Probability theory notes

Kotatsu

TOALDO COL **CAZZO** CHE DIVENTI ORDINARIO

Preface

This document stems from the fact that I just seem unable to pass the Probability Theory exam for the life of me. I regret with every ounce of my being the fact that I enrolled to the Stochastics and Data Science master degree a year ago. Since dear Professor Toaldo never really thrilled me with his insightful lectures about this delightful topic, I resorted to watch the old lectures by Professor Polito, who at least seems to know the subject and to be determined to explain it.
Unlike many among my esteemed colleagues I have NOT a background in mathematics so there will be a lot of repetitions and possibly mistakes. Do what you want with this information. YES I KNOW that there are the whiteboard registrations of his lectures but if I DECIDED TO DO THIS it was because I couldn't comprehend shit with only those notes.
I'll also try to compile the notes made by Professor Sacerdote, in the vain attempt to overcome the drowsiness that is congenitally entwined with every event that contemplates her uttering any words.
I take much pride in my custom environment and in my packages. If you don't like them I will be very sad.
It is strongly recommended to play Metal Gear Solid, Metal Gear Solid 2 and Metal Gear Solid 3 before reading these notes to fully understand the subject treated.

Kotatsu

Contents

1 Basics of probability

We start with the probability triplet: $(\Omega, \mathcal{H}, \mathbb{P})$ Here Ω is the set of sample space, $\mathcal{H}$ is the σ-algebra built upon Ω and $\mathbb{P}$ is the probability measure. Since $\mathbb{P}$ is a measure, it will take values in $\mathbb{R}$.
We are interested in probability measure, which means:

- $\mathbb{P}$ is a **finite measure** and $\mathbb{P}(\Omega) = 1$;
- $\omega \in \Omega$ will be called **outcomes**.

So consider the example of the roll of the die. If we roll it,

$$\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$$

And if we considere the elements $A \in \mathcal{H}$ (which will be subsets of Ω) will be called **events**.

We want to quantify the possibility that the event A occurs: we want to measure, through $\mathbb{P}$, the set A: from a measure theory point of view, it's only sets in the σ-algebra.
The probability measure has the following properties:

- $\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\varnothing) = 0$

- **monotonicity of** $\mathbb{P}$: take 2 events $H, K \in \mathcal{H}$ such that $H \subset K$. Then $\mathbb{P}(H) \leqslant \mathbb{P}(K)$[1].
- **finite additivity**: take $H, K \in \mathcal{H}$ such that $H \cap K = \varnothing$. The $\mathbb{P}(H \cup K) = \mathbb{P}(H) + \mathbb{P}(K)$;
- **countable additivity**: this requires that we consider collection of events. We denote them in this way:
$$(H_n)_{n\in\mathbb{N}} \subset \mathcal{H}$$
with $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ and $\mathbb{N}^\star = \{1, 2, 3, 4, \ldots\}$ such that they are disjoint pairwise (except identical pairs). Then
$$\mathbb{P}\left(\bigcup_n H_n\right) = \sum_n \mathbb{P}\left(H_n\right)$$
- **Boole inequality (sub-additivity)**: if we have a collection $(H_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ (not necessarily disjoint) then
$$\mathbb{P}\left(\bigcup_n H_n\right) \leqslant \sum_n \mathbb{P}\left(H_n\right)$$
- **sequential continuity**: consider the sequence $(H_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ such that $H_n \nearrow H \in \mathcal{H}$ (H_n is an increasing sequence of numbers that has H as limit) then $\mathbb{P}(H_n) \nearrow \mathbb{P}(H)$. Moreover, if $(F_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ such that $F_n \searrow F \in \mathcal{H}$ then $\mathbb{P}(F_n) \searrow \mathbb{P}(F)$. The second property is actually true because $\mathbb{P}$ is finite (it is not true for infinite measures).

In measure theory we encounter the concept of **negligible sets**: these are sets of measure zero or non measurable sets included in measure zero sets. In probability theory, sets are **events**: so we have negligible events (events with probability 0 or non measurable events included in events with probability 0). Analogously, in measure theory a property which holds **almost everywhere** is allowed not to hold on negligible sets. In probability theory a property which holds **almost surely** is allowed not to hold on negligible events. We also have, in measure theory, *measurable functions* that in probability theory are **random variables**. Let's have a look back into what the absolute fuck a measurable function is. Also what is an integral? This course deals with distributions, measures and other hellish machinery that servers the sole purpose to confuse you.

Revise with Kotatsu!

Definition 1.1

Let $(E, \mathcal{E}$ and $(F, \mathcal{F})$ be measurable spaces. A mapping $f : E \mapsto F$ is said to be **measurable** relative to $\mathcal{E}$ and $\mathcal{F}$ if
$$f^{-1}(B) \in \mathcal{E} \qquad \forall\, B \in \mathcal{F}.$$

There is an useful property to measurable functions. Take a function $f : E \mapsto F$. In order for f to be measurable relative to $\mathcal{E}$ and $\mathcal{F}$ it is necessary and sufficient that
$$f^{-1}(B) \in \mathcal{E} \qquad \forall\, B \in \mathcal{F}_0$$
where $\mathcal{F}_0$ is a collection that generates $\mathcal{F}$, i.e. $\mathcal{F} = \sigma(\mathcal{F}_0)$.

1.1 Random variables

Consider a measurable space $(E, \mathcal{E})$.

Definition 1.2

A mapping $X : \Omega \to E$ is called **random variable taking values in** E if X is measurable relative to $\mathcal{H}$ and $\mathcal{E}$.

[1] note that the notation is loose since we have proper subset on one side and leq on the other side. But this is not much of a problem, since i will kill myself very soon.

What does it mean[2]? The inverse image of the set A through X ($X^{-1}A$) with $A \in \mathcal{E}$ is actually the set of the ωs such that $X(\omega)$ arrives to A. So

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

so that $X^{-1}A$ is an event for all A in $\mathcal{E}$.

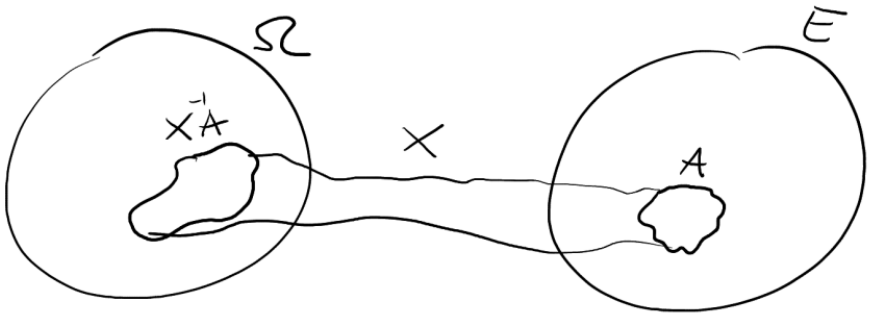


Figure 1: this is an early reminder of the fact that I will take my own life very soon.

So if $X^{-1}A$ is measurable by $\mathbb{P}$ then it is in $\mathcal{H}$: otherwise it is not in $\mathcal{H}$. So

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

The message is that I am interested/able to evaluate $\mathbb{P}$ over the set only if what I am evaluating is indeed an event (which means: it belongs to $\mathcal{H}$[3]). If something is not in $\mathcal{H}$ get it off my fucking face man and kill yourself NOW[4]. This is the only restriction for a random variable. E can be whatever we need it to be: a graph, a tree, your mom being absolutely [redacted] by me. But most of the times, we have $E = \mathbb{R}$ or $E = \mathbb{R}^d$ with respectively $\mathcal{E} = \mathcal{B}$[5]$(\mathbb{R}) = \mathcal{B}_{\mathbb{R}}$ and $\mathcal{B}_{\mathbb{R}^d}$.

Remark

The simplest random variables are indicator functions of events. Example: take $H \in \mathcal{H}$. Define the function

$$\mathbb{1}_H : \Omega \to \mathbb{R}$$
$$\mathbb{1}_H(\omega) = \begin{cases} 0 & \omega \notin H \\ 1 & \omega \in H \end{cases}$$

Remark

A random variable is said to be **simple** if it takes only finitely many values in $\mathbb{R}^d$.

Remark

A random variable is said to be **discrete** if it takes only countably many values.

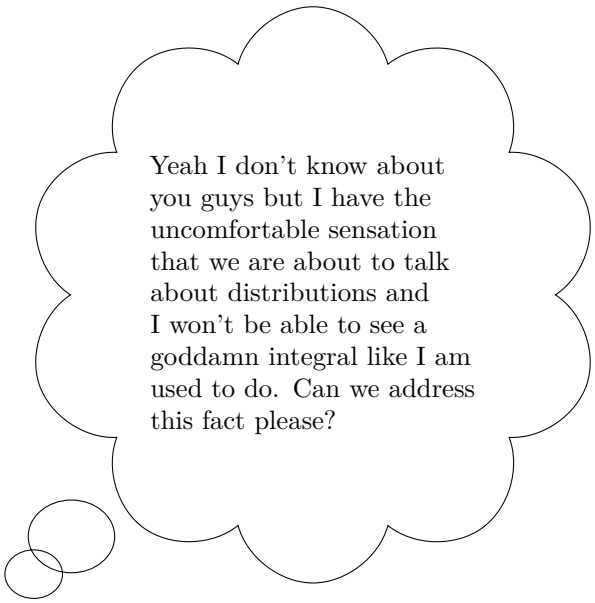
We are now ready to define the concept of *distribution of a random variable*. But first...

[2] who asked

[3] il lettore più arguto avrà notato che, a questo punto, il dio è ormai irrimediabilmente cane.

[4] 

[5] Borel σ-algebra. You don't know what a Borel σ-algebra is? https://en.wikipedia.org/wiki/Borel_set



Sure. Let's have a look back to Lebesgue integration.

Revise with Kotatsu!

Consider a measure space $(E, \mathcal{E}, \mu)$. $\mathcal{E}$ can be seen as the collection of all $\mathcal{E}$-measurable functions $f : E \mapsto \overline{\mathbb{R}}$ on E that can be denoted with an abuse of notation[a] by $f \in \mathcal{E}$ and by $d \in \mathcal{E}_+$ if the functions are positive. Our aim is to define integrals of measurable functions with respect to the measure μ so that:

$$\mu f = \mu(f) = \int_E f(x)\mu(\mathrm{d}x) = \int_E f \,\mathrm{d}\mu$$

which is written as the product of μ and f. It is interesting to note, in the last part of the equation, that the integral reads something like: "integrate f over E with respect to the measure μ". What is this measure?? This is the question. Turns out that the good old Riemann integral is just a particular case of the Lebesgue integral when a certain measure is chosen.
We consider them as the generalization of vectors and hence the scalar product becomes a sum, which transforms into an integral. We will define the Lebesgue integral in three steps:

1. Simple and positive functions:

   Definition 1.3

   The function f is called a **simple and positive function** if it can be written as

   $$\sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$$

   where $A_i \in \mathcal{E}$ and $a_i \geqslant 0 \in \mathbb{R}$ for $i = 1, 2, \ldots, n$.

Definition 1.4

For simple and positive functions, we define the Lebesgue integral as

$$\mu f := \sum_{i=1}^{n} a_i \mu(A_i)$$

2. Positive and measurable functions:

Theorem 1.1

Let $f \in \mathcal{E}_+$. Then there exists a sequence of simple and positive functions f_n such that $f_n \nearrow f$.

Thanks to this theorem, we can well pose the following definition"

Definition 1.5

Let $f \in \mathcal{E}_+$. We define

$$\mu f := \lim_n \mu f_n$$

where f_n is a sequence of simple and positive functions such that $f_n \nearrow f$.

3. Recall a general fact for real-valued functions.

Remark

Let f be a real-valued function. Then we can write

$$f = f^+ - f^-$$

With $f^+ := f \vee 0 = max\{f, 0\}$, called **positive part** and $f^- := -(f \wedge 0) = -min\{f, 0\}$, called **negative part**. Both of them are real and positive functions and f is measurable if and only if f^+ and f^- are real and positive functions.

We are now ready to define the Lebesgue integral for measurable functions in $\mathbb{R}$. The trick is to separate the positive and the negative part of the function, to treat them as the limit of sequence of simple functions and then lose ourselves in the bliss of measure theory.

Definition 1.6

Let $f \in \mathcal{E}$. We define

$$\mu f := \mu(f^+) - \mu(f^-)$$

Provided that at least one of the integrals is finite in order to be defined and not incur into indefinite forms like $+\infty$ or $-\infty$.

This definition can be easily converted if f is a complex function: we only have to remember that we can decompose any complex number in its real and imaginary part. Both of them will be measurable real functions.

$$f = \mathfrak{Re}^+ f - \mathfrak{Re}^- f + i(\mathfrak{Im}^+ f - \mathfrak{Im}^- f).$$

From now on we will use this notation for the Lebesgue integral on $(E, \mathcal{E})$:

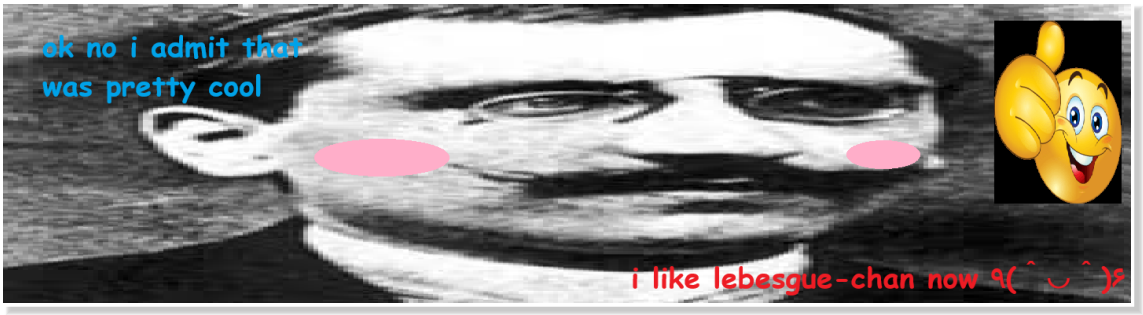
$$\mu f = \int_E f(x)\mu(\mathrm{d}x) \qquad \text{with } f \in \mathcal{E}$$

and if we choose $f = \mathbb{1}_B$ with $B \in \mathcal{E}$. then

$$\mu f = \mu \mathbb{1}_B = \int_E \mathbb{1}_B(x)\mu(\mathrm{d}x) = \mathbb{1}_B \mu(\mathrm{d}x) = \mu(B).$$

So this last equivalence helps us to understand one thing. Integrals are a device that needs a measure and a function to work. In the notation above, $\mathrm{d}x$ has the meaning of an infinitesimal amount of the variable x that is fed into the function f. Writing $\mu(\mathrm{d}x)$ means measuring an infinitesimal amount of x using the measure μ.
In Riemann integration, $\mathrm{d}x$ represents an infinitesimal segment of the x-axis multiplied by the height of the function at x (which is, of course, $f(x)$) and summed ($\int_a^b$) with all the other infinitesimal segments of the x-axis over the interval $[a, b]$.
Here it's really the same thing with the difference that we multiply the height of the function $f(x)$ by calculating the "weight", or "measure" of a smaller and smaller part of the domain that "causes that function to be of that height", according to our method of measure of choice. We do this over the set E.

[a] I am the only one being abused here.



The main difference between Riemann and Lebesgue integration is, in a certain way, *what* we are slicing. In the Riemann approach be basically do the following:

1. slice the x-axis in smaller and smaller slices;
2. compute $f(x)$;
3. sum all the cute little rectangles you got.

In the Lebesgue approach we basically *start by choosing different slices of the* range*of the function*, that is the co-domain. These little "slabs" of the range of the functions are nothing else but the "stepped" simple function version of our function:

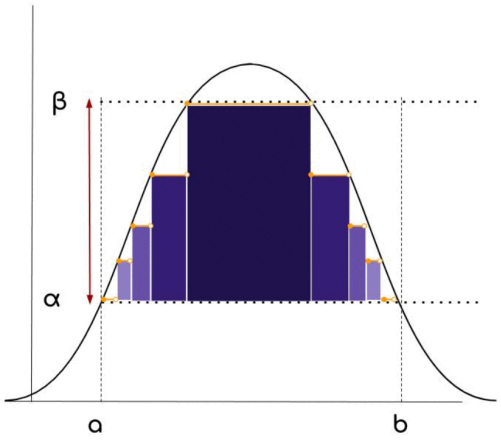


Figure 2: Imagine the steps getting smaller and smaller...

Since we are dealing with simple functions, we are effectively approaching the problem from the y-axis. This means that since we are choosing slices of height *first* our "slabs" may have different length when it comes to the x-axis. Anyway was it really SO DIFFICULT to explain? I don't think so. Fuck you mathematicians.
So... will there ever be a measure and a set for which we will be able to circle back to our definition of Riemann integral? Hmm...

Definition 1.7

Distribution of a random variable. Let X be a random variable taking values in $(E, \mathcal{E})$ and let μ be the image of $\mathbb{P}$ under X, that is,

$$\mu(A) = \mathbb{P}(X^{-1}A) = \mathbb{P}(X \in A) = \mathbb{P} \circ X^{-1}(A)^{a}, \ A \in \mathcal{E}.$$

Then μ is a probability measure on $(E, \mathcal{E})$ and it is called **distribution of X**.

[a] you would know this if you knew fucking measure theory I guess

So we map, by means of X, sets belonging to $\mathcal{E}$ into $\mathcal{H}$ and then evaluates this sets by means of the measure $\mathbb{P}$. This is what we mean when we say that distributions are ultimately built with the probability measure and the random variable.
Distribution is itself a measure. To be exact it is a measure that we employ with a function (that in our case is a random variable) to form a Lebesgue integral just like we have seen in the revise box above. As we said, integrals are a machine that needs a function and a measure; in the case of probability theory these elements are respectively the **random variable** and the **probability distribution**.
Right now we can start to see the light at the end of the tunnel[6] and start to have an intuition for all the ingredients to create this soup called "probability theory". Distributions are NOT cumulative density functions and neither they are probability density functions... They are something that transcends these "specialized" concepts and goes to the heart of how we evaluate (how we weigh; how we **measure**) a probability in a certain scenario.

Distributions are probability measures.

Remark

You should remember (LOL) that when we want to specify a measure on a σ-algebra, it's enough to do it on a π-*system*[a] generating that σ algebra: by means of the monotone class theorem we are then able to extend the measure to the σ-algebra.
This means that to specity μ it is enough to specify it on a π-*system* which generates $\mathcal{E}$. For example, consider $E = \overline{\mathbb{R}}, \mathcal{E} = \mathcal{B}_{\overline{\mathbb{R}}}$. Consider the collection of sets $[-\infty, x]$, $x \in \mathbb{R}$ which is of course a π-system because it is closed under intersection. Moreover, this shit generates the Borel sigma algebra on $\overline{\mathbb{R}}$.
If we want to define a distribution, that is a measure, it is enough to define it on this π-system. Imagine that we apply our distribution measure to one set of this π-system

$$c(x)^{b} = \mu\left([-\infty, x]\right) = \mathbb{P}(X \leqslant x), \qquad x \in \mathbb{R}$$

by the monotone class theorem. So we have now specified the measure on the π-system. The part $\mathbb{P}(X \leqslant x)$ reminds us of the undergraduate times[c]: it is a distribution function! This is what our professor did implicitly to avoid using measure theory[d].

[a] a π-system is a simpler object than a σ-algebra: it is simply a collection of sets closed under intersection
[b] because it is a function of x
[c] I already wanted to kill myself at that time.
[d] I have noticed that my life has not benefited in ANY form since I have been introduced to measure theory.

Revise with Kotatsu!

But what is the *monotone class Theorem*? First, we need the definition of *monotone class*:

[6] This is only the first chapter.

Definition 1.8

A collection of functions $\mathcal{M}$ is called **monotone class** provided that:

1. it includes the constant function 1;
2. taken f and $g \in \mathcal{M}_b$ (with $\mathcal{M}_b$ being the subcollection of bounded functions in $\mathcal{M}$) and $a, b \in \mathbb{R}$, then $af + bg \in \mathcal{M}$;
3. if the sequence $(f_n)_n$ is contained in $\mathcal{M}_+$ (with $\mathcal{M}_+$ being the subcollection consisting of positive functions in $\mathcal{M}$) and $f_n \nearrow F$ then $f \in \mathcal{M}$.

