Vol. 1896). Springer-Verlag.

McDiarmid, C. (1998). Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, & B. Reed (Eds.), ***Probabilistic methods for algorithmic discrete mathematics*** (pp. 195–248). Springer, New York.

Pollard, D. (2002). ***A user's guide to measure theoretic probability*** (Vol. 8, p. xiv+351). Cambridge University Press, Cambridge.

Ross, N. (2011). Fundamentals of Stein's method. ***ArXiv e-Prints***. Retrieved from https://arxiv.org/abs/1109.1880

Ross, Nathan. (2011). Fundamentals of Stein's method. ***Probab. Surv.***, *8*, 210–293. https://doi.org/10.1214/11-PS182

Widder, D. V. (2015). ***Laplace transform (PMS-6)***. Princeton university press.

Wilf, H. S. (2005). ***Generatingfunctionology***. AK Peters/CRC Press.

Williams, D. (1991). ***Probability with martingales*** (p. xvi+251). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511813658