Theorem 1.2

Monotone class Theorem:
Let $\mathcal{M}$ be a monotone class of functions on E. Suppose, for some π-system $\mathcal{C}$ generating $\mathcal{E}$, that $\mathbb{1}_A \in \mathcal{C}$ for every $A \in \mathcal{C}$. Then $\mathcal{M}$ incluedes all positive $\mathcal{E}$-measurable functions and all bounded $\mathcal{E}$-measurable functions.

So, turning back to the previous remark: in that case $\mathcal{E}$ consists of the Borel σ-algebra on the extended real line ($\mathcal{B}_{\overline{\mathbb{R}}}$); our π-system is capable of generating the Borel σ-algebra (because every Borel set can be constructed with the combination $[-\infty, x]$ for all $x \in \mathbb{R}$[7]); we defined the measure μ on the π-system $[-\infty, x]$ for all $x \in \mathbb{R}$; the monotone class theorem states that if a class of sets (in this case, the class of sets where μ is well-defined) contains a π-system (✓) and is closed under monotone limits (i.e. is a monotone class), then it contains the σ-algebra generated by the π-system: this means that the class of sets where the distribution μ is well-defined will include the Borel σ-algebra $\mathcal{B}_{\overline{\mathbb{R}}}$. This is kinda cool, I'll have to admit. Unfortunately, I don't really care about this.

1.2 Functions of random variables

Consider X, a random variable taking values in $(E, \mathcal{E})$ and consider further a measurable space $(F, \mathcal{F})$. Let $f : E \to F$ be a measurable function relative to $\mathcal{E}$ and $\mathcal{F}$[8]. This function should me measurable by means of $\mathbb{P}$, otherwise we couldn't do anything useful with it. Consider the composition

$$Y = f \circ X \qquad \text{such that } Y(\omega) = f \circ X(\omega) = f\big(X(\omega)\big),\ \omega \in \Omega.$$

This composition is a random variable taking values in $(F, \mathcal{F})$ which comes from the fact that measurable functions of measurable functions are still measurable.

Definition 1.9

Consider two random variables X, Y taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively. Consider the pair

$$Z = (X, Y) : \Omega \to E \times F.$$

Why would we want to call it Z? It's because, beside being a random vector, it is in turn a random variable:

$$Z(\omega) = (X(\omega), Y(\omega)).$$

Since $E \times F$ is a product space, we should attach it the product σ-algebra. So Z is a random variable taking values in $E \times F$.

Note that the product space $E \times F$ is endowed with the σ-algebra$\mathcal{E} \otimes \mathcal{F}$, that is the product σ-algebragenerated by the collection of all possible rectangles between E and F. We frequently have to look to special cases like random vectors that must take values in measurable spaces for them to make sense. This measurable space is naturally generated by the product σ-algebra(but it may be generated by other σ-algebras[9]!).

[7] I know, I know: the fuck is a Borel set? A Borel set is every set that can be formed by the countable union or countable intersection or complementation from any open or closed set. You see that every Borel set you can imagine can be constructed by $[-\infty, x]$.

[8] This basically means that this bitch won't do anything evil. The whole point of measure theory, σ algebras and all other shit is to ensure everything behaves.

[9] Repeatedly inflicting painful kicks on my gonads.

Definition 1.10

We call **joint distribution** of X and Y the distribution of Z.

This is interesting, since we know that this variable has the specific structure of a random vector: we identify the distribution of this vector as the joint distribution of its two coordinates[10].

Remark

The product σ-algebra $\mathcal{E} \otimes \mathcal{F}$ is generated by the π-system of measurable rectangles.

On the product space, it is enough to only specify it on this π-system.
Let denote with π the joint distribution of X, y. It is sufficient to specify

$$\pi(A \times B) = \mathbb{P}(X \in A, Y \in B) \qquad \forall\, A \in \mathcal{E}, B \in \mathcal{F}.$$

We exploited the measurability of X and Y

Definition 1.11

Given the joint distribution π, consider sets $A \in \mathcal{E}, B \in \mathcal{F}$. Then we call **marginal distribution of X**

$$\mathbb{P}(X \in A) = \pi(A \times F) \qquad \forall\, A \in \mathcal{E}$$

and we call **marginal distribution of Y**

$$\mathbb{P}(Y \in B) = \pi(E \times B) \qquad \forall\, B \in \mathcal{F}.$$

We call it distribution because it is actually a measure! So we can call it with the notation of measure

$$\mu(A) = \mathbb{P}(X \in A) = \pi(A \times F) \qquad \forall\, A \in \mathcal{E}$$

and

$$\nu(B) = \mathbb{P}(Y \in B) = \pi(E \times B) \qquad \forall\, B \in \mathcal{F}.$$

This actually means that the second coordinate is fixed in being the whole space F. Think about integrating the second coordinate along the real line when doing marginal distributions... this is the same thing here.
Now that we have joint and marginal distributions, what is the next step[11]?

Definition 1.12

Let X, Y be random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$ respectively and let μ and ν be their respective distributions. Then X and Y are said to be **independent** if their joint distribution is the product measure formed by their marginals.

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad \forall\, A \in \mathcal{E}, B \in \mathcal{F}.$$

This also means that

$$\pi = \mu\nu$$

Here the marginals do not interact with each other. This is true for two random variables but we need[12] something more general.

Definition 1.13

Let $(X_1, X_2, \dots, X_n)$ be a finite collection of random variables. The collection is said to be an **independency** if the distribution of $(X_1, X_2, \dots, X_n)$ is the product of $\mu_1, \mu_2, \dots, \mu_n$ where μ_i is the distribution of X_i, for $i = 1, \dots, n$.

Çinlar is stupid I wish him dead to be frank for this independecy shit. Independecy is not even an english word. What the fuck? Anyway, what about infinite collections?

[10] it eludes me how anyone could find this interesting. We have to think about the whole vector as being distributed like its components separately
[11] Abandoning myself in the sweet embrace of Death, methinks.
[12] No.

Definition 1.14

Let $(X_n)_n$ be an infinite collection of random variables. It is said to be an **independency** if every finite sub-collection of it is an independecy.

We now turn to stochastic processes[13]! But first...

1.3 Infinite product spaces

Let T be an arbitrary (countable or uncountable) set. We will think about this set as an "index" set. For each $t \in T$ consider the measurable $(E_t, \mathcal{E}_t)$. So we have a space for each index (plenty of measurable spaces hanging around). Consider a point x_t in E_t for each $t \in T$. The collection[14] $(x_t)_{t\in T}$. If $(E_t, \mathcal{E}_t) = (E, \mathcal{E})$ then $(x_t)_{t\in T}$ is actually a function of T taking values on $(E, \mathcal{E})$.
The set F of all possible functions $x = (x_t)_{t\in T}$ is called the **product space** $((E_t\mathcal{E}_t))_{t\in T}$.
This is the natural generalization of what we do when we construct product spaces, albeit with a different notation. Usually F is denoted by $X_{t\in T}E_t$. But we know we also need a σ-algebra...
A **rectange** in F is a subset of the form

$$\{x \in F : x_t \in A_t \ \forall t \in T\}$$

Where A_t differs from E_t for only a finite number of t. So I want to consider subsets of F (the space of functions) of the form above. I want only the functions x in F such that each coordinate belongs to A_t, a subset of E_t for each $t \in T$. It seems that we have a restriction on all the coordinates... But this may bring to problems when we have an uncountable number of coordinates and therefore an uncountable number of restrictions. But we can says that if $A_t = E_t$ (the whole space) we don't apply any restriction. So in this case X_t belongs to E_t so we can choose whatever X_t we like. So only a finite number of coordinates are restricted while the other infinite ones are free to vary[15].

The σ-algebra generated by the collection of all measurable rectangles is denoted by

$$\bigotimes_{t\in T} \mathcal{E}_t.$$

This is the product σ-algebra in any infinite-dimensional space. So, the (natural) resulting measurable space in the end will be

$$\bigtimes_{t\in T} E_t, \bigotimes_{t\in T} \mathcal{E}_t.$$

This is not in contrast with what we already know for finite product space, since these already have a finite number of restrictions. So this concept of rectangle, which can be a bit different from the one regarding the famous and well-tested geometrical shape[16], is not restricted on all the coordinates (like the shape[17]) but only on a finite number of them.
We also have an alternative notation for this measurable space!

$$\bigotimes_{t\in T} (E_t, \mathcal{E}_t).$$

In the case that $(E_t, \mathcal{E}_t) = (E, \mathcal{E}) \ \forall t \in T$ the product space is denoted by

$$(E, \mathcal{E})^T$$

or

$$(E^T, \mathcal{E}^T)$$

These are not real powers but it's just notation... Anyway these are all different notations to indicate the infinite product space with the product σ-algebra built upon the π-system which is the collection of all possible rectangle defined in the way we saw above[18].

1.4 Stochastic processes

[13] Please no.
[14] We could consider it a function of t but that wouldn't be exactly correct since each t has a different measurable space. We may have the same space but it's not true in general... I am thrilled to say the least.
[15]  → my honest reaction.
[16] Oh thank god someone finally said it. I was starting to get scared.
[17] I swear to god.
[18] NO I WON'T USE LABELS AND NUMBERED EQUATIONS.

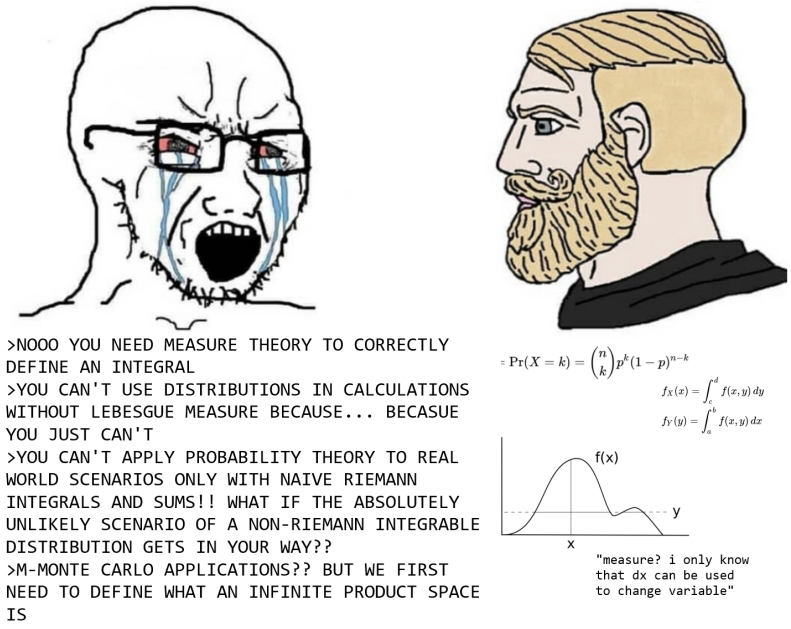


Figure 3: I'm sorry but here you're the soyjack and I'm the chad.

Definition 1.15

Let $(E, \mathcal{E})$ be a measurable space and consider an index set T (as before, an arbitrary set countable or uncountable).
Let also X_t be a random variable taking values in $(E, \mathcal{E})$. Then the collection of those random variables $(X_t)_{t \in T}$ is called a **stochastic process** with state space $(E, \mathcal{E})$ and parameter set T.

Note that there is no mention about time here. Just think about the index set, which indexes the stochastic process. If we interpret T as time then we have the most common interpretation of stochastic processes. But it could also be space (imagine $\mathbb{R}^2$) or your mom being [redacted]. Anyway the most natural interpretation is time.

Now take a $\omega \in \Omega$ and evaluate all these random variables on the same ω. What we get is

$$t \mapsto X_t(\omega)$$

which is a function from T to $(E, \mathcal{E})$. So if we see it as a function of t for each ω we get a function which is an element of E^T. So what is a stochastic process, to sum it up? It's just a random variable taking values in the infinite product space E^T. That's why it is a problematic object: it's because mathematicians deserve to experience the sadness and evil they unleashed upon the world. Ever noticed how similar the words "measurable" and "miserable" are? I didn't think so. Yeah technically more structure helps us modeling real phenomena more accurately but who the FUCK cares.

1.5 Example of random variables

Consider some examples of simple random variables:

Example 1.1

Poisson random variables.
This random variable takes vales in $\mathbb{N}$ (it's a one dimensional random variable). We consider the power set of $\mathbb{N}$[a]. We know that power sets *are* σ-algebras that we can use (but we could encounter some trouble with uncountable elements, for which we would need smaller σ-algebras[b]).

What is the distribution of this random variable?

$$\mu(A) = \mathbb{P}(X \in A) = \sum_{n \in A} \mathbb{P}(X = n) \qquad A \subset \mathbb{N}$$

$$\text{with } \mathbb{P}(X = n) = e^{-c}\frac{c^n}{n!},\ n \in \mathbb{N},\ c > 0.$$

So imagine we have this kind of random variable. We consider a subset of the natural number and we want to evaluate the measure of this subset that we chose. We know that we define the random variable by defining the distribution. For each n we get a number $e^{-c}\frac{c^n}{n!}$. Another interesting implication is that

$$\sum_{n \in A} \mathbb{P}(X = n) = \sum_{n \in \mathbb{N}} \delta_n(A)\mathbb{P}(X = n)$$

where $\delta_n(A)$ is the **Dirac measure** sitting at n. So, n is a parameter and

$$\delta_n(A) = \begin{cases} 1 & n \in A \\ 0 & n \notin A \end{cases}.$$

The Dirac measure is similar to the indicator function (they behave basically in the same way) but the difference is that this one is a *measure* and the latter is a *function*. The Dirac measure has n as a parameter, while the indicator function has the set as a parameter ($\mathbb{1}_A(n)$).

[a] Subset of all subsets of $\mathbb{N}$.
[b] No one really cares, not even Federico Polito.

Example 1.2

Exponential random variable

This random variable is again one-dimensional but this time this random variable is *absolutely continuous*. What does it mean? It actually means that the variable is absolutely continuous with respect to the Lebesgue measure[a]. This is evident when we write down the distribution. Consider a random variable taking values in $\mathbb{R}_+$ and further consider $\mathcal{B}_{\mathbb{R}_+}$. We have

$$\mu(\mathrm{d}x) = \underbrace{\mathrm{d}x}_{Leb(\mathrm{d}x)} ce^{-cx}, \qquad c > 0,\ x \in \mathbb{R}_+.$$

Ok no wait hold your fucking horses, cowboy. Why did we write $\mathrm{d}x$ instead of just x? Also weren't densities, like, a fucking measure of some set in the form of $\mu(A)$? I like the fact that these densities resemble more closely the probability density function I was taught to work with during my sad Economics degree but there are many many things that creep me out. We have an answer for this, but we need to do a bit of backtracking.

[a] tacci tua.

Revise with Kotatsu!

First of all: what does "absolutely continuous" *even means*?

Definition 1.16

let μ and ν be measures on a measurable space $(E, \mathcal{E})$. Then, measure ν is said to be absolutely continuous with respect to measure μ if, for every set $A \in \mathcal{E}$,

$$\mu(A) = 0 \implies \nu(A) = 0.$$

Huh. That was pretty simple. Well, turns out we can exploit this fact to "switch" between different measures inside of integrals...

Theorem 1.3

Radon-Nikodyn Theorem. Suppose that measure μ is σ-finite and measure ν is absolutely continuous with respect to μ. Then there exists a positive $\mathcal{E}$-measurable function p such that

$$\int_E \nu(\mathrm{d}x) f(x) = \int_E \mu(\mathrm{d}x) p(x) f(x) \qquad f \in \mathcal{E}_+.$$

If we use the alternative notation:

$$\int_E f \,\mathrm{d}\nu = \int_E pf \,\mathrm{d}\mu \qquad f \in \mathcal{E}_+.$$

Moreover, p is unique up to equivalence: if the equation above holds for another $\hat{p} \in \mathcal{E}_+$ then $\hat{p}(x) = p(x)$ for μ-almost every[a] $x \in \mathcal{E}_+$. This is an if and only if statement!
Also, p is called the **Radon-Nikodyn derivative** of ν with respect to μ:

$$\frac{\nu(\mathrm{d}x)}{\mu(dx)} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = p.$$

[a] This means that all the setes where this condition doesn't hold are neglegible when weighted with measure μ.

With all this alternative notation this thing honestly feels like trying to understand the Metal Gear Solid plot, where identical characters named Snake keep cloning each other and being triple crossed by everyone until you finally understand that the storyline never made sense in the first place and that Hideo Kojima writes his games like a fucking fanfiction.

Now everything should make more sense[19]. If we know that a given random variable (say, the exponential random variable) has a distribution $\mu(\mathrm{d}x)$ then we will be able to transform this in a distribution of the form Lebesgue measure$\cdot p(x)$ (we can lose f if f is constant). In this formulation $\mathrm{d}x$ stands for the Lebesgue measure and the second part of the equation (ce^{-cx}) is the $p(x)$, called **density function**. We can do this because we can see from the formula that this distribution, or measure, is indeed absolutely continuous with respect to the Lebesgue measure since we can express it in the form stated by the Radon-Nikodyn theorem. So $p(x) = ce^{-cx}$, $x \in \mathbb{R}_+$ is the density relative to μ. This should serve us as a demonstration that if we define the random variable we get the distribution/measure (remember! distributions are measures!) and vice versa.

It is interesting[20] to see that also discrete random variable turns out to be absolutely continuous... But not with respect to the Lebesgue measure. To exact, discrete random variables are absolutely continuous with respect to the *counting* measure. And here's why to do all this shit we need the Lebesgue integral: by changing the measure we are using to compute the integral, we can use just one object (the probability distribution) to treat both discrete (using a counting measure, which gives us the cumulative distribution function in the form of a sum) and continuous random variables (using the Lebesgue measure, which gives us the cumulative distribution function in the form of a Riemann integral.)

[19] Enviable optimism.
[20] Debatable claim.

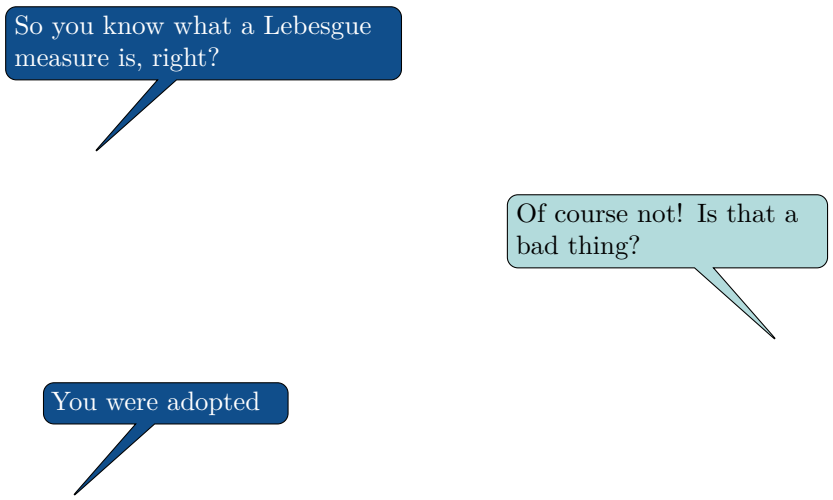


Figure 4: Actual conversation happened between me and Professor Lods.

So we're due for a little refresh on what the hell a Lebesgue measure is. I'm sorry[21] for our mathematician friends but I need this to be written loud and clear. This is from Professor Lods' Lecture notes from the pre-course in Measure Theory, with the hope that one day I'll be skilled like he is with LaTeX typesetting.

Revise with Kotatsu!

Let's have a quick refresh about the Lebesgue measure over the measurable space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Let $S = \mathbb{R}$. First of all, what is an algebra?

Definition 1.17

A collection Σ_0 of subsets of S is called an algebra on S if:

- $S \in \Sigma_0$;
- if $A \in \Sigma_0$ then $A^c \in \Sigma_0$ where $A^c = S \setminus A$ is the complementary of A;
- if $A, B \in \Sigma_0$ then $A \cup B \in \Sigma_0$.

We also need the concept of pre-measure, which is basically a measure but defined on an algebra (instead of a σ-algebra):

Definition 1.18

Let Σ_0 be an algebra on S (not necessarily a σ-algebra. A mapping $\ell : \Sigma_0 \mapsto [0, \infty]$ is said to be a **pre-measure** on Σ_0 if $\ell(\varnothing) = 0$ and for any pairwise disjoints $\{A_n\}_n \subset \Sigma_0$ with $\bigcup_n A_n \in \Sigma_0$ it holds:

$$\ell\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \ell A_n.$$

Moreover, a pre-measure ℓ is said to be σ-finite on Σ_0 if there exists a sequence $\{A_n\}_n \subset \Sigma_0$ with $\bigcup_n A_n \in \Sigma_0$ and $\ell(A_n) < \infty$ for any $n \in \mathbb{N}$.

We can immediately see that $\bigcup_n A_n \in \Sigma_0$ is an additional assumption: in σ-algebras this assumption is always met. So if Σ_0 is a σ-algebra any measure on Σ_0 is a pre-measure. We also need one more thing: the **Caratheodory's extension Theorem**.

[21] Not really. I mean, I'm sorry for the fact that they *are* mathematicians but that's where my compassion starts and ends.

Theorem 1.4

Charatheodory's extension Theorem: Let S be a given set and let Σ_0 be an algebra on S and $\Sigma = \sigma(\Sigma_0)$. If $\ell : \Sigma_0 \mapsto [0, \infty]$ is a pre-measure on (S, Σ_0) then there exists a measure μ on (S, Σ) such that

$$\mu(A) = \ell(A) \qquad \forall\, A \in \Sigma_0.$$

Moreover, if ℓ is a σ-finite pre-measure on Σ_0, then such a measure μ on (S, Σ) is unique and σ-finite.

Apparently this is one of the principal results in measure theory since it allows to construct measures well-adapted to practical situations: once such measures are constructed, Caratheodory's theorem can go fuck itself off. But the most important question is: why do we care about these total nerds? Because we can now define

$$\mathcal{C}_0 = \{[a, b) : -\infty \leqslant a \leqslant b \leqslant \infty \in \mathbb{R}\}$$

and let

$$\Sigma_0 = \left\{ \bigcup_{j=1}^{N} I - J : I_j \in \mathcal{C}_0 \; \forall\, j, I_i \cap I_j = \varnothing \text{ if } i \neq j, N \in \mathbb{N} \right\}$$

We can prove without major difficulty that Σ_0 is an algebra on $\mathbb{R}$. Let's define a pre-measure on Σ_0 by setting:

- $\ell([a, b)) = b - a$ for any $b \geqslant a$;
- $\ell((-\infty, b)) = \ell((a, \infty)) = \ell(\mathbb{R}) = +\infty$;
- $\ell\left(\bigcup_{j=1}^{N} I_j\right) = \sum_{j=1}^{N} \ell(I_j)$ if $\{I_j\}_{j=1,\dots,N} \subset \mathcal{C}_0$ are pairwise disjoints.

It can be checked that this newly defined measure is σ-finite Remember that $\sigma(\Sigma_0) = \sigma(\mathcal{C}_0) = \mathcal{B}_{\mathbb{R}}$, which is the Borel σ-algebra. Consider, additionally, the result of the Charatheodory's extension Theorem. By stitching all of these amenities together we get:

Theorem 1.5

There exists a unique measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that we denote λ (or $\mathfrak{m}$) and such that

$$\lambda([a, b)) = b - a \qquad \forall\, a < b.$$

We call this measure the **Lebesgue measure** on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

So we just learnt how to FIND A FUCKING INTERVAL ON THE REAL LINE?

Look, I know that can be seen as useless formalism. And indeed it is! We just formalized the notion of *length*. If we take it to 2-dimensional spaces we end up with the notion of area, in 3-dimensional spaces is the notion of volume... and so on.

Huh. This makes sense. So this is the notion of length when everything, including the real line, is a set. Kinda seems like the solution to a problem we ourselves created...

Remark

We can define in the same way the Lebesgue measure on $(I, \mathcal{B}_I)$ for all $I \subset \mathbb{R}$.

Remark

The measure space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \lambda)$ is σ-finite since $([-n, n))_n \nearrow \mathbb{R}$ but is not finite since

$$\lambda(\mathbb{R}) = \lim_n \lambda([-n, n)) = \lim_n 2n = \infty$$

Yeah, that mysterious measure I was talking about before to connect Riemann and Lebesgue integration was the Lebesgue measure. We often just write $\mathrm{d}x$ to express the Lebesgue measure (which is what we did on example 1.2 about the exponential random variable), but the meaning is always the same, with a striking similarity to the concept of Riemann integration: take the Lebesgue measure of a smaller and smaller element of our $\mathcal{B}_{\overline{\mathbb{R}}}$ set, use it to "weight" (read: multiply) the value of the function for that element and then sum it all up together. What we end up with is basically a serie of simple functions that slice "horizontally" the co-domain of the function. Keep reducing the size and you end up with the Lebesgue integral. Of course, when the measure is the Lebesgue measure, the Riemann and the Lebesgue integral have a really similar interpretation.

Back to our topic: the exponential random variable is absolutely continuous with respect to the Lebesgue measure. It is interesting to see that also discrete random variables turn out to be absolutely continuous: the difference is that they are not absolutely continuous to the Lebesgue measure, but the *counting* measure. At the undergrad level we are used to say that a random variable is either discrete or absolutely continuous, buy this was ultimately a lie[22].

Example 1.3

Gamma distribution[a]: Consider a random variable taking values in $\mathbb{R}_+$ and consider as a σ-algebra the Borel σ-algebra $\mathcal{B}_{\mathbb{R}_+}$. The distribution of the Gamma random variable is the following:

$$\mu(\mathrm{d}x) = \mathrm{d}x \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}, \qquad \text{with} \quad \begin{matrix} a > 0, \\ c > 0, \\ x \in \mathbb{R}_+ \end{matrix} .$$

Here $\Gamma(a)$ is the *Gamma function:*

$$\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} \, \mathrm{d}x.$$

[22] Measure theory turns truths into lies. Truly a demonic machinery.

The Gamma function is one of the most famous special functions that comes up almost everywhere. This definition of Gamma function is valid just for positive values and can be seen as a *Laplace transform*[b] or as a *Mellin transform*[c]. The first parameter a is called *shape parameter*; the parameter c is called *scale parameter*. This distribution is also continuous with respect to the Lebesgue measure.
We have some special cases of the Gamma distribution but Federico Polito doesn't really care about. Just know that the χ^2 distribution is a special case of the Gamma random variable.

[a] She factorials on my γ 'till I β.
[b] $\mathcal{L}\{f\}(s) = \int_0^{+\infty} f(t)e^{st}\,\mathrm{d}t$ where s is a complex number $s = a + ib$.
[c] $\mathcal{M}[f; s] \equiv F(s) = \int_0^{+\infty} f(t)t^{s-1}\,\mathrm{d}t$ where $s = a + ib$.

Example 1.4

This is a certified hood classic: **Gaussian distribution**.
Consider a random variable takinv values in $\mathbb{R}$. Of course we consider $\mathcal{B}_{\overline{\mathbb{R}}}$ and the distribution is (notice how also this one is absolutely continuous with respect to the Lebesgue measure):

$$\mu(\mathrm{d}x) = \mathrm{d}x \cdot \underbrace{\frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-1)}{2b}}}_{p(x)}, \qquad \begin{array}{l} a \in \mathbb{R}, \\ b > 0 \in \mathbb{R}, \\ x \in \mathbb{R}_+ \end{array} .$$

Of course, a is called the *mean* of the distribution and b is called the *variance*.

Example 1.5

This is a random variable that stems from two independent random variables having Gamma distribution. Consider γ_a (distribution of a Gamma random variable with parameters a and $c = 1$) and γ_b (distribution of a Gamma random variable with parameters b and $c = 1$). So, two gammas with different shape parameter.
Let $X \sim \gamma_a$ and $Y \sim \gamma_b$. Moreover, let them be independent. This is a random vector (X, Y) with two components... What is its distribution?

$$\pi(\mathrm{d}x, \mathrm{d}y) = \underbrace{\gamma_a(\mathrm{d}x) \cdot \gamma_b(\mathrm{d}y)}_{\text{because of independency}} = \mathrm{d}x\,\mathrm{d}y \frac{e^{-x}x^{a-1}}{\Gamma(a)} \cdot \frac{e^{-y}y^{b-1}}{\Gamma(b)}.$$

So it's easy to build joint distributions when the random variables are independent[a].

[a] Well no shit. Even I can multiply two numbers

Example 1.6

Gaussian random variable with exponential variance.
Here the variance is random and is distributed exponentially[a]. Consider a random variable X taking values in $\mathbb{R}_+$ and a random variable Y taking values in $\mathbb{R}$.
Here we are again in presence of a random vector. The distribution is the following:

$$\pi(\mathrm{d}x, \mathrm{d}y) = \mathrm{d}x\,\mathrm{d}y \cdot ce^{-cx} \frac{1}{\sqrt{2\pi x}} e^{-\frac{y^2}{2x}} \qquad \begin{array}{l} x \in \mathbb{R}_+, \\ y \in \mathbb{R} \end{array} .$$

Remark

π in this case has a special form: it has the form

$$\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y).$$

In particular, here $\mu(\mathrm{d}x)$ is

$$\mathrm{d}x\,\mathrm{d}y \cdot ce^{-cx}\frac{1}{\sqrt{2\pi x}}e^{-\frac{y^2}{2x}},$$

which is a docile exponential function, and $K(x, \mathrm{d}y)$ is

$$\mathrm{d}x\,\mathrm{d}y \cdot ce^{-cx}\frac{1}{\sqrt{2\pi x}}e^{-\frac{y^2}{2x}}.$$

In this case $K(x, \mathrm{d}y)$ cannot be the distribution fo Y, because it has some x in it. So this distribution is not simply the product of marginal distribution. But let's take a closer look to the form $\mu(\mathrm{d}x)K(x, \mathrm{d}y)$. $\mu(\mathrm{d}x)$ is certainly a measure, but what about $K(x, \mathrm{d}y)$?

[a] because humans should never have the hubris to meddle with the horrorific world of random necessities that the Gods have laid before us.

2 Transition kernels

Turns out that $K(x, \mathrm{d}y)$, depending on x, is connected to the other measure $\mu(\mathrm{d}x)$.
The object $K(x, \mathrm{d}y)$ is called **transition kernel**[23] and it's very important. Not that here $K(x, \mathrm{d}y)$ has the form

$$B \mapsto K(x, B)$$

It should be seen as a set function. Why? Just look back to $\pi(\mathrm{d}x, \mathrm{d}y)$. It is in differential form, but if we integrate the joint distribution against one set in the product space we get π evaluated on that subset of the product function (it is a measure... measures are always set functions). So also $\mu(\mathrm{d}x)K(x, \mathrm{d}y)$ should be a set function. $\mu(\mathrm{d}x)$ is of course a set function (it is a measure...[24]) and so should be $K(x, \mathrm{d}y)$. The presence of x in $K(x, \mathrm{d}y)$ is what "links" the X coordinate with the Y coordinate of the random vector. The Gamma example was about two independent random variables: here it is impossible to obtain the product of measures. Transition kernels are incredibly important because they are ultimately connected with the *structure of dependency* between random variables: that's why this course is going to bust our ball into the oblivion about them[25].

Figure 5: Yeah, no more Metal Gear Solid jokes after this one. I promise[26].

[23] Kernel? Colonel? I thought we were over with the Metal Gear Solid jokes.
[24] All this passive-aggressiveness for what?
[25] Professor Polito jokes that every year people complain about the abstractness of kernels and laughs about it. I'm happy that his sense of humor has been left untouched by my slightly scathing EDUMETER review of this course.

Remember in the undergraduate courses: when there was dependence we usually expressed it with the *conditional probability*.
Transition kernels let us manage random vectors that are not trivial and that have a dependence relationship with other random vectors.
Think about the previous example: the marginal distribution ν of Y has the form

$$\begin{aligned}\nu(B) = \pi(\mathbb{R}_+ \times B) &= \int_{\mathbb{R}_+} \mu(\mathrm{d}x) K(x, B), \qquad B \in \mathcal{B}_{\overline{\mathbb{R}}} \\ &= \int_{\mathbb{R}_+ \times B} \pi(\mathrm{d}x, \mathrm{d}y) = \int_{\mathbb{R}_+ \times B} \mathrm{d}x\, \mathrm{d}y \cdot ce^{-cx} \frac{1}{\sqrt{2\pi x}} e^{-\frac{y^2}{2x}} \\ &= \int_B \mathrm{d}y\, n(y), \qquad \text{where } n(y) = \int_0^\infty \mathrm{d}x \cdot ce^{-cx} \frac{1}{\sqrt{2\pi x}} e^{-\frac{y^2}{2x}}.\end{aligned}$$

So we have that the marginal distribution $\nu(B)$ is written as $\int_B \mathrm{d}y\, n(y)$ and is therefore absolutely continuous with respect to the Lebesgue measure. We could actually solve this integral ($n(y)$ is a closed form called *two-sided exponent*). So we now have the marginal of Y and the marginal of X and we immediately realize that if we multiply the two densities we do not obtain the joint density (because they are dependent[27]).

So we are now ready to define the concept of transition kernel.

Definition 2.1

Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Let K be a mapping from $E \times \mathcal{F}$ into $\overline{\mathbb{R}}_+$. Then K is called **transition kernel** from space $(E, \mathcal{E})$ into space $(F, \mathcal{F})$ if:

- the mapping $x \mapsto K(x, B)$ is $\mathcal{E}$-measurable $\forall\, B \in \mathcal{F}$;
- the mapping $B \mapsto K(x, B)$ (the second mapping of the kernel, the one regarding the set) is a measure $\forall\, x \in E$.

We can consider the transition kernel a hybrid object: if we look at it with respect to the first variable it is a *measurable function*, if we look at it with respect to the second variable it is a *measure*.

Example 2.1

Take ν, a finite measure on $(F, \mathcal{F})$ and take k, a positive function on $(E \times F)$ which is measurable with respect to $\mathcal{E} \otimes \mathcal{F}$, the product σ-algebra. Then, we integrate

$$\int_B \nu(\mathrm{d}y) k(x, y) \qquad \begin{matrix} B \in \mathcal{F} \\ x \in E \end{matrix}$$

We see how this object depends on x and on the choice of B (a function of x and B...). It defines a transition kernel

$$K(x, B) = \int_B \nu(\mathrm{d}y) k(x, y) \qquad \begin{matrix} B \in \mathcal{F} \\ x \in E \end{matrix}$$

from $(E, \mathcal{E})$ into $(F, |F)$.

Theorem 2.1

This theorem tells us what we can do with a kernel. Let K be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then,

1. we have
$$\int_F K(k, \mathrm{d}y) f(y) \qquad \text{with } x \in E,\ f \in \mathcal{F}_+.$$
This operation defines a function $Kf \in \mathcal{E}_+\ \forall\, f \in \mathcal{F}_+$.

[26] This is not, in fact, the last Metal Gear Solid joke.
[27] OK I GET IT.

Note the notation!

The notation $f \in \mathcal{F}_+$ in Çinlar is either the σ-algebra $\mathcal{F}$ or the set of functions measurable with respect to the σ-algebra $\mathcal{F}$. Another one of the many ways Çinlar chooses to sadden my day. But my revenge is on the way.

First of all we integrate the kernel with respect to the second variable (which basically means we use it as a measure on F). The integrand is F and since this function is a measure with respect to the second variable I can integrate $\mathcal{F}_+$-measurable functions. Remember that $\mu f = \mu(f) = \int_E f(x)\mu(\mathrm{d}x) = \int_E f \,\mathrm{d}\mu$.
The integration over F "takes out" one part of the kernel from the equation (the measure part) so that we can write:

$$Kf(x) = \int_F K(k, \mathrm{d}y) f(y);$$

2. we want to use the kernel with respect to the first variable (obtaining a $\mathcal{E}$-measurable function), so we integrate

$$\int_E \mu(\mathrm{d}x) K(x, B)$$

Remember that we must integrate the kernel with respect to a measure μ that is attached to the space $(E, \mathcal{E})$. This operation (since we remove the "function" part of the kernel) defines a *measure* μK on $(F, \mathcal{F})$ for each measure μ on $(E, \mathcal{E})$ and we can write:

$$\mu K(B) = \int_E \mu(\mathrm{d}x) K(x, B);$$

3. now we want to integrate everything: we consider the measure μK, take f and calculate its integral with respect to the measure μK

$$(\mu K) f.$$

With $f \in \mathcal{F}_+$. We can now link the function obtained in step 1 and the measure obtained in step 2:

$$(\mu K) f = \mu(Kf) = \int_E \mu(\mathrm{d}x) \cdot \int_F K(x, \mathrm{d}y) f(y)$$

for every choice of measure μ on $(E, \mathcal{E})$ and for every choice of $f \in \mathcal{F}_+$

Remember that here $(\mu K) f$ is shortened notation: μf means $\int_E f(x)\mu(\mathrm{d}x)$. We are NOT applying the measure to the function!
Right... but if we choose $f \in \mathcal{E}$ (f is $\mathcal{E}$-measurable) as the indicator function $f = \mathbb{1}_B, \quad B \in \mathcal{E}$ (the most simple example of $\mathcal{E}$-measurable function), this happens:

$$\begin{aligned} \mu \mathbb{1}_B &= \int_E \mathbb{1}_B(x)\mu(\mathrm{d}x) \\ &= \int_B \mu(\mathrm{d}x) = \mu(B) \end{aligned}$$

This is the reason behind this kind of notation that switches between measures and integrals. This is interesting and useful, if "interesting and useful" was slang for "boring and useless". So technically yeah, we *are* applying the measure to the function in the sense that we are weighing the function with the measure μ.

Proof

1 First of all, remember that Kf resulting from the integral is a well defined function but that is not enough: we need a $\mathcal{E}$-measurable function of Kf. We need to proceed in a constructive way: we have to think about $\mathcal{E}$-measurable functions, in whose space there are different type of functions: indicators, simple function, positive functions, positive or negative function. We start from simple function and then extend this

result to positive functions and then to a broader class.
First consider f to be a simple function with its canonical representation:

$$f = \sum_{i=1}^{n} b_i \mathbb{1}_{B_i}$$

for given weights b_i and given sets B_i. For this function we consider

$$Kf(x) = \sum_{i=1}^{n} b_i \underbrace{K(x, B_i)}_{\substack{\mathcal{E}\text{-measurable} \\ \text{with respect to } x}}.$$

So it is a linear combination of $\mathcal{E}$-measurable functions and therefore $Kf \in \mathcal{E}_+$ when f is simple.

We now want to extend the proof to the subclass of positive measurable functions. Take f positive. We know that we can approximate positive functions by means of simple functions, but reducing the "step" of the simple functions (discretizing the original function) by means of an auxiliary function called **dyadic function** (see below), thus producing a sequence of simple functions. So we have $f \in \mathcal{F}_+$ and $f_n = \mathscr{d}_n \circ f$.

Lemma 2.1

Each $\mathscr{d}_n$ is an increasing right-continuous simple function on $\overline{\mathbb{R}}_+$ and $\mathscr{d}_n(r)$ increases to $r \; \forall r \in \overline{\mathbb{R}}_+$ as $n \to \infty$.

So as soon as n goes to infinity we get the original function. Moreover, remember the theorem that gives measurability of a function:

Revise with Kotatsu!

Theorem 2.2

A positive function on $(E, \mathcal{E})$ is $\mathcal{E}$-measurable if and only if it is the limit of an increasing sequence of positive simple functions.

Theorem 2.3

Let $(f_n)_n$ be a sequence of positive measurable functions in (S, Σ) with respect to a measure Σ such that $f_n \nearrow f$ almost everywhere on the set S, Then

$$\lim_n \int_S f_n \, \mathrm{d}\mu = \int_S f \, \mathrm{d}\mu$$

where the sequence $\left(\int_S f_n \, \mathrm{d}\mu\right)_n$ is a nondecreasing sequence.

We consider, as we said before, the discretization

$$f_n = \mathscr{d}_n \circ f.$$

What happens to $Kf(x)$? In this case, it is defined as:

$$Kf(x) = \lim_{n \to \infty} Kf_n(x) \qquad \forall x$$

and this is true for the monotone convergence theorem for the measure $B \mapsto K(x, B)$. Then Kf is $\mathcal{E}$-measurable being the limit of the $\mathcal{E}$-measurable sequence of functions $(Kf_n)_n$.

2–3 Fix a measure μ on $(E, \mathcal{E})$ and define a functional

$$L : \mathcal{F}_+ \mapsto \overline{\mathbb{R}}_+$$

by setting

$$L(f) = \mu(Kf)$$

that is, we integrate Kf with respect to the measure μ. We note that

i.) if $f = 0$ then $L(f) = 0$;

ii.) if $f, g \in \mathcal{F}_+$ with $a, b \in \mathbb{R}_+$ then

$$\begin{aligned} L(af + bg) &= \mu\big(K(af + bg)\big) \\ &= a\mu(Kf) + n\mu(Kg) \\ &= aL(f) + bL(g). \end{aligned}$$

So the functional is a linear function.

iii.) if $(f_n)_n \subset \mathcal{F}_+$ and $f_n \nearrow f$ then $Kf_n \nearrow Kf$ and this is true by monotone convergence theorem with respect to integrals with respect to the measure $B \mapsto K(x, B)$.

iv.) note that

$$L(f_n) = \mu(Kf_n) \nearrow \mu(Kf)$$

again because of the monotone convergence theorem with respect to the measure μ.

So given that i.), ii.), iii.) and iv.) hold (and recurring to the theorem 4.21 of the Cinlar book[a]) we have that there exists a measure ν on $(F, \mathcal{F})$ such that

$$L(f) = \nu f \qquad \forall\, f \in \mathcal{F}_+,$$

Note that if we specifically take the function $f = \mathbb{1}_B$, $B \in \mathcal{F}$ we see that

$$\begin{aligned} \nu(B) = \nu\mathbb{1}_B = L(\mathbb{1}_B) = \mu(K\mathbb{1}_B) &= \mu\left(\int_B K(x, \mathrm{d}y)\right) \\ = \mu K(x, B) &= \int_E K(x, B)\mu\,\mathrm{d}x = \mu K(B). \end{aligned}$$

Then $\nu \equiv \mu K$, that is μK is a measure on $(F, \mathcal{F})$ and

$$(\mu K)f = \nu f = L(f) = \mu(Kf).$$

□

[a] yeah not gonna check that.

Definition 2.2

The **dyadic function** is defined as:

$$\mathscr{d}_n(r) = \sum_{k=1}^{n\cdot 2^n} \frac{k-1}{2^n} \mathbb{1}_{\left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)}(r) + n\mathbb{1}_{[n, +\infty)}(r), \qquad r \in \overline{\mathbb{R}}_+$$

So we basically have two different sections in this function: after n the value of this function is equal to n, otherwise it is a step function in the sequence of interval from the start to n. To see the shape of this function we could see some examples.

Example 2.2

Take $n = 1$. We can calculate this dyadic function obtaining

$$\begin{aligned} \mathscr{d}_1(r) &= 0, \qquad r \in \left[0, \frac{1}{2}\right) \\ \mathscr{d}_1(r) &= \frac{1}{2}, \qquad r \in \left[\frac{1}{2}, 1\right) \\ \mathscr{d}_1(r) &= 1, \qquad r \geqslant 1. \end{aligned}$$

The function is right-continuos and a step function.

Example 2.3

$n = 1$. Try to do it by hand.

$$\begin{aligned}
\mathscr{d}_2(r) &= 0, && r \in \left[0, \frac{1}{4}\right) \\
\mathscr{d}_2(r) &= \frac{1}{4}, && r \in \left[\frac{1}{4}, \frac{1}{2}\right) \\
\mathscr{d}_2(r) &= \frac{1}{2}, && r \in \left[\frac{1}{2}, \frac{3}{4}\right) \\
\mathscr{d}_2(r) &= \frac{3}{4}, && r \in \left[\frac{3}{4}, 1\right) \\
\mathscr{d}_2(r) &= 1, && r \in \left[1, \frac{5}{4}\right) \\
\mathscr{d}_2(r) &= \frac{5}{4}, && r \in \left[\frac{5}{4}, \frac{3}{2}\right) \\
\mathscr{d}_2(r) &= \frac{3}{2}, && r \in \left[\frac{3}{2}, \frac{7}{4}\right) \\
\mathscr{d}_2(r) &= \frac{7}{4}, && r \in \left[\frac{7}{4}, 2\right) \\
\mathscr{d}_2(r) &= 2, && r \geqslant 2.
\end{aligned}$$

2.1 Products of kernels

Let K be a kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ and let K be a kernel from $(F, \mathcal{F})$ into $(G, \mathcal{G})$.

Definition 2.3

The product of K and L is the transition kernel from $(E, \mathcal{E})$ into $(G, \mathcal{G})$ such that $(KL)f = K(LF)$ for $f \in \mathcal{G}_+$.

Let's now fill in some information for transition kernels.

Definition 2.4

A transition kernel from $(E, \mathcal{E})$ into $(E, \mathcal{E})$ is actually called **transition kernel on** $(E, \mathcal{E})$.

This is actually the most common transition kernel in probability[28], since random variables take values in the same space.

Definition 2.5

A transition kernel on $(E, \mathcal{E})$ is called **Markov kernel** if

$$K(x, E) = 1 \qquad \forall\, x \in E.$$

It is called **sub-Markov** if

$$K(x, E) \leqslant 1 \qquad \forall\, x \in E.$$

So here we are, we summoned Markov for the first time in this course.

[28] Oh right we are studying probability theory. Thanks for reminding!

Figure 6: Andrej Andreevič Markov if he was cool.

Definition 2.6

Given a transition kernel on $(E, \mathcal{E})$ its **powers** are define recursively as follows:

$$\begin{aligned} K^0 &= I \\ K^1 &= K \\ &\vdots \\ K^n &= K \cdot K^{n-1} \end{aligned}$$

I is the identity kernel, i.e. $I(x, A) = \delta_x(A) = \mathbb{1}_A(x) \quad \forall\, x \in E,\ A \in \mathcal{E}$

Remark

$$If = f;\ \mu I = \mu;\ \mu If = \mu F;\ IK = KI = K$$

Definition 2.7

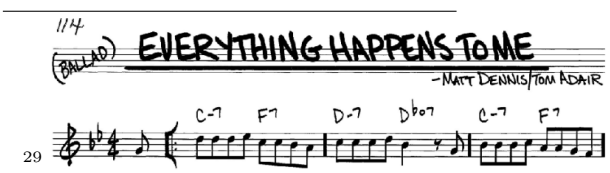
A transition Kernel K from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ is said to be:

- **finite** if $K(x, F) < \infty$ for $\forall\, x \in E$;
- **bounded** if $x \mapsto K(x, F)$ is bounded;
- **σ-finite** if $B \mapsto K(x, B)$ is σ-finite for $\forall\, x \in E$;
- **σ-bounded** if it exists a measurable partition $(F_n)_n$ of F such that $x \mapsto K(x, F_n)$ is bounded for $\forall\, n$;
- **Σ-finite** if $K = \sum_{n=1}^{\infty} K_n$ for some sequence of finite kernels $(K_n)_n$.
- **Σ-bounded** if the K_n can be chose to be bounded.

Definition 2.8

If $K(x, F) = 1\ \forall\, x \in E$ then the kernel K is said to be a **transition probability kernel**.

We now turn to extending measure to product spaces with respect to kernels[29]. In order to formally solve this problem we need the following proposition.



Proposition 2.1

Let K be a Σ-finite kernel (the most general property we can think of) from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. We consider measurable functions with respect to the product space: for every positive funtion $f \in \mathcal{E} \otimes \mathcal{F}$ we have that:

$$Tf(x) = \int_F K(x, \mathrm{d}y) f(x, y) \qquad x \in E$$

And this object defines a function $Tf \in \mathcal{E}_+$. This is a similar operation to the previous theorem. Moreover the transformation $T : (\mathcal{E} \otimes \mathcal{F})_+ \mapsto \mathcal{E}_+$ is linear and continuous under increasing limits, that is:

a) if we take $f, g \in (\mathcal{E} \otimes \mathcal{F})_+$ and $a, b \in \mathbb{R}_+$ we have

$$T(af + bf) = aTf + bTg;$$

b) $Tf_n \nearrow Tf$ for $\forall$ sequence $(f_n)_n \subset (\mathcal{E} \otimes \mathcal{F})_+$ with $f_n \nearrow f$.

So, similarly to the previous theorem we have constructed this operator Tf which operates on the function set $(\mathcal{E} \otimes \mathcal{F})_+$ giving us a positive $\mathcal{E}$-measurable function. So we can start from this function to build a method to construct measures on the product space $(\mathcal{E} \otimes \mathcal{F})_+$ with its related σ-algebra.

Theorem 2.4

Extension of measures on product spaces.
Let μ be a measure on the measurable space $(E, \mathcal{E})$. Let K be a Σ-finite[a] transition kernel from space $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then:

1. if we take our function $f(x, y)$, integrate it against our kernel $K(x, \mathrm{d}x)$ over F and then integrate again against measure μ over E, the operation

$$\pi f = \int_E \mu(\mathrm{d}x) \int_F K(x, \mathrm{d}y) f(x, y)$$

defines a measure π on $(E \times F, \mathcal{E} \otimes \mathcal{F})$;

2. if μ is σ-finite and K is σ-bounded then π is σ-finite and it is the unique measure on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ satisfying

$$\pi(A \times B) = \int_A \mu(\mathrm{d}x) K(x, B) \qquad \forall\, A \in \mathcal{E}, B, \in \mathcal{F}.$$

[a] Erm... what the sigma?

So by means of kernels we are able to define measures on product spaces in this way[30].

Remark

When the kernel that we used to extend the measure has the form

$$K(x, B) = \nu(B)$$

for some Σ-finite measure ν on $(F, \mathcal{F})$, which means that it only depends on B and is therefore a measure, then we obtain the **product measure**

$$\pi = \mu\nu.$$

But what should we do when we have more than 2 spaces[31]? On finite product spaces we introduce in the same manner the product measure

$$\pi = \mu_1 \mu_2 \cdots \mu_n$$

[30] I crave to be released from this prison of flesh.
[31] I have an idea.

where μ_i is Σ-finite on $(E_i, \mathcal{E}_i)$ for $\forall\, i = 1, \dots, n$. So we can induce the presence of another measure from the product measure using the kernels, without assuming independence in the construction.

Example 2.4

Here we tackle $n = 3$. Take μ_1 on $(E_1, \mathcal{E}_1)$, the transition kernel K_2 from $(E_1, \mathcal{E}_1)$ into $(E_2, \mathcal{E}_2)$ and the transition kernel K_3 from $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ into $(E_3, \mathcal{E}_3)$. Not how we only defined the first measure and then only the transition kernels. We will use these fucking kernels to "move" from measure to measure and from space to space. We have

$$\pi f = \int_{E_1} \mu(\mathrm{d}x_1) \int_{E_2} K_2(x, \mathrm{d}x_2) \int_{E_3} K\Big((x_1, x_2), \mathrm{d}x_3\Big) f(x_1, x_2, x_3)$$

with f positive and $(\mathcal{E}_1 \otimes \mathcal{E}_2 \otimes \mathcal{E}_3)$-measurable. So we defined a measure π which is different from the product measure we recalled earlier (actually that product measure is a special case of this measure) and π is a measure on the 3-dimensional product space. Writing π in differential form:

$$\begin{aligned} \pi(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) &= \mu(\mathrm{d}x_1) K_2(x, \mathrm{d}x_2) K\Big((x_1, x_2), \mathrm{d}x_3\Big) \\ &= \mu_1 K_2 K_3. \end{aligned}$$

This is for finite product spaces but we can also extend this to infinte product spaces[32]. What should we expect now?

3 Expectation

Well that was a cheap joke. We already know the meaning of expectation from our undergraduate courses... but here we will rock our world and learn some new interpretations. Let's start with measure theory[33]. In probability the concept of expectation is strictly tied with the concept of integral: we could say they are almost the same thing.

Definition 3.1

Let X be a real valued ($\overline{\mathbb{R}}$) random variable on the probability space $(\Omega, \mathcal{H}, \mathbb{P})$. The **expectation** or **expected value** of X is

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

Note the notation!

$$\mathbb{E}X = \int_\Omega X(\omega) \mathbb{P}(\mathrm{d}\omega) = \int_\Omega x \, \mathrm{d}\mathbb{P}.$$

There is always the compact notation we use for integrals

$$\mathbb{P}X.$$

We are integrating the function over the space where it is defined, but since we are using Lebesgue integration we need to integrate with respect to a measure... which is the probability measure $\mathbb{P}(\mathrm{d}\omega)$. This is a little bit different from the classic definition, but this is more correct.

Remark

The expectation of X exists if and only of the related integral exists. So the existence of the expectation is the existence of the integral.

[32] Why. Stop.
[33] We have a great time ahead, I see.

Consider the random variable X, with its positive part X^+ and its negative part X^-. Moreover, remember that

$$X = X^+ - X^-$$

where both the positive part *and* the negative part are positive functions (remember?[34]). If we apply the expectation to X, by the linearity of the integral operator we have that:

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

Where, in particular:

$$\mathbb{E}X^+ = \int_\Omega X^+ \mathbb{P}(\mathrm{d}\omega) = \begin{cases} < +\infty \\ +\infty \end{cases}$$

and

$$\mathbb{E}X^- = \int_\Omega X^- \mathbb{P}(\mathrm{d}\omega) = \begin{cases} < +\infty \\ +\infty \end{cases}$$

The problem arises when both the expectation of the positive part and the expectation of the negative part are simultaneously infinite.
Further, we say that $\mathbb{E}X$ exists finit e when at the same time both expectations are finite:

Figure 7: And then he turned himself into a pickle, funniest shit I've ever seen.

$$\mathbb{E}X^+ < +\infty, \quad X^- < +\infty.$$

In this case

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^- < \infty.$$

Finally, remember from undergraduate courses:

$$\mathbb{E}X^+ \mathbb{E}X^- = \mathbb{E}|X|$$

So if we check that the expectation of the absolute value is finite this will imply that the expectation of X is finite in both the negative and the positive part. A random variable with finite expectations is said to be **integrable**.

Remark

This is connected to the definition of expectation of a random variable that we already know. Consider the "change of variable" formula for Lebesgue integrals (see Çinlar , formula 5.3 page 30): consider $f \in \mathcal{E}$ and $h : (F, \mathcal{F}) \mapsto (E, \mathcal{E})$. We integrate

$$\int_F \nu(\mathrm{d}x) f\Big(h(x)\Big) = \int_E \mu(\mathrm{d}y) f(y)$$

where μ is a measure on $(E, \mathcal{E})$ and is the *image measure* of ν through the function h. For us:

- $h \equiv X$;
- $(F, \mathcal{F}) \equiv (\Omega, \mathcal{H})$;
- $\nu = \mathbb{P}$;
- μ is the distribution of X;

[34] No. Happy?

- $(E, \mathcal{E}) \equiv (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

So the formula becomes

$$\int_\Omega \mathbb{P}(\mathrm{d}\omega) f\Big(X(\omega)\Big) = \int_{\mathbb{R}} \mu(\mathrm{d}x) f(x)$$

with f Borel-measurable. Now the last step is to choose a specific function f, for which we choose the *identity function* so that

$$\int_\Omega \mathbb{P}(\mathrm{d}\omega) X(\omega) = \int_{\mathbb{R}} \mu(\mathrm{d}x) x.$$

Here's the formula that we used to calculate the expectation in the undergraduate years! To be honest, that's what you'll ever really use in the sad case of you working in this field. If the distribution is absolutely continuous with respect to the lebesgue measure we get

$$\int_{\mathbb{R}} \mu(\mathrm{d}x) x = \int_{\mathbb{R}} f_X(x)\, \mathrm{d}x \cdot x$$

otherwise, if it is absolutely continuous with respect to a counting measure we get

$$\int_{\mathbb{R}} \mu(\mathrm{d}x) x = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) \cdot x_i.$$

Note the notation!

Forget riemann integrals and sums, fucker, from now on you must learn to use

$$\int_{\mathbb{R}} \mu(\mathrm{d}x) x.$$

3.1 Properties of expectation

- **Positivity**:
$$X \geqslant 0 \implies \mathbb{E}X \geqslant 0.$$
Remember that we are talking about random variables, so when we say $X \geqslant 0$ we actually mean "$X \geqslant 0$ almost surely with respect to $\mathbb{P}$"[35].

- **Monotonicity**:
$$X \geqslant Y \geqslant 0 \implies \mathbb{E}X \geqslant \mathbb{E}Y.$$

- **Linearity**:
$$X, Y \geqslant 0 \implies \mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y.$$

- **Insensitivity**:
$$X, Y \geqslant 0,\ X = Y \text{ almost surely} \implies \mathbb{E}X = \mathbb{E}Y.$$

- **Monotone convergence**: if we have, for $X_n \geqslant -0$,
$$(X_n)_n \nearrow X \implies \mathbb{E}X_n \nearrow \mathbb{E}X.$$

- **Fatou's Lemma**: for $X_n \geqslant 0$ we have
$$(X_n)_n \geqslant 0 \implies \mathbb{E} \liminf X_n \leqslant \liminf \mathbb{E}X_n.$$

- **Dominated convergence**: if $(X_n)_n$ is a sequence of random variables such that $\forall n |X_n| \leqslant Y$ and Y has finite expectation (it is integrable) and $\lim_{n\to\infty} X_n$ exists, then
$$\mathbb{E} \lim_n X_n = \lim_n \mathbb{E}X_n.$$

[35] Because reality crumbles in front of the chaotic horror of randomness. A similar fate is reserved to my balls.

- **Bounded convergence**: if we have a sequence of random variables $(X_n)_n$ such that $|X_n| \leqslant b < \infty$ and $\lim_{n\to\infty} X_n$ exists, then

$$\mathbb{E} \lim_n X_n = \lim_n \mathbb{E} X_n.$$

Remark

X is positive (or non negative) and $\mathbb{E}X = 0$ if and only if $X = 0$ almost surely.

Remark

If we restrict $\mathbb{E}X$ and $\mathbb{E}Y$ on the subset H we get:

$$\text{if } \mathbb{E}X\mathbb{1}_H \geqslant \mathbb{E}Y\mathbb{1}_H \quad \forall H \implies X \geqslant Y \text{ a.s.}$$

Theorem 3.1

Let X be a random variable taking values in $(E, \mathcal{E})$ and be measurable relative to $\mathcal{H}$ and $\mathcal{E}$. If μ is the distribution of X then

$$\mathbb{E}f \circ X = \mu f, \qquad \forall f \in \mathcal{E}_+ \tag{$\star$}$$

Conversely, if $\star$ holds for some measure μ on $(E, \mathcal{E})$ and $\forall f \in \mathcal{E}_+$, then μ is the distribution of X.

Proof

The proof should be simple. The first statement is basically the change of variable formula, rephrasing the theorem on integration with respect to image measures. The second converse statement requires more thought. If $\star$ holds $\forall f \in \mathcal{E}_+$, then it holds also for $f = \mathbb{1}_A$ for $A \in \mathcal{E}$ and we have that if we want to calculate the measure

$$\mu(A) = \mu\mathbb{1}_A = \mathbb{E}\mathbb{1}_A \circ X = \mathbb{P}(X \in A).$$

So we have identified this measure with the distribution of the random variable and hence μ is the distribution of X. □

Example 3.1

1. The **variance** of the random variable X is

$$\mathbb{V}\text{ar}\, X = \mathbb{E}(X - \mathbb{E}X)^2.$$

2. Consider the expectation of an exponential transform of X:

$$\tilde{\mu}_r = \mathbb{E}e^{-rX} = \int_{\mathbb{R}_+} e^{-rx}\mu(\mathrm{d}x), \qquad r \in \mathbb{R}_+, x > 0.$$

This is known as the **Lapalce transform** of distribution μ. It is connected to the moment generating function of a distribution.

3. We could be interested in the expectation of another exponential transform:

$$\hat{\mu}_r = \mathbb{E}e^{irX} = \int_{\mathbb{R}} \mu(\mathrm{d}x)e^{irx}$$

and, exploiting the representation of complex numbers,

$$\hat{\mu}_r = \int_{\mathbb{R}} \mu(\mathrm{d}x)\cos(rx) + i\int_{\mathbb{R}} \mu(\mathrm{d}x)\sin(rx)$$

and this is called the **characteristic function** of μ.

4. Consider X, a random variable taking values in $\overline{\mathbb{N}}$. Consider

$$\mathbb{E}z^X = \sum_{n=1}^{\infty} z^n \cdot \mathbb{P}(X = n), \qquad z \in [0, 1].$$

This is called **probability generating function of X**.

These are all special kinds of expectations.

Remark

The expectation $\mathbb{E}X$ is in some sense "optimal": what the fuck? It is "optimal" because it is our best estimate of X. Imagine we are given the following integral:

$$f(a) = \int_\Omega \Big(X(\omega - a)\Big)^2 \mathbb{P}(\mathrm{d}\omega).$$

Let us now derive the minimum value for the number a, that is the best value that cancels out the random variable X. Just take the derivative with respect to a:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a} f(a) &= \int_\Omega \mathbb{P}(\mathrm{d}\omega)\Big(-2\Big(X(\omega) - a\Big)\Big) = 0 \\ &= 2a \int_\Omega \mathbb{P}(\mathrm{d}\omega) - 2\int_\Omega \mathbb{P}(\mathrm{d}\omega) X(\omega) = 0 \\ \implies a &= \underbrace{\int_\Omega \mathbb{P}(\mathrm{d}\omega) X(\omega)}_{\text{definition of } \mathbb{E}X}.\end{aligned}$$

We must now recall some famous inequalities.

- **Markov inequality**: this is for positive, real-valued random variables. Let X be a positive and real-valued random variable. We may be interested in the probability that X exceeds some positive value b. We know that

$$\mathbb{P}(X > b) \leqslant \frac{1}{b}\mathbb{E}X \qquad \forall\, b > 0.$$

- **Chebyshev inequality**: this is too for positive, real-valued random variables. Let X be a positive and real-valued random variable. What we have is that

$$\mathbb{P}\Big(|X - \mathbb{E}X| > \varepsilon\Big) \leqslant \frac{\mathbb{V}\mathrm{ar}\, X}{\varepsilon^2}, \qquad \varepsilon > 0.$$

- **Jensen inequality**: Let X be a real-valued random variable with finite expectation (integrable) and let f be a convex function on $\mathbb{R}$. Then

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

If the function is concave the inequality is the opposite.

Definition 3.2

Let X, Y be two real-valued random variables with finite variance. The **covariance** between X and Y is

$$\mathbb{C}\mathrm{ov}\,(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

If X and Y are independent, then $\mathbb{C}\mathrm{ov}\,(X, Y) = 0$ (try to prove it but it is easy even for me[36]).

[36] Proceeds *not* to do it.

Remark

If we want to calculate the variance of two real-valued random variables then

$$\mathbb{V}\mathrm{ar}\,(X+Y) = \mathbb{V}\mathrm{ar}\,(X) + \mathbb{V}\mathrm{ar}\,(Y) + 2\mathbb{C}\mathrm{ov}\,(X,Y).$$

This can be proven very easily. Of course, if X and Y are an independency then $\mathbb{C}\mathrm{ov}\,(X,Y) = 0$ and $\mathbb{V}\mathrm{ar}\,(X+Y) = \mathbb{V}\mathrm{ar}\,(X) + \mathbb{V}\mathrm{ar}\,(Y)$.

Consider $\mathbb{E}|X| < +\infty$. We think of random variables with finite expectation in absolute value as "well-behaved", without any crazy jumps or fluctuations. So it could be interesting to look for a way to treat all of these random variables with finite expectation in the same way.

Remark

A random variable (real-valued) random variable with finite expectations is called **integrable random variable**.

We are going back to functional analysis now. Brace yourselves.

3.2 L^p spaces

We start, as usual, with the $(\Omega, \mathcal{H}, \mathbb{P})$ probability space, with the real-valued random variable X and the parameter $p \in [1, \infty]$. Why doesn't p start from 0? It's dumbass mathematical reason we won't and shouldn't care about[37].
Define the so-called L^p-norm of X as:

$$\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}, \qquad p < +\infty.$$

Remember that expectations are integrals. We can also define a *sup-norm*:

$$\|X\|_\infty = \inf(b \in \mathbb{R}_+ : |X| \leqslant b \text{ a.s.}).$$

So, taking a random variable X and fixing p we can calculate a number to attach to this random variable.

Example 3.2

Take $p = 1$. Our L^p-norm would then be

$$\|X\|_1 = \mathbb{E}|X|.$$

No shit. But as long as p is finite we can technically write the L^p norm of the random variable. The same goes for $p = \infty$, but with the different definition. We will call the *sup-norm* **essential supremum of** X. This name has a meaning: imagine the random variable X "folded" by the absolute variable. We are interested in the number b for which the probability mass is more or left all to the left of this number.

Note the notation!

$$\mathrm{essup}(X) = \|X\|_\infty.$$

Remark

If $\|X\|_p = 0 \implies X = 0$ almost surely. Remember that the L^p-norm has an absolute value, so if the norm is 0 then the random variable must be 0 everywhere.

If we fix a number $c \geqslant 0$ we have

$$\|cX\|_p = c\,\|X\|_p.$$

[37] L in L^p stands for Lebesgue space. If $p < 0$ these spaces do not have the property of Lebesgue spaces so we don't care.

If we have $1 \leqslant p \leqslant q \leqslant +\infty$ then

$$0 \leqslant \|X\|_p \leqslant \|X\|_q \leqslant +\infty.$$

Definition 3.3

The collection of real-valued random variables X with $\|X\|_p < +\infty$, $p \in [1, \infty]$ is called $\mathbf{L^p}$.

So we are creating sub-collections of real-valued random variables such that they have a finite norm. The space L^p is a subset of the space of all random variables.

Remark

X is in L^p, $p \in [1, \infty)$ if and only if $|X|^p$ is integrable and X is in L^∞ if and only if X is almost surely bounded.

Remark

If $1 \leqslant p \leqslant q \leqslant +\infty$ then

$$L^q \subset L^p$$

So the L^p spaces are one inside the other, with L^∞ being the smallest of them all and L^1 being the biggest. Think about the meaning: L^1 is the space of all integrable random variables; L^2 is the space for which $|X|^2$ is integrable, but if $|X|^2$ is integrable so is $|X|$... and so on. We can thus prove certain properties for the whole class and this will extend to smaller classes.
The regularity of these functions is evident when we look at L^∞: the functions that belong in this space are very regular because they are almost surely bounded and therefore their support doesn't explode towards infinity[38].

Remark

Some useful inequalities...

- **Minkowsky inequality**:

$$\|X + Y\|_r \leqslant \|X\|_p - \|Y\|_p .$$

- **Hölder inequality**:

$$\|XY\|_r \leqslant \|X\|_p \|Y\|_q$$

  where $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$ for $p, q, r \in [1, \infty)$.

- **Jensen inequality**: Consider:
  - a convex domain D in $\mathbb{R}^d$;
  - a continuous and concave function $f : D \mapsto \mathbb{R}$;
  - a sequence of random variables $X_1, X_2, \ldots, X_d$ integrable (in L^1 space) for which $(X_1, X_2, \ldots, X_d) \in D$ almost surely.

  Then

$$\mathbb{E} f(X_1, X_2, \ldots, X_d) \leqslant f\Big(\mathbb{E}(X_1, X_2, \ldots, X_d)\Big).$$

When we talk about integrability, we have this useful lemma:

Lemma 3.1

Let X be a real-valued random variable. X is integrable if and only if

$$\lim_{b \to \infty} \mathbb{E}|X| \mathbb{1}_{(|X| > b)} = 0.$$

[38] Professor Polito said: "Good guys". :3

This lemma states that this is an equivalent condition for integrability. The random variable $|X|\mathbb{1}_{(|X|>b)} = 0$ is equal to 0 until b and equal to the absolute value of X after b. So we are effectively truncating the random variable but maintaining the right tail. This means that if we integrate only the tails of the random variable we wouldn't get much information (we can say that the tail is *thin*).

Proof

Let's call Z_b the truncated random variable $|X|\mathbb{1}_{(|X|>b)} = 0$. Note that $Z_b \leqslant |X|$ and that $Z_b \xrightarrow[b\to\infty]{} 0$. Since X is integrable, we use the dominated convergence theorem:

$$\lim_{b\to\infty} \mathbb{E}|X|\mathbb{1}_{(|X|>b)} = \mathbb{E}\Big[\lim_{b\to\infty} \mathbb{1}_{(|X|>b)}\Big] = 0.$$

The first statement is proved. Conversely, if $\mathbb{E}Z_b \to 0$, we can choose b large enough such that

$$\mathbb{E}Z_b \leqslant 1.$$

Then

$$|X| \leqslant b + Z_b.$$

This is not immediate but it's true. Due to linearity of expectation, we know that

$$\mathbb{E}|X| \leqslant b + \mathbb{E}Z_b$$

but we know that $\mathbb{E}Z_b \leqslant 1$ so

$$\mathbb{E}|X| \leqslant b + 1 \leqslant \infty.$$

□

3.3 Uniform integrability

Definition 3.4

A collection of random variables K is said to be **uniformly integrable** if

$$k(b) = \sup_{X\in K} \left(\mathbb{E}|X|\mathbb{1}_{(|X|>b)}\right)$$

goes to 0 as $b \to \infty$.

Note that it is a function of b. We consider the supremum of the collection to be "conservative": if the supremum goes to 0 then I know the whole collection will.

Remark

1. If K is finite and each $X \in K$ is integrable then K is uniformly integrable.
2. If K is dominated by an integrable random variable Z then it is uniformly integrable.
3. Uniform integrability of a collection K implies the so-called L^1-boundedness, which means
$$k \subset L^1 \quad \text{and} \quad k(0) = \sup_K \mathbb{E}|X| < \infty.$$
Note that $k(0)$ considers the whole random variable without truncation.

Proof

$$\mathbb{E}|x| \leqslant b + k(b) \qquad \forall\, X \in K.$$

So, since now every random variable is in K and K is uniformly integrable, we use this property of K to choose a value for b such that $k(b) \leqslant 1$ and therefore it is finite. □

4. We know that uniform integrability implies L^1... but the converse is not true. We can prove it by a counterexample.

Proof

Consider the probability space

$$\Big((0,1), \mathcal{B}_{(0,1)}, \underbrace{\lambda}_{\text{Lebesgue measure}}\Big).$$

Normally the Lebesgue measure is infinite on the whole support, but if we restrict the measure on the unit interval then the lebesgue measure has maximum value 1 and so it is a probability measure.
Consider the collection

$$K = (X_n)_{n\geqslant 1} \qquad \text{s.t. } X_n = \begin{cases} n, & \omega \leqslant \frac{1}{n} \\ 0, & \text{otherwise} \end{cases}$$

Note that

$$\forall\, n \quad \mathbb{E}X_n = 1$$

That is, K is L^1-bounded.
But if we calculate $k(b)$ we realize it is equal too 1 for each b, so

$$\mathbb{E}X_n \mathbb{1}_{(X_n > b)} = \mathbb{E}X_n = 1 \qquad \forall n > b$$

Therefore the collection K is *not* uniformly integral. □

5. If K is L^p-bounded with $p > 1$ then it is uniformly integrable with $f(x) = x^p$.
To prove this we recur to the following proposition:

Proposition 3.1

Suppose it exists a positive borel function f on $\mathbb{R}_+$ such that

$$\lim_{x\to\infty} \frac{f(x)}{x} = \infty$$

and write

$$c = \sup_{X\in K} \mathbb{E} f \circ |X|. \qquad (\star)$$

If $c \leqslant \infty$ then K is uniformly integrable.

Proof

We are talking about the absolute value of X, so we don't lose generality if we talk about positive random variables. Let us assume, then, that $X \in \mathbb{R}_+$. Let us replace f with $f \vee 1$ if necessary and assume $f \geqslant 1$.
Let $g(x) = \frac{x}{f)=(x)}$ and note that if we are looking at the tail of the random variable we get

$$X\mathbb{1}_{(X_n>b)} = (f \circ X)(g \circ X)\mathbb{1}_{(X_n>b)}.$$

This is true because $(f \circ X)(g \circ X) = \cancel{f(x)}\,\frac{x}{\cancel{f(x)}}$.
Let us now bound this quantity:

$$\begin{aligned} X\mathbb{1}_{(X_n>b)} &= (f \circ X)(g \circ X)\mathbb{1}_{(X_n>b)} \\ &\leqslant (f \circ X) \sup_{x>b} g(x). \end{aligned}$$

~~If now $X \in \mathbb{R}$ (not necessarily positive) we can write this inequality as~~

$$|X|\mathbb{1}_{(|X|>b)} \leqslant (f \circ |X|) \sup_{x>b} g(x).$$

Take the expectation and the $\sup_K$, using equation $\star$ (remember that $k(b) = \sup_{X\in K}(\mathbb{E}|X|\mathbb{1}_{(|X|>b)})$):

$$k(b) \leqslant c \cdot \sup_{x>b} g(x).$$

~~To study the behaviour in the limit for $b \to \infty$, we must study separately c (that is a finite number) and $\sup_{x>b} g(x)$ (that for $b \to \infty$, and therefore for $x \to \infty$, goes to 0 as $g(x) = \frac{x}{f(x)} = \frac{1}{\frac{f(x)}{x} \to \infty}$).~~ Hence, K is uniformly integrable. □

Let's see one more[39] theorem about uniform integrability.

Theorem 3.2

The following are equivalent:

1. K is uniformly integrable;
2. $h(b) = \sup_K \int_b^{+\infty} \mathrm{d}y\mathbb{P}(|X| > y) \xrightarrow[b\to\infty]{} 0$;
3. $\sup_K \mathbb{E}f \circ |X| < +\infty$ for some increasing convex function f on $\mathbb{R}_+$ such that

$$\lim_{x\to\infty} \frac{f(x)}{x} = +\infty.$$

So we can consider all of these as alternative definition of uniform integrability and this is important because it is the extension of the concept of integrability for random variables. For example, we will treat only uniformly integrable martingales, which are good boys who's mommy good boy you are.

4 σ-algebras and random variables

There is a way to connect σ-algebras and random variables and we will see how linked the two objects are. Consider a random variable X taking values in $(E, \mathcal{E})$. Then consider the collection of all inverse image of the set A through the random variable X:

$$\sigma X = X^{-1}\mathcal{E} = \left\{X^{-1}A : A \in \mathcal{E}\right\}$$

[39] There will be consequences.

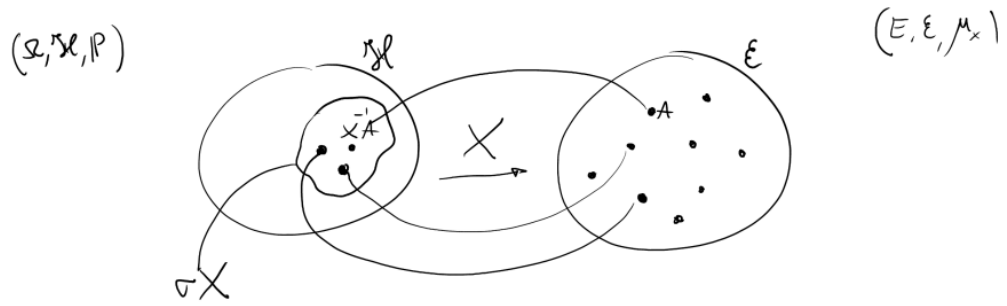


Figure 8: Remember that $\mathcal{H}$ is made of the subsets of Ω and that $\mathcal{E}$ is made of the subsets of E. Also, σX is a subset of $\mathcal{H}$.

Figure 9: I'm just a gurl[40].

We call it σX because it is the **σ-algebra generated by X**. It is basically the set of all inverse images of X that sit in $\mathcal{E}$. This object is simple... Every time we introduce a random variable we have a connection to two spaces: there is the probability space (for instance $(\Omega, \mathcal{H}, \mathbb{P})$), the arrival space $(E, \mathcal{E})$ and the σ-algebra σX. By means of the measurability of X (remember the definition of measurability?) we are simultaneously constructing all the inverse images of the element of the σ-algebra $\mathcal{E}$ in the arrival space. If I consider all of them together we can prove that the resulting set is actually a σ-algebra (the one generated by X) and it is a subset of $\mathcal{H}$, as we see in figure 8. This is very important, for reasons that elude me.

Remark

σX is the smallest σ-algebra $\mathcal{G}$ on Ω such that X is measurable relative to $\mathcal{G}$ and $\mathcal{E}$.

Take now T as an arbitrary index set (which can be countable or uncountable). Consider $\forall\, t \in T$, X_t being a random variable taking values in the measurable space $(E_t, \mathcal{E}_t)$. Then

$$\sigma\left\{X_t : t \in T\right\} = \sigma(X_t)_{t\in T} = \sigma\left(\bigcup_{t\in T} \sigma X_t\right).$$

This is just the union of the σ-algebras generated by the by the random variable X: we consider the σ-algebra of this union just to make sure that this union is, indeed, a σ-algebra.

Note the notation!

Remember that the σ-algebra generated by the collection $(X_t)_{t\in T}$ is noted as

$$\bigvee_{t\in T} \sigma X_t = \sigma\left(\bigcup_{t\in T} \sigma X_t\right)$$

Furthermore, it is the smallest σ-algebra $\mathcal{G}$ on Ω such that $\forall\, t \in T$ the random variable X_t is measurable relative to $\mathcal{G}$ and the corresponding $\mathcal{E}_t$.

Remark

Consider the collection $X = (X_t)_{t\in T}$ to be the random variable taking values in the product space

$$\bigotimes_{t\in T} (E_t, \mathcal{E}_t).$$

Define now $X(\omega)$ to be the point $\Big(X_t(\omega)\Big)_{t\in T}$. The mapping

$$\omega \mapsto X_t(\omega)$$

[40] This image will mean absolutely nothing to you unless you were terminally online in the summer of 2024. Otherwise it is hilarious.

is a random variable and is called **t-coordinate of X**.

Proposition 4.1

If $X = (X_t)_{t\in T}$ then

$$\sigma X = \sigma\Big(X_t\Big)_{t\in T}$$

So if we generate the σ-algebra from the collection as $\bigvee_{t\in T}\sigma X_t$ and we generate the σ-algebra from a collection seen a single random variable (as $\sigma X = X^{-1}\mathcal{E}$) we end up with the same object. Professor Polito think this is a nice thing. I am very keen on that man, to be honest. His enthusiasm is almost contagious.

Remember: when we define a random variable we actually do two things:

- we induce a distribution (that is a probability measure) to the arrival space:

$$(\Omega, \mathcal{H}, \mathbb{P}) \underset{X}{\longrightarrow} (E, \mathcal{E}, \mu_x);$$

- we induce the existence of a special σ-algebra that links the σ-algebra $\mathcal{H}$ and the sa $\mathcal{E}$ (that is the σ-algebra σX).

We can define a new random variable V which is measurable with respect to the σ-algebra σX. Since we defined V as a random variable then it is automatically measurable with respect to $\mathcal{H}$ and $\mathcal{E}$, but in particular when V is also measurable with respect to σX then it has a special representation. Before going through the theorem, let's revise once more a couple of measure theory concepts:

Revise with Kotatsu!

Definition 4.1

The collection of functions $\mathcal{M}$ is called **monotone class** if:

a) $\mathbb{1} \in \mathcal{M}$;

b) let $f, g \in \mathcal{M}_b$ (bounded functions in $\mathcal{M}$) and let $a, b \in \mathbb{R}$. Then $af + bg \in \mathcal{M}$;

c) $(f_n) \subset \mathcal{M}_+$ with $f_n \nearrow f$. Then $f \in \mathcal{M}$.

Theorem 4.1

Monotone class theorem for functions. Let $\mathcal{M}$ be a monotone class of functions on E. Suppose that $\mathbb{1}_A \in \mathcal{M}$ for each $A \in \mathcal{C}$ where $\mathcal{C}$ is some π-system generating $\mathcal{E}$. Then, $\mathcal{M}$ includes all positive $\mathcal{E}$-measurable functions and all bounded $\mathcal{E}$-measurable functions.

We are going to use this theorem in the proof of the following theorem..

Theorem 4.2

"Theorem Δ":
Let X be random variable taking values in the measurable space $(E, \mathcal{E})$. A mapping

$$V : \Omega \mapsto \overline{\mathbb{R}}$$

belongs to σX if and only if

$$V = f \circ X$$

for some deterministic function $f \in \mathcal{E}$.

The proof will be long as fuck.

Proof

(Sufficiency)
Consider a collection $\mathcal{M}$ of all V of the form

$$V = f \circ X, \qquad f \in \mathcal{E}.$$

We would like to prove that $\mathcal{M} \subset \sigma X$, that is to say: if we have a function in the form of V the it is σX-measurable. Measurable functions of measurable functions are still measurable: f is $\mathcal{E}$-measurable and X is a random variable: therefore X is $\mathcal{H}$-measurable but X is also measurable with respect to the σ-algebra generated by itself.
Hence

$$V \text{ is } \sigma X\text{-measurable.}$$

(Necessity)
We are planning to show that $\mathcal{M} \supset \sigma X$.

1. We show that $\mathcal{M}$ is a monotone class of functions: we must check the three properties of the class;
   (a) $\mathbb{1}_E \in \mathcal{M}$ because $\mathbb{1}_E = f \circ X$ with $f(x) = 1 \; \forall x \in E$.
   (b) let $U, V \in \mathcal{M}_b$. Let $a, b \in \mathbb{R}$. Plainly

   $$U = f \circ X, \qquad V = g \circ X$$

   for the right choice of $f, g \in \mathcal{E}$. Consider now the combination

   $$aU + bV = h \circ X$$

   with $h = af + bg$. Note that $h \in \mathcal{E}$. Hence

   $$aU + bV \in \mathcal{M}.$$

   H
   (c) consider $(V_n) \subset \mathcal{M}_+$ and $V_n \nearrow V$. For each n, $\exists\, f_n \in \mathcal{E}$ such that $V_n = f_n \circ X$. We now want to check that $V \in \mathcal{M}$.
   Consider

   $$f = \sup_n f_n \in \mathcal{E}$$

   and

   $$V = \sup_n V_n = \sup_n f_n(X) = f(X)$$

   Hence $V \in \mathcal{M}$.
2. Consider $H \in \Omega$ such that it is in σX. If it is in σX then this means that H comes from the inverse image of a set A in $\mathcal{E}$:

   $$H = X^{-1} A$$

   for some $A \in \mathcal{E}$. So

   $$\mathbb{1}_H = \mathbb{1}_A \circ X \in \mathcal{M}.$$

   So the class $\mathcal{M}$ contains all the indicators of the events that are contained in σX. We need to check, for the monotone class theorem, what happens to the indicators of the elements of the π-system which in this case would be σX... that is a σ-algebra but being a σ-algebra means also being a π-system.
   We have proven that $\mathcal{M}$ contains all the indicators in σX and applying the monotone class theorem for functions we have that $\mathcal{M}$ contains all positive random variables that are σX-measurable.

3. Let $V \in \sigma X$ be arbitrary. Then of course $V^+ \in \sigma X$ and it is positive (the same can be said for v^-). Hence $V^+ = g \circ X$ for some $g \in \mathcal{E}$ and $V^- = h \circ X$ for some $h \in \mathcal{E}$. Then

$$V = V^+ - V_- = f \circ X$$

where

$$f(x) = \begin{cases} g(x) - h(x), & \text{if } g(x) \wedge h(x) = 0 \\ 0, & \text{otherwise} \end{cases}$$

and $f \in \mathcal{E}$.

This concludes $\sigma X \subset \mathcal{M}$. □

Corollary

For each $n \in \mathbb{N}^*$, let X_n be a random variable taking values in $(E_n, \mathcal{E}_n)$. A mapping

$$V : \Omega \mapsto \overline{\mathbb{R}}$$

is $\sigma(X_n)_{n\in\mathbb{N}^*}$-measurable if and only if

$$V = f(X_1, X_2, X_3, \ldots)$$

For some deterministic functtion $f \in \bigotimes_n \mathcal{E}_n$.

Remember that N^* are the natural numbers without 0.

Remark

The above corollary can also be extended to uncountable collections. Should it? I don't know. You tell me.

Why did we destroyed our balls with this section about σ-algebras?

4.1 Filtrations

Let T be a subset of $\mathbb{R}$. Let $\mathcal{F}_t$ be a sub-σ-algebra of $\mathcal{H}$ $\forall\, t \in T$. The family $\mathcal{F} = (\mathcal{F}_t)_{t\in T}$ is called **filtration** if $\mathcal{F}_s \subset \mathcal{F}_t$ for every $s < t$. In figure 10 we can see how $\mathcal{H}$ contains all events we are

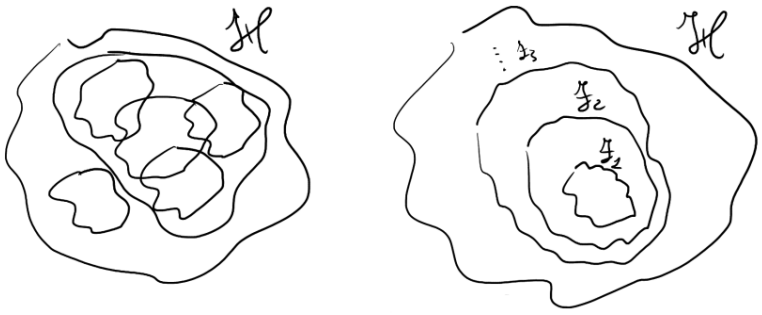


Figure 10: The first one is not a filtration, the second is for $T = \mathbb{N}^*$.

concerned with: it is built upon the *experience* of the event we are facing. Take, for example, $\mathcal{F}_1$: it is much smaller than $\mathcal{H}$ and if we say that a random variable is measurable with respect to $\mathcal{F}_1$ it means that we do not have much knowledge, but if we expand to $\mathcal{F}_2$ we are able to gain more knowledge about the random variable. If we interpret the index set T as time we may actually think about filtrations as our knowledge of the phenomenon as time passes.

Example 4.1

Consider a filtration generated by a stochastic process. Let $X = (X_t)_{t\in T}$. Of course each X_t takes value in the same state space. Define

$$\mathcal{F}_t = \sigma(X_s)_{\substack{s\leqslant t\\ s\leqslant T}}.$$

Here we are generating a σ-algebra from all the random variables from the beginning of time until t. Consider now the family of σ-algebras

$$\mathcal{F} = (\mathcal{F}_t)_{t\in T} :$$

this is a filtration. This means that for this stochastic process, as time passes, we are gaining more knowledge about $\mathcal{H}$ through the σ-algebras.

4.2 Independency for a finite class of sub-σ-algebras

We already know the concept of independence between random variables and the effect of independence with respect to the induced measure of probability. If we look at Çinlar , it actually introduces independence of random variables through the independence of σ-algebras. To do so we need to introduce the concept of **independency for a finite class of sub-σ-algebras**.

Definition 4.2

Consider the class $\mathcal{H}$ as the collection of sub-σ-algebras $\mathcal{F}$:

$$\mathcal{H} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n\} \qquad \mathcal{F}_i = \text{sub-}\sigma\text{-algebra } \forall\, i = 1, \dots, n.$$

$\mathcal{H}$ is said to be an **independency** if

$$\mathbb{E}[V_1 \cdot V_2 \cdot \dots V_n] = \mathbb{E}V_1 \cdot \mathbb{E}V_2 \cdot \dots \mathbb{E}V_n \qquad ((\star))$$

for $\forall$ random variable $V_1, V_2, \dots, V_n$ belonging respectively to $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n$.

But what does this definition mean? The definition is given in the terms of what happens to the random variables that are measurable with respect to the σ-algebras. If for each of them we get the factorization of the expectation, then our sub-σ-algebras are independency... and so, of course, are their algebras (all σ-algebras are algebras.)
This definition is valid only for a *finite* class of sub-σ-algebras, but we can extend it to infinite classes (Çinlar 's book extends it to countable collections of sub-σ-algebras and even to uncountable collections of sub-σ-algebras, so we shouldn't.). This definition, though, was just an intuition of the fact that independency of σ-algebras and independency of random variables are *not* unrelated, but are simply the same concept seen from two different points of view.

Proposition 4.2

Let $\mathcal{F}_1, \mathcal{F}_2 \dots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$, with $n \geqslant 2$. For each $i \leqslant n$, let $\mathcal{C}_i$ be a π-system generating $\mathcal{F}_i$. Then the collection $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n\}$ is an independency if and only if

$$\mathbb{P}(H_1 \cap H_2 \cap \dots \cap H_n) = \prod_{i=1}^{n} \mathbb{P}(H_i) \qquad \forall\, H_i \in \overline{\mathcal{C}}_i = \mathcal{C}_i \cup \{\Omega\},\ i = 1, \dots, n.$$

We need to have clear that there is not an unique starting point in defining the concept of independency, so all of them are quite equivalent. We will build upon this fact and see other deep relations between these notions[41].

Proposition 4.3

Every partition of an independency is an independency.

[41] Can't wait.

Take, for example, the collection $(\mathcal{F}_t)_{t\in T}$ of sub-σ-algebras and the partition $(T_i)_{i\in\mathbb{N}^*}$ of T: these two together form the sub-collection of sub-σ-algebras $\mathcal{F}_{T_i} = (\mathcal{F}_t)_{t\in T}$ with $i \in \mathbb{N}^*$.

Note the notation!

$$\mathcal{F}_{T_i} = \bigvee_{j\in T_i} \mathcal{F}_j.$$

We are interested in checking whether $(\mathcal{F}_{T_1}, \mathcal{F}_{T_2}, \dots)$ (remember: they are all σ-algebras!) are an independency... and they are.

Proposition 4.4

The notion of independence between random variables (taking values in general measurable spaces) is given in terms of independence of the σ-algebras generated by those random variables.

Example 4.2

Consider X_1, X_2 where X_1 takes values in $(E_1, \mathcal{E}_1)$ and X_2 takes values in $(E_2, \mathcal{E}_2)$. If we are concerned about independence of the vector (X_1, X_2) we should look at the vector of the collection of generated σ-algebras $(\sigma X_1, \sigma X_2)$ and check whether it is an independency or not.

Proposition 4.5

Let $X_1, X_2, \dots, X_n$ be random variables taking values respectively in $(E_i, \mathcal{E}_i)$ for $i = 1, \dots, n$. They are independent if and only if

$$\mathbb{E}\left[f_i \circ X_1, f_2 \circ X_n, \dots, f_1 \circ X_n\right] = \mathbb{E}f_i \circ X_1, \mathbb{E}f_2 \circ X_n, \dots, \mathbb{E}f_1 \circ X_n$$

for every $f_1, \dots, f_n$ positive and measurable with respect to $\mathcal{E}_1, \dots, \mathcal{E}_n$ respectively.

We need to prove this proposition. But before, just because we like it (like the little sluts we are), let's state an equivalent proposition.

Proposition 4.6

Let $X_1, X_2, \dots, X_n$ be random variables taking values in $(E_i, \mathcal{E}_i)$ for $i = 1, \dots, n$ respectively. Then they form an independency if and only if their joint distribution equals the product of their marginal distribution.

Well we already knew that... but thanks. Now we see more clearly how these two properties are basically the same thing.

Proof

We will prove the first proposition. We go back to formula of independence $(\star)$ and show that it holds $\forall\, V_1, V_2, \dots, V_n$ positive and measurable with respect to

$$\sigma X_1, \sigma X_2, \dots, \sigma X_n$$

respectively. We know that $(\star)$ holds because of Theorem Δ.

We will now prove the second proposition. From the first proposition we know that:

$$\int_{E_1\times\ldots\times E_n} \underbrace{\pi(\mathrm{d}x_1, \dots, \mathrm{d}x_n)}_{\text{joint distribution of } (X_1,\dots,X_n)} f_1(x_1)\cdot\ldots\cdot f_n(x_n)$$
$$= \int_{E_1} \mu_1(\mathrm{d}x_1) f_1(x) \cdot \int_{E_2} \mu_2(\mathrm{d}x_2) f_2(x) \cdot \ldots \int_{E_n} \mu_n(\mathrm{d}x_n) f_n(x)$$

Plainly, considering positivity of $f_1, f_2, \ldots, f_n$, to have the above equality we obtain that

$$\pi = \mu_1 \times \mu_n.$$

□

You should convince yourself that the second proposition is just the first proposition rewritten. So, the independence of random variables is very important and right now we will see an application.

4.3 Sums of independent random variables

The simplest structure that we can construct with random variableis just adding them up. Let X, Y be independent real-valued random variables with distribution μ, ν respectively. Then the distribution $\mu \star \nu$ of $X + Y$ (i.e. the effect of $\mu \star \nu$ over f) is given by

$$(\mu \star \nu)f = \mathbb{E}f(X+Y) = \int_{\mathbb{R}} \mu(\mathrm{d}x) \int_{\mathbb{R}} \mu(\mathrm{d}y) f(x+y).$$

This is called the **convolution** of the measures μ and ν.
So if we are asked to calculate the distribution it is interesting to choose a specific f to see why the distribution is defined in this way: let us choose $f = \mathbb{1}_{(-\infty,z]}(x+y)$ so that when we calculate the expectation we get

$$\begin{aligned}
\mathbb{E}f(X+Y) = \mathbb{E}\mathbb{1}_{(-\infty,z]}(X+Y) &= \mathbb{P}(X+Y \leqslant z) \\
&= \int_{\mathbb{R}} \mu(\mathrm{d}x) \int_{\mathbb{R}} \nu(\mathrm{d}y) \mathbb{1}_{(-\infty,z]}(x+y) \\
&= \int_{\mathbb{R}} \nu(\mathrm{d}y) \int_{\mathbb{R}} \mathbb{1}_{(-\infty,z-y]}(x) \mu(\mathrm{d}x) \\
&= \int_{\mathbb{R}} \underbrace{F_X(z-y)}_{\text{distribution function of } X} \nu(\mathrm{d}y)
\end{aligned}$$

Which is what, supposedly, we saw in our undergraduate course[42]. From this last series of equations we can also have a look to a special case where X is absolutely continuous with respect to the Lebesgue measure with density $h(x)$:

$$\begin{aligned}
\mathbb{P}(X+Y \leqslant z) = \int_{\mathbb{R}} F_X(z-y)\nu(\mathrm{d}y) &= \int_{\mathbb{R}} \nu(\mathrm{d}y) \int_{-\infty}^{z-y} h(t)\,\mathrm{d}t \\
&\overset{w=t+y}{=} \int_{\mathbb{R}} \nu(\mathrm{d}y) \int_{-\infty}^{z} h(w-y)\,\mathrm{d}w \\
&= \int_{\infty}^{z} \underbrace{\int_{\mathbb{R}} h(w-y)\nu(\mathrm{d}y)}_{\text{density of} X+Y} \mathrm{d}w.
\end{aligned}$$

Further, if Y is also absolutely continuous wuth density g we get that the density of $X+Y$ becomes

$$\int_{\mathbb{R}} h(w-y)g(y)\,\mathrm{d}y$$

Which is the convolution formula for calculating the density of $X+Y$. For a further example see ex. 2.1.3 of Çinlar page 43 (convolution of Gamma densities).

We can further consider a sequence of random variables $(X_n)_{n\in\mathbb{N}^*}$ and start adding them up:

$$S_n = X_1 + X_2 + X_3 + \ldots + X_n$$

so we stop at a finite time. But the we have a new sequence, which is the sequence $(S_n)_{n\in\mathbb{N}^*}$ where S_n is called *partial sum*. The partial sum is in itself a random variable, made by little random variables added up and this is why S_n is also called **random walk**. The question is, as always,

"Should we care?"

[42] I didn't see anything remotely similar to this in my economics degree but what do I know. Economics is not real anyway and economists should kill themselves NOW.

and while the answer is often

"No"

this time it actually is

"Kinda"

which is probably the most astounding result of this whole course. This is one of the first kinds of real problems (albeit very simplistic) modeled through the use of random variables. To do this we need the notion of *tail σ-algebra*.

4.4 Tail σ-algebra

Did you think that we were about to dive into some interesting and useful practical applications of the The name alone suggests a concept of something away from us. Think of an experiment involving $n \geqslant 1$ trials, like the repeated roll of a die or toss of a coin. What matters is that we have a trial repeated infinite times and the result of all these trials (which are random variables) makes up our experiment, which is a random variable in itself.
The experiment is defined on $(\Omega, \mathcal{H}, \mathbb{P})$. Consider a sequence $(\mathcal{G}_n)_{n\in\mathbb{N}^*}$ of sub-σ-algebras of $\mathcal{H}$ such that $\mathcal{G}_n$ is the information on $\mathcal{H}$ revealed by the n-th trial.

Example 4.3

Imagine that the trials are given by the sequence of trials is given by

$$(X_n)_{n\in\mathbb{N}^*}$$

and the sub-σ-algebras are defined by

$$(\sigma X_n)_{n\in\mathbb{N}^*}$$

So the σ-algebras generated by each X_n are actually the *trajectories* of the random variables. Imagine we are sitting a time n. Consider now $\mathcal{G}_m$ for $m > n$: I want to take the union of all these σ-algebras, but since I am not sure that we would end up with a σ-algebra, I take the σ-algebra of that union.

$$\tau_n = \sigma\left(\bigcup_{m>n} \mathcal{G}_m\right) = \bigvee_{m>n} \mathcal{G}_m.$$

This σ-algebra depends on n and it represents the *information about the future*, since we are now sitting at the time point n. Remember that when we talk about *revealed information* we are always talking about σ-algebras! This object is a bit strange... it is about the information that models the future. What the fuck? We now want to take the intersection of the various τ_n (which is surely a σ-algebra since we are intersecting):

$$\tau = \bigcap_n \tau_n.$$

This is now the *information about the remote future*: we are doing the *intersection*, not the *union*! This is not exactly what happens in the future, but the information revealed by all the infinite trials in the future. The σ-algebra τ is called **tail σ-algebra**. Remember that we always remain inside $\mathcal{H}$.

Example 4.4

Consider $(X_n)_{n\in\mathbb{N}^*}$ and $(\mathcal{G}_n)_{n\in\mathbb{N}^*} = (\sigma X_n)_{n\in\mathbb{N}^*}$. We have

$$S_n = \sum_{i=1}^{n} X_i.$$

1. consider the event

$$\left\{\omega : \lim_n S_n(\omega) \text{ exists.}\right\}$$

If we think about the tail σ-algebra, it is the intersection of the σ-algebras after n: it is composed by events whose occurence is not influenced by the happenings in finite

time! If we look at this event it consists of all the realization of the random walk until infinity: so if I were to remove the first n random variables in this sequence then the limit of this sequence wouldn't be affected: so this event *belongs* to τ.

2. consider the set $B \in \mathcal{B}_{\mathbb{R}}$ and consider the event

$$\{\omega : X_n(\omega) \in B \text{ i.o.}\}.$$

"*i.o.*" means *infinitely often*: this means that

$$\{\omega : X_n(\omega) \in B \text{ i.o.}\} = \left\{\omega : \sum_n \mathbb{1}_B \circ X_n(\omega) = +\infty\right\}.$$

So we infinitely many $\mathbb{1}$'s in B. Again, it is clear that this event belongs in the tail σ-algebra because to test the divergence of the series $\sum_n \mathbb{1}_B \circ X_n(\omega)$ we need to consider the behavior in the limit. We don't care what happens in finite time, so if we remove the contribution of the first n random variables:

$$\{\omega : X_n(\omega) \in B \text{ i.o.}\} \in \tau.$$

3. consider again $B \in \mathcal{B}_{\mathbb{R}}$ and consider

$$\{\omega : S_n(\omega) \in B \text{ i.o.}\}.$$

This event does not belong in τ! Why? Let's consider a special case of $B = \{0\}$ and therefore we consider the event

$$\{\omega : S_n(\omega) = 0 \text{ i.o.}\}$$

and we also specialize the "jumps":

$$(X_n)_{n\in\mathbb{N}^*} \text{ is i.i.d. with } \begin{array}{l}\mathbb{P}(X_1 = 1) = p \\ \mathbb{P}(X_1 = -1) = 1 - p\end{array}$$

with $p \in [0, 1]$. In this case $\{\omega : S_n(\omega) = 0 \text{ i.o.}\}$ is not a tail event. Let's realize the sequence of the "jumps" choosing an ω such that

$$X(\omega) = (+1, -1, +1, -1, +1, \ldots)$$

Clearly $\omega \in \{\omega : S_n(\omega) = 0 \text{ i.o.}\}$:

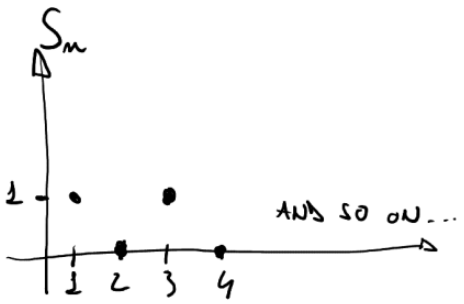


Figure 11: The random walk randomly walking

Now we choose a different $\tilde{\omega}$ such that:

$$X(\tilde{\omega}) = (+1, +1, +1, +1, -1, +1, -1, \ldots)$$

So in this case we have a "ramp" before the oscillation: $\tilde{\omega} \notin \{\omega : S_n(\omega) = 0 \text{ i.o.}\}$! Try to make the same drawing: the walk will go up and the 0 will never be visited!
Observe that ω and $\tilde{\omega}$ agree from the fourth coordinate on: but this means that the

first events are decisive when it comes to determine whether a certain ω belongs to $\{\omega : S_n(\omega) = 0 \text{ i.o.}\}$. This means that

$$\{\omega : S_n(\omega) = 0 \text{ i.o.}\} \notin \underbrace{\sigma(X_4, X_5, X_6, \ldots)}_{\text{We need more information to decide!}} \implies \notin \tau_3$$

So of course $\{\omega : S_n(\omega) = 0 \text{ i.o.}\} \notin \tau$. The nature of the event is changed by a finite number of occurences...

4. let $(X_1, X_2, \ldots)$ be independent and $S_n = \sum_{i=1}^n X_i$. Then

$$\left\{\omega : \frac{S_n(\omega)}{n} \text{ converges (exists finite)}\right\} \in \tau.$$

Consider

$$Z_1 = \liminf_{n\to+\infty} \frac{S_n}{n}, \; Z_2 = \limsup_{n\to+\infty} \frac{S_n}{n}$$

so that our event becomes

$$\{\omega : Z_1(\omega) = Z_2(\omega)\}.$$

Rewrite

$$\frac{S_n}{n} = \underbrace{\frac{1}{n}\sum_{i=1}^{m} X_i}_{S_{n1}} + \underbrace{\frac{1}{n}\sum_{i=m+1}^{n} X_i}_{S_{n2}} \qquad m \leqslant n$$

So that we "split" the partial sums in two sums before and after time m. We call these two partial sums S_{n1} and S_{n2}. Note that when $n \to \infty$ then $S_{n1} \to 0$ and therefore Z_1 and Z_2 do not depend upon the first n values of $(X_1, X_2, \ldots, X_m)$. Therefore

$$\{Z_1 = Z_2\} \in \tau.$$

Theorem 4.3

Kolmogorov's 0-1 law

This is a theorem about independence. Let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ be independent. Then

$$\mathbb{P}(H) = \begin{cases} 0 \\ 1 \end{cases} \qquad \forall\, H \in \tau.$$

So when the outcomes of a random variable are independent then the tail is made of almost sure or almost impossible events! The proof is short.

Proof

Remember that partition of independencies are independencies. Our independency is formed by the sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ and our partition is

$$(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n, \tau_n).$$

This is an independency $\forall\, n \in \mathbb{N}^*$. Since $\tau \subset \tau_n \; \forall n$, then $(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n, \tau)$ is an independency for each n. But since we now have an independency made of finite elements for each n we have an independency of infinite components

$$(\tau, \mathcal{G}_1, \mathcal{G}_2, \ldots)$$

by definition of independency for countably infinite sequences. But again, if we partition that last independency we still get an independency:

$$(\tau, \tau_0)$$

is still an independency. Now consider two σ-algebras $H \in \tau, G \in \tau_0$. Since they are independent, we have that $\mathbb{P}(H \cap G) = \mathbb{P}(H)\mathbb{P}(G)$. But since $\tau \subset \tau_0$ then we can choose

$G = H$ and we get

$$\mathbb{P}(H) = \mathbb{P}(H \cap H) = \mathbb{P}(H)\mathbb{P}(H)$$

and the solution to this equation can only be 1 or 0. □

We must remember that this is a special case where we have independence: this, in an experimental scenario, is modeled by independent trials. This result can be applied of concept of convergence. The idea is to build up instruments useful to the analysis of real random variable in their limits

5 Convergence of random variables and asymptotic behavior

5.1 Convergence of real sequences

Let $(x_n)_n$ be a sequence in $\mathbb{R}, n \in \mathbb{N}^*$. Then

$$\liminf_n x_n = \sup_m \inf_{n \geqslant m} x_n$$

and the

$$\limsup_n x_n = \inf_m \sup_{n \geqslant m} x_n.$$

Both are well defined numbers (and can possibly be infinity). We know that if

$$\liminf_n x_n = \limsup_n x_n$$

then the limit

$$\lim_n x_n$$

exists and the sequence is **convergent**.

Note the notation!

Let $\varepsilon \in \mathbb{R}_+$. We let i_ε be the indicator function of (ε, ∞):

$$i_\varepsilon(x) = \mathbb{1}_{(\varepsilon,\infty)}(x) = \begin{cases} 1, & x > \varepsilon \\ 0, & x \leqslant \varepsilon. \end{cases}$$

We need two remarks about the convergence of x_n:

Remark

- Let $(x_n)_n$ be a sequence in $\mathbb{R}$. Then it converges to $x \in \mathbb{R}$ if and only if

$$|x_n - x| \underset{n}{\to} 0.$$

- Let $(x_n)_n$ be a sequence in $\mathbb{R}$. Then it converges to 0 if and only if

$$\forall \varepsilon > 0 \exists k \text{ s.t. } x_n \leqslant \varepsilon \; \forall n \geqslant k.$$

This is the usual definition of the limit and it means that

$$x_n \underset{n}{\to} 0 \iff \sum_n i_\varepsilon < +\infty \quad \forall \varepsilon > 0.$$

This means that we have a finite number of x_n that are smaller than ε, so summing up the indicator function when we are in an interval larger than ε we will get less than infinity. This concept connects the convergence of real-valued positive sequences with the convergence to 0 of real-valued positive sequences.
Further, the fact that

$$\sum_n i_\varepsilon(x_n) < +\infty \iff \limsup_n i_\varepsilon(x_n) = 0 \iff \lim_n i_\varepsilon(x_n) = 0$$

Proposition 5.1

Cauchy's criterion:
Let $(x_n)_n$ be a sequence in $\mathbb{R}$. Then $(x_n)_n$ converges if and only if

$$\lim_{m,n\to\infty} |x_n - x_m| = 0.$$

This criterion is interesting because it gives us a criterion to check whether the sequence converges without looking at the actual limit of the sequence...

Proposition 5.2

If there exists a positive sequence $(\varepsilon_n)_n$ such that

$$\sum_n \varepsilon_n < +\infty, \qquad \sum_n i_{\varepsilon_n}(|x_{n+1} - x_n|) < +\infty$$

Then $(x_n)_n$ is convergent.

This is another way to prove convergence: we just need to prove that the two series described above are finite.

Definition 5.1

Given a sequence $(x_n)_n$, the sequence $(y_n)_n$ is said to be a **subsequence of** $(x_n)_n$ if there exists an increasing sequence $(k_n)_n \in \mathbb{N}^*$ with $\lim_n k_n = \infty$ such that

$$y_n = x_{k_n} \qquad \forall\, n.$$

So the subsequence is like a tool to extract elements from a sequence that are still a sequence!

Note the notation!

We will usually denote $(y_n)_n$ as $(x_n)_{n\in N}$ where $N = (k_n)_n$.

Remark

We say that $(x_n)_n$ converges **along** N to x if $(x_n)_{n\in N}$ converges to x. This basically means that we are extracting a subsequence from x and check if it converges.

Remark

$(x_n)_n$ converges to x if and only if every subsequence has the same limit x. The converse is also true!

Remark

Let $(x_n)_n$ be a bounded real sequence. Then it is always possible to extract from it a convergent subsequence.

So if we prove that $(x_n)_n$ is bounded we can always get a convergent subsequence...

Proposition 5.3

Selection principle:
If every subsequence that has a limit has the same limit value x, then the sequence that we are considering tends to the same limit x (that can be finite or infinite). If the sequence is bounded and every convergent subsequence has the same limit x, then the sequence converges to x.

We now consider a lemma about the behavior of lim sup and lim inf.

Lemma 5.1

Let $(x_n)_n$ be a sequence of positive real numbers and let $\overline{x}_n = \frac{\sum_{i=1}^n x_i}{n}$. Let $N = (n_k)_k$ be a subsequence of $\mathbb{N}$ with

$$\lim_k \frac{n_{k+1}}{n_k} = r > 0.$$

If $(\overline{x}_n)_n$ converges along N to x then

$$\frac{x}{r} \leqslant \liminf_n \overline{x}_n \leqslant r \cdot x.$$

This is interesting because it gives us an idea of the fluctuation of the sequence of averages in terms of the behaviour of the subsequence N and the limit x. We will make use of this lemma soon...

Theorem 5.1

Helly's theorem:
For every sequence $(c_n)_n$ of distribution functions (this is not a real-valued sequence anymore) there exists a subsequence of distribution function $(b_n)_n$ and a limiting distribution function x such that

$$\lim_{n\to\infty} b_n(t) = c(t) \qquad \forall\, t \text{ at which } c \text{ is continuous.}$$

Lemma 5.2

Kronecker's lemma:
Let $(x_n)_n$ a real-valued sequence. let $(a_n)_n$ be a strictly positive sequence increasing to ∞. Finally, write

$$y_n = \frac{\sum_{k=1}^n x_k}{a_k}.$$

Then if (y_n) is convergent,

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{k=1}^n x_k = 0.$$

We are now ready to talk about what we actually care about in probability theory.

5.2 Almost sure convergence

Definition 5.2

A real-valued sequence of random variables $(X_n)_n$ on $(\Omega, \mathcal{H}, \mathbb{P})$ is said to be **almost sure convergent** (a.s. convergent) is the numerical sequence

$$\Big(X_n(\omega)\Big)_n$$

converges for almost all $\omega \in \Omega$.
It is said to converge to X if X is an almost sure real-valued random variable and

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

for almost all $\omega \in \Omega$.

Let us define

$$\Omega_0 = \left\{\omega \in \Omega : \liminf_n X_n(\omega) = \limsup_n X_n(\omega)\right\}.$$

Of course this is an event and we note that $(X_n)_n$ is almost sure convergent if and only if Ω_0 is an almost sure event (i.e. $\mathbb{P}(\Omega_0) = 1$).

Theorem 5.2

Characterization of almost sure convergence:
A sequence of real valued random variables $(X_n)_n$ converges to X almost surely if and only if, for every $\varepsilon > 0$,

$$\sum_n i_\varepsilon \circ |X_n - X| < \infty \qquad (\star\star)$$

almost surely.

Proof

Implication ($\Longrightarrow$).
Start from the fact that

$$X_n \to X \text{ a.s.}$$

Write $Y_n = |X_n - X|$: for every $\omega \in \Omega$, by the analogous result for real sequences, we have that

$$\sum_n i_\varepsilon \circ Y_n(\omega) < \infty \qquad \forall\, \varepsilon > 0$$

and therefore $(\star\star)$ holds.
Then there is the converse implication ($\Longleftarrow$).
Let $(\star\star)$ hold and let us consider $(\varepsilon_k)_k$ be a strictly decreasing sequence converging to 0.
Also, let

$$N_k = \sum_n i_{\varepsilon_k} \circ |X_n - X|.$$

Then

$$\mathbb{P}(N_k < \infty) = 1 \qquad \forall\, k$$

and this is true because $(\star\star)$ holds! Now note that

$$\varepsilon_{k+1} < \varepsilon_k \implies i_{\varepsilon_{k+1}} \geqslant i_{\varepsilon_k} \implies N_{k+1} \geqslant N_k.$$

But now we have to pay attention to the nature of the events involved: consider the event

$$\{N_k < +\infty\} \qquad \forall\, k.$$

The sequence $(\{N_k < +\infty\})_k$ decreases with k towards the intersection

$$\bigcap_k \{N_k < +\infty\} = \left\{ \omega \in \Omega : \sum_n i_\varepsilon \circ Y_n(\omega) \quad \forall\, \varepsilon > 0 \right\}$$
$$= \Omega_0.$$

Hence we want to evaluate

$$\mathbb{P}(\Omega_0) = \mathbb{P}(\lim_k \{N_k < +\infty\}) = \underbrace{\lim_k \mathbb{P}(N_k < +\infty)}_{\text{sequence of 1}}$$

Where the last equality is true because of sequential continuity. So $\forall \omega \in \Omega_0$ we have $X_n(\omega) \to X(\omega)$ which is the definition of almost surely convergence. □

Let's study more what happens in the limits of these random walks. Remember that Kolmogorov's 0-1 law tells us something nice: if all the random variables are independent then events in the asymptotic future will either be certain or impossible. But we know nothing about situations in which random variables are dependent...

5.3 Borel-Cantelli Lemmas

Let $(H_n)_n$ be a sequence of events. Imagine evaluating the probability of all these events: we would end up with a sequence of real numbers. We may ask ourselves how the sum of the probability

behaves.

Theorem 5.3

First Borel-Cantelli Lemma Let $(H_n)_n$ be a sequence of events. Then

$$\sum_n \mathbb{P}(H_n) < +\infty \implies \sum_n \mathbb{1}_{H_n} < +\infty \quad \text{a.s.}$$

Remember that the second sum is equivalent to

$$\mathbb{P}(H_n \text{ i.o.}) = 0 \qquad \text{or} \qquad \mathbb{P}(\limsup_n H_n) = 0.$$

If you want to know more see Çinlar page 101. Why do we need it to be almost surely? Well, the sum $\sum_n \mathbb{1}_{H_n}$ contains an indicator of events (which is nothing else but a Bernoulli random variable) so it is in itself a random variable. So we are asking yourself whether the sum converges *almost surely*. This theorem is important because if we can sum the probabilities then we can directly infer information about the presence of events! Let's see the proof[43].

Proof

Denote

$$N = \sum_n \mathbb{1}_{H_n}$$

So that

$$\sum_n \mathbb{P}(H_n) = \mathbb{E}N.$$

Our claim now becomes:

"If $\mathbb{E}N < +\infty$ then $N < \infty$ almost surely."

But this is obvious: if the expectation is finite then the random variable is finite almost surely (mening that the support is finite). □

That's it? That was cool and all. But let's see the first Borel-Cantelli lemma in action in the next proposition.

Proposition 5.4

Let

$$\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < +\infty \qquad \forall \varepsilon > 0.$$

Then

$$X_n \xrightarrow{\text{a.s.}} X.$$

Here we are checking the convergence of the sum of a sequence of real numbers and we are given a result on random variables. This is pretty neat!

Proof

Consider

$$H_n = \{|X_n - X| > \varepsilon\}.$$

Then, we apply Borel-Cantelli Lemma and we calculate

$$\sum_n \mathbb{1}_{\{|X_n - X| > \varepsilon\}} < +\infty.$$

But now we are in the situation described by ($\star\star$), so we know that

$$X_n \xrightarrow{\text{a.s.}} X$$

□

[43] Professor Polito says that we should as interested in the result as in the proof. I do not agree and wish to die.

So we can proof almost sure convergence by means of the application. So when we need to prove convergence always think about Borel-Cantelli lemmas!

Proposition 5.5

Suppose that there exists a sequence $(\varepsilon_n)_n$ decreasing to 0 such that

$$\sum_n \mathbb{P}(|X_n - X| > \varepsilon_n) < +\infty.$$

Then

$$X_n \xrightarrow{\text{a.s.}} X.$$

So we don't need ε to be constant, but just to be decreasing to 0.

Proposition 5.6

Suppose that there exists a sequence of positive numbers $(\varepsilon_n)_n$ such that

$$\sum_n \varepsilon_n < +\infty, \quad \sum_n \mathbb{P}(|X_{n+1} - X_n| > \varepsilon_n) < +\infty$$

Then X_n converges almost surely.

Let's state again the two Borel-Cantelli lemmas side by side (and with a slightly different formulation).

Theorem 5.4

Borel-Cantelli lemmas

a) Let $(B_n)_n$ be a sequence of Bernoulli random variables.

$$\sum_n \mathbb{E}B_n < +\infty \implies \sum_n \implies \sum_n B_n < +\infty \text{ a.s.}$$

b) If

$$\sum_n \mathbb{E}B_n = \infty \text{ and } (B_n)_n \text{ are pairwise independent}$$

then

$$\sum_n B_n = +\infty \text{ a.s.}$$

We will prove b).

Proof

Note the notation!

$$p_n = \mathbb{E}B_n, \quad a_n = \sum_{i=1}^n \mathbb{E}B_i = \sum_{i=1}^n p_i$$

and we get the partial sum and the limits

$$S_n = \sum_{i=1}^n B_i, \quad S = \lim_n S_n$$

First, we know that, due to pairwise independency in the hypothesis

$$\operatorname{Var} S_n = \sum_{i=1}^n \operatorname{Var} B_i$$

but we know that they are Bernoulli random variables, so we can write the variance explicitly:

$$\mathbb{V}\mathrm{ar}\, S_n = \sum_{i=1}^{n} p_i(1-p_i) \leqslant \sum_{i=1}^{n} p_i = a_n$$

where the inequality is just arithmetical calculations.
Second, fix $b \in (0, +\infty)$. We know that $(a_n)_n$ increases to ∞ by hypothesis:

$$\left(a_n - \sqrt{ba_n}\right)_n.$$

We are basically subtracting to a_n something that goes to ∞ as well, but slower than a_n. So the sequence $\left(a_n - \sqrt{ba_n}\right)_n$ stays also increasing towards infinity. But if that is the case the event $\{S < +\infty\}$ is the limit of the increasing sequence of events

$$\left\{S < a_n - \sqrt{ba_n}\right\}.$$

Since $S_n \leqslant S \; \forall\, n$ consider further the following modified sequence of events:

$$\left\{S_n < a_n - \sqrt{ba_n}\right\} \supset \left\{S < a_n - \sqrt{ba_n}\right\}.$$

Next, we also have

$$\left\{S_n < a_n - \sqrt{ba_n}\right\} \subset \left\{|S_n - a_n| > \sqrt{ba_n}\right\}.$$

Now, let's switch to the probability point of view. Remember that the events are included one in the other so we can think about weak inequalities of probabilities.

$$\mathbb{P}(S < a_n - \sqrt{ba_n}) \leqslant \mathbb{P}(S_n < a_n - \sqrt{ba_n}) \leqslant \mathbb{P}(|S_n - a_n| > \sqrt{ba_n})$$

and we take the $\limsup_n$:

$$\limsup_n \mathbb{P}(S < a_n - \sqrt{ba_n}) \leqslant \limsup_n \mathbb{P}(|S_n - a_n| > \sqrt{ba_n})$$

But the left hand side of the inequality is

$$\begin{aligned}\limsup_n \mathbb{P}(S < a_n - \sqrt{ba_n}) &= \mathbb{P}\left(\lim_n \left\{\mathbb{P}(S < a_n - \sqrt{ba_n})\right\}\right) \\ &= \mathbb{P}(S < \infty).\end{aligned}$$

We get, thus,

$$\mathbb{P}(S < +\infty) \leqslant \limsup_n \mathbb{P}(|S_n - a_n| > \sqrt{ba_n}).$$

But now we can apply Chebyshev's inequality:

$$\mathbb{P}(S < +\infty) \leqslant \limsup_n \frac{\mathbb{V}\mathrm{ar}\, S_n}{ba_n}$$

and we can now exploit the result we had reached before:

$$\mathbb{P}(S < +\infty) \leqslant \limsup_n \frac{\mathbb{V}\mathrm{ar}\, S_n}{ba_n} \leqslant \limsup_n \frac{\cancel{a_n}}{b\cancel{a_n}} = \frac{1}{b}.$$

If we let $b \to \infty$ we should get the same result and immediately obtain

$$\mathbb{P}(S < +\infty) \leqslant \frac{1}{b} \xrightarrow[b\to\infty]{} 0$$

and hence

$$\mathbb{P}(S = +\infty) = 1.$$

□

Damn. I hate this. Let's see an example about coin toss!

Example 5.1

Imagine considering the probability of getting a head[a]:

$$\mathbb{P}(H) \in (0,1), \qquad \Omega = \{H,T\}^{\infty}.$$

We are infinitely tossing coin and the probability of having head is always the same. Our probability space is thus the space of infinity sequences of H and T. Consider a chunk of tosses and define B_1 as the first subsequence of length k, B_2 as the second subsequence of length k and so on. So we get infinitely many chunks of length k.
Fix

$$s = (s_1, s_2, \dots, s_k) \in \{H,T\}^k.$$

Now consider the events

$$\begin{aligned} D_1 &= \{B_1 = s\} \\ D_2 &= \{B_2 = s\} \\ D_3 &= \{B_3 = s\} \\ &\vdots \end{aligned}$$

and suppose further that the sequence s has actually m heads and $k-m$ tails. Then

$$\mathbb{P}(D_i) = [\mathbb{P}(H)]^m \cdot [\mathbb{P}(T)]^{k-m}.$$

Imagine we want to calculate the sum:

$$\sum_i \mathbb{P}(D_i) = +\infty.$$

But now we can use the Borel-Cantelli lemma (point b))! So we get that

$$\sum_i \mathbb{1}_{D_i} = \infty$$

which means that

$$\mathbb{P}(D_1 \text{ i.o.}) = 1$$

or, in other words, for any given sequence of k outcomes we observe it infinitely often.

[a] ...so no head? Imma head out.

5.4 Convergence in probability

This definition is strictly related to almost sure convergence. In measure theory this is called *convergence in measure* (makes sense...).

Definition 5.3

Let $(X_n)_n$ be a sequence of real-valued random variables. Then $(X_n)_n$ converges to a further real-valued random variable **in probability** if

$$\lim_n \mathbb{P}(|X_n - X| > \varepsilon) = 0 \qquad \forall \varepsilon > 0.$$

So this means that the probability of X_n being at most ε away from X gets smaller and smaller as n gets bigger. It is quite different than the notion of almost sure convergence... and it depends heavily on the probability measure $\mathbb{P}$! What we get now is a sequence of probabilities that we need to evaluate.

Example 5.2

Consider

$$\Omega = [0,1],\ \mathcal{H} = \mathcal{B}_{[0,1]},\ \mathbb{P} = \underbrace{\lambda}_{\text{Lebesgue measure}}.$$

Consider the sequence of random variables

$$X_1, X_2, X_3, \dots$$

to be indicators of $(0,1], (0,\frac{1}{2}], (\frac{1}{2},1], (0,\frac{1}{3}], (\frac{1}{3},\frac{2}{3}], (\frac{2}{3},1], \dots$ respectively.
Then $\forall\, \varepsilon > 0, \varepsilon \in (0,1)$, $\mathbb{P}(X_n > \varepsilon)$ forms the sequence

$$(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \dots) \xrightarrow[n\to\infty]{} 0.$$

Hence

$$X_n \xrightarrow{\mathbb{P}} 0.$$

However,

$$X_n(\omega) = \begin{cases} 0 \\ 1 \end{cases} \qquad \forall\, \omega \in \Omega$$

and therefore we have that

$$\liminf_n X_n = 0,\ \limsup_n X_n = 1$$

and we do not have almost sure convergence... So

$$\Omega_0 = \left\{ \omega \in \Omega : \liminf_n X_n(\omega) = \limsup_n X_n(\omega) \right\}$$

is empty!

But how are these two different types of convergence related?

Theorem 5.5

Characterization theorem for convergence in probability.

i) if $(X_n)_n$ converges to X almost surely, then it converges to X in probability;

ii) if $(X_n)_n$ converges in probability to X, then it has a subsequence converging to the same random variable X almost surely;

iii) if every subsequence of the main sequence has a further subsequence converging to X almost surely, then the main sequence converges to X in probability.

We will only prove point *i)* because it is the most useful.

Proof

Let $X_n > 0\ \forall n$ and $X = 0$: this means replacing $|X_n - X|$ with X_n (as long as X_n is positive and converges to 0). Recall the definition of indicator $i_\varepsilon + \mathbb{1}_{(\varepsilon,\infty)}$ and define

$$p_n = p_n(\varepsilon) = \mathbb{E}[\underbrace{i_\varepsilon \circ X_n}_{\text{Bernoulli r.v.}}] = \mathbb{P}(X_n > \varepsilon).$$

Here we switch between expectation and probability because of the property of expectations with indicator functions in this theorem. We know that $X_n \xrightarrow{\text{a.s.}} 0$. Fix $\varepsilon > 0$: now we have

$$\mathbb{1}_\varepsilon \circ X_n \xrightarrow{\text{a.s.}} 0.$$

Now take the expectation

$$p_n = \mathbb{E} i_\varepsilon \xrightarrow{\text{a.s.}} 0$$

Hence

$$\mathbb{P}(X_n > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

that is to say

$$X_n \xrightarrow{\mathbb{P}} 0$$

□

Proposition 5.7

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be continuous. Then if $X_n \xrightarrow{\mathbb{P}} X$ in probability then

$$f(X_n) \xrightarrow{\mathbb{P}} f(x)$$

That is to say $f \circ N_n \xrightarrow{\mathbb{P}} f \circ X$.

Proof

We know that $X_n \xrightarrow{\mathbb{P}} X$. Let N be a subsequence of $\mathbb{N}^*$. Then $X_n \xrightarrow{\mathbb{P}} X$ along N. For the point *ii)* of the previous theorem there exists a subsequence N' along which the convergence takes place almost surely:

$$X_n \xrightarrow{\text{a.s.}} X.$$

Then along N'

$$f \circ X_n \xrightarrow{\text{a.s.}} f \circ X$$

for the continuity of f: looking at the limit of f we exploit the continuity of f:

$$\lim f\Big(X_n(\omega)\Big) = f\Big(\lim X_n(\omega)\Big).$$

So saying that f is continuous is enough to prove the convergence of its composition with X. Now by point *iii)* of the previous theorem we conclude that

$$f \circ X_n \xrightarrow{\mathbb{P}} F \circ N.$$

□

Remark

Convergence in probability is preserved under arithmetic operations.

So, for example if we have $X \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$ then $(X_n + Y_n)_n \xrightarrow{\mathbb{P}} X + Y$.
We will now introduce a metric for convergence in probability: since we are talking about *distance*, we may think[44] that this has something to do with *metric spaces*, where we measure the distance between different objects. Indeed there is a connection between measure spaces and metric spaces...

Let us introduce now a metric for convergence in probability. If we want to calculate a metric between random variables X and Y we can define the following metric:

$$d(X, Y) = \mathbb{E}(|X - Y| \wedge 1).$$

Remark

1. $d(X, Y) = 0 \iff X = Y$ a.s.
2. $d(X, Y) + d(Y, Z) \geqslant d(X, Z)$.

d is a metric on the space of real-valued random variables if X and Y are indentified as the same random variable if $X = Y$ almost surely.

[44] A remark that reeks of overestimation.

Proposition 5.8

$$X_n \xrightarrow{\mathbb{P}} X \iff d(X_n, X) \xrightarrow[n\to\infty]{} 0.$$

Instead of proving this version, we will prove a slightly different version of the proposition, that is:

Theorem 5.6

$$X_n \xrightarrow{\mathbb{P}} X \iff \lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n - X|}{1 + |X_n - X|}\right] = 0$$

Proof

Without losing generality we fix $X = 0$ a.s. We are concerned about the following claim:

$$X_n \xrightarrow{\mathbb{P}} 0 \iff \lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right].$$

($\implies$) Suppose that $X_n \xrightarrow{\mathbb{P}} 0$. This means that

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) = 0, \qquad \forall \varepsilon > 0.$$

Note that

$$\frac{|X_n|}{1 + |X_n|} \leqslant \frac{|X_n|}{1 + |X_n|} \mathbb{1}_{(|X_n| > \varepsilon)} + \varepsilon \mathbb{1}_{(|X_n| < \varepsilon)}$$

Because if $|X_n|$ is smaller than ε then also the fraction above is smaller than ε (duh). But we see that

$$\frac{|X_n|}{1 + |X_n|} \leqslant \underbrace{\frac{|X_n|}{1 + |X_n|}}_{<1} \mathbb{1}_{(|X_n| > \varepsilon)} + \varepsilon \underbrace{\mathbb{1}_{(|X_n| < \varepsilon)}}_{\leqslant 1}$$
$$\leqslant \mathbb{1}_{(|X_n > \varepsilon|)} + \varepsilon$$

Then take the expectation:

$$\mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] \leqslant \underbrace{\mathbb{E}\mathbb{1}_{(|X_n > \varepsilon|)}}_{\mathbb{P}(|X_n| > \varepsilon)} + \varepsilon =$$

and the limit

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] \leqslant \varepsilon$$

Because we know that $\mathbb{P}(|X_n| > \varepsilon)$ goes to 0 as n gets larger. But since ε is chosen arbitrarily, then

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] = 0.$$

($\impliedby$) Let

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] = 0.$$

We want to prove convergence in probability. Note that $f(x) = \frac{x}{1+x}$ is strictly increasing and therefore I can write the following inequality:

$$\frac{\varepsilon}{1 + \varepsilon} \mathbb{1}_{(|X_n > \varepsilon|)} \leqslant \frac{|X_n|}{1 + |X_n|} \leqslant \mathbb{1}_{(|X_n > \varepsilon|)} \leqslant \frac{|X_n|}{1 + |X_n|}.$$

Now take the expectation and the limit for $n \to \infty$:

$$[\frac{\varepsilon}{1 + \varepsilon} \lim_{n\to\infty} \mathbb{P}(|X_n| > \varepsilon) \leqslant \lim_{n\to\infty} \mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] = 0$$

where the last "=0" is by hypothesis. Since ε is chosen arbitrariliy then

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

so we have proven the converse implication. □

We said before that X_n converges in probability to X if the metric $d(X_n, X) \xrightarrow[n\to\infty]{} 0$. The last proof refers to the function $f(x) = \frac{|x|}{1+|x|}$ but the same holds for every bounded, non decreasing, continuous function g on $\mathbb{R}_+$ with $g(0) = 0, g(x) > 0 \,\forall\, x > 0$. For example, $g(x) = |x| \wedge 1$ (which is similar to the metric we estabilished before).

5.5 Convergence in L^p spaces

Let's start with the proper definition.

Definition 5.4

A sequence $(X_n)_n$ of real-valued random variables is said to be convergent to the real-valued random variable X in L^p if, for each n, $X_n \in L^p, X \in L^p$ and

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

Revise with Kotatsu!

Recall the p-norm of X:

$$\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

Then L^p is a normed vector space if we identify all the random variables that are almost surely equal as one single random variable: this means that the norm works up to an equivalence class (if A and B are equal a.s. then for *us* they are the same random variable as long as we work in L^p). In this case, the L^p-convergence is to be written as

$$\lim_{n\to\infty} \|X_n - X\|_p = 0.$$

Remark

What happens if we have two different sequences converging to the same limit in L^p? If $(X_n)_n$ converges to X in L^p then X is unique (up to equivalence).

Proof

Consider

$$X_n \xrightarrow{L^p} X$$
$$X_n \xrightarrow{L^p} Y$$

Revise with Kotatsu!

Recall the Minkowsky's inequality

$$\|X - Y\|_p \leqslant \|X - X_n\|_p + \|X_n - Y\|_p \xrightarrow[n\to 0]{} 0$$

But both $\|X - X_n\|_p$ and $\|X_n - Y\|_p$ tend to 0 as $n \to \infty$, we have that

$$X = Y \text{ a.s.}$$

□

So, since they are the same random variable in L^p then they differ only for elements of measure 0.

Remark

$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X.$$

This is quite strong! Convergence in L^p is stronger! I wish the same could be said of my will to live.

Proof

We use Markov's inequality:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leqslant \left(\frac{1}{\varepsilon}\right)^p \mathbb{E}|X_n - X|^p \xrightarrow[n\to 0]{} 0$$

□

But is the converse true? NO IT IS NOT IT IS NOT TRUE IT IS A LIE. Consider this counterexample:

Example 5.3

We show that the converse of what it is stated in the previous remark does NOT hold in general. Consider $\Omega = [0,1], \mathcal{H} = \mathcal{B}_{[0,1]}, \mathbb{P} = \lambda$. Consider $(X_{n_n}$ a sequence of indicators of the intervals

$$(0,1], \left(0, \frac{1}{2}\right], \left(\frac{1}{2}, 1\right], \left(0, \frac{1}{3}\right], \left(\frac{1}{3}, \frac{2}{3}\right], \left(\frac{2}{3}, 1\right], \dots$$

Note that we already proved that this sequences converges to 0 in probability. Note, moreover, that $X_n \xrightarrow{L^p} 0$ because $\mathbb{E}|X_n|$ forms the sequence

$$(1 + 0 \cdot p, 1 \cdot p + 0 \cdot q, \dots) = \left(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \dots\right)$$

(remember that our $\mathbb{P}$ is the Lebesgue measure, which is the length of the interval(). This sequence goes to zero, clearly.
Now let's modify the sequence of random variables:

$$(\hat{X}_n)_n = (X_1, 2X_2, 2X_3, 3X_4, 3X_5, 3X_6, \dots)$$

and evaluate

$$\mathbb{P}(\hat{X}_n > \varepsilon) = \mathbb{E}|X_n|$$

because for indicators our probability coincides with the expectation. Now take the limit and see whether $(\hat{X}_n) \xrightarrow{\mathbb{P}} 0$. However, if we do the calculations we get

$$\mathbb{E}|\hat{X}_n| = 1 \qquad \forall n.$$

Hence

$$\hat{X}_n \not\xrightarrow{L^1} 0$$

so there exists a p for which $\hat{X}_n$ does not converge to 0.

Theorem 5.7

Characterization theorem for convergence in L^1.
Let $(X_n)_n$ be a sequence of real-valued random variables then the following are equivalent:

1. $(X_n)_n$ converges in L^1;
2. $(X_n)_n$ converges in probability and is uniformly continuous;

3. $(X_n)_n$ is Cauchy for the L^1-convergence:

$$\lim_{m,n\to\infty} \mathbb{E}|X_m - X_n| = 0.$$

Proposition 5.9

$$X_n \xrightarrow{L^1} X \implies \lim_{n\to\infty} \mathbb{E}X_n Y - \mathbb{E}XY$$

And this is true for every bounded random variable Y.

So it is not true in generale, but requires boundedness. Quite a strict need. See what I did there?

Remark

Setting $Y = 1$ a.s. we get that

$$X_n \xrightarrow{L^1} X \implies \lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X.$$

5.6 Weak convergence and convergence in distribution

Definition 5.5

The sequence of probability measures $\mu_1, \mu_2, \mu_3, \dots$ on $\mathbb{R}$ is said to **converge weakly** to μ if

$$\lim_{n\to\infty} \underbrace{\mu_n f = \mu f}_{\text{integrals!!}} \qquad \forall f \in \mathbb{C}_b$$

where $\mathbb{C}_b$ is the set of continuous and bounded functions from $\mathbb{R}$ to $\mathbb{R}$.

So we are talking about convergence of probability measures and this is done through integrals. If the sequence of integrals converge towards a limiting integral then we say that the convergence happens for the *sequence of measures*: this shouldn't come as a surprise, since YOU should remember that measures are what defines integrals. But remember that random variables are ultimately characterized by their distributions (which are nothing else but probability measures!) so there must be some kind of connection between the distributions converging and their random variables doing the same thing...
Consider a sequence $(X_n)_n$ of random variables characterized by their sequence of distributions $(\mu_n)_n$.

Definition 5.6

The sequence $(X_n)_n$ is said to **converge in distribution** to a limiting random variable X with distribution μ if

$$(\mu_n)_n \xrightarrow{\text{weak}} \mu$$

that is to say if

$$\lim_{n\to\infty} \mathbb{E}f \circ X_n = \mathbb{E}f \circ X \qquad \forall f \in \mathbb{C}_b.$$

Note the notation!

We usually denote convergence in distribution as

$$X_n \xrightarrow{\text{d}} X.$$

Remark

$$X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{\mathrm{d}} X$$
$$X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathrm{d}} X$$
$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathrm{d}} X$$

So we see: convergence in distribution is the weakest of them all. It is related to the behavior in the limit of the probability distribution... and just having the same probability distribution is not saying much about a random variable. Imagine that we have $Y \sim N(0,1)$ and $Z = -Y$. Z is still a normal random variable, but clearly $Y(\omega) = a \implies Z(\omega) = -a$ so they are very different! But there is a special case...

Remark

If $X_n \xrightarrow{\mathrm{d}} X$ and $X = a$ a.s. (degenerate random variable[a]). Then

$$X_n \xrightarrow{\mathbb{P}} X.$$

[a] Yeah I know some degenerates "random variables" around here.

What the actual fuck? Now convergence in distribution implies convergence in probability? I will end my own life RIGHT NOW!

Proof

Let us choose

$$f(x) = \frac{|x-a|}{1-|x-a|}.$$

Note that $f \in \mathbb{C}_b$. So we have

$$\lim_{n\to\infty} \mathbb{E}\frac{|X_n - a|}{1-|X_n-a|} = \mathbb{E}f \circ X = \mathbb{E}\frac{|a-a|}{1-|a-a|} = 0.$$

Lol! Now

$$X_n \xrightarrow{\mathbb{P}} a.$$

□

Proposition 5.10

If μ and ν are probability measures on $\mathbb{R}$ and

$$\mu f = \nu f \qquad \forall f \in \mathbb{C}_n \implies \mu = \nu.$$

Definition 5.7

Take a distribution function $C(x) = \mathbb{P}(X \leqslant x)$ and construct:

$$q(u) = \inf\{x \in \mathbb{R} : C(x) > u\}.$$

This is known as **quantile function.**

So the quantile functions finds the first point at which the distribution function exceeds the level u. This is just the compositional inverse of C, if we think about it.

Proposition 5.11

Convergence of distribution functions and quantile functions.
The following points are equivalent:

1. $(\mu_n)_n \xrightarrow{\text{weak}} \mu$;
2. take $C_n(x) = \mathbb{P}(X_n \leqslant x)$ (distribution function: since it depends on n we get a sequence of functions with respect to the index n). We have
$$C_n(x) \to C(x) = \mathbb{P}(X \leqslant x)$$
for each point of continuity x of the distribution function C;
3. $q_n(u) \to q(u)$ for every point of continuity u of the quantile function q.

Theorem 5.8

Consider a sequence $(\mu_n)_n$ of probability measures such that
$$\mu_n \xrightarrow{\text{weak}} \mu,$$
where μ is a further probability measure. The above convergence is equivalent to the existance of a probability space
$$(\Omega', \mathcal{H}', \mathbb{P}')$$
and some random variables
$$Y_1, Y_2, \dots, Y$$
on it such that the distribution of Y_i is $\mu_i, i \in \mathbb{N}$ and of Y is μ and we have that
$$Y_n \xrightarrow{\text{a.s.}} Y$$
on $(\Omega', \mathcal{H}', \mathbb{P}')$.

This theorem is really important. We start with a sequence of weakly convergent probability measures, so we already know that on our initial space $(\Omega, \mathcal{H}, \mathbb{P})$ there are some random variables $X_1, X_2, \dots, X$ with those (convergent) measures. These random variables are still defined on $(\Omega, \mathcal{H}, \mathbb{P})$, but this theorem tells us that there exists *another space* $(\Omega', \mathcal{H}', \mathbb{P}')$ with other random variables $Y_1, Y_2, \dots, Y$ that are actually convergent almost surely! So if we have convergence in distribution and we want convergence almost surely we can have it, but at the cost of working in a different probability space.

Corollary

Skorokhod representation.
$(X_n) \xrightarrow{\text{d}} X$ if and only if there exist random variables $Y_1, Y_2, \dots, Y$ on some probability space such that
$$Y_n \overset{d}{=} X_n \quad \forall\, n, Y \overset{d}{=} x \text{ and } Y_n \xrightarrow{\text{a.s.}} Y.$$

I could have stuffed another Metal Gear Solid joke here with the similarity between Skorokhod and Shagohod but I won't. Maybe. Anyway, the greentext in figure 12 is enough for now.

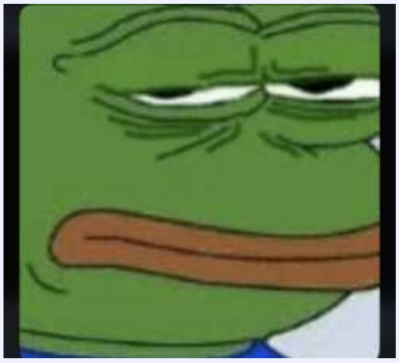
Proposition 5.12

If $X_n \xrightarrow{\text{d}} X$, the following are equivalent:

1. $(X_n)_n$ is uniformly integrable;
2. X_n and X are integrable and
$$\mathbb{E}|X_n| \to \mathbb{E}|X|.$$

But first let's introduce Fourier transform because... because fuck you. Let f_{nn} be the sequence

File: mfw.jpg (711 KB, 1170x1061)



Anonymous 08/07/24(Wed)17:24:38 No.16313511 ▶ >>16313525

>be me
>converging in distribution in my probability space $(\Omega, \mathscr{H}, \mathbb{P})$, not a big deal
>meet chick in uni
>ask her out
>got rejected because she only dates almost surely converging chads
>depression.jpg
>finally decide to get almost sure convergence to cope
> starting to realize that my sequences $(X_n)_n$ of random variables converging to X seem different now
>mfw now I am in $(\Omega', \mathscr{H}', \mathbb{P}')$
>"anon... what the fuck is that probability space?"

>i am now the laughing stock of my uni

is it over for me bros?

Anonymous 08/07/24(Wed)17:34:13 No.16313525 ▶

>>16313511 (OP)
absolutely brutal. As a 28 year old virgin who knows its over for me, this still hurts reading

Figure 12: Anon realizes he was deceived by probability theory[45].

of Fourier transforms of $(\mu_n)_n$, $\forall n$:

$$f_n(r) = \int_{\mathbb{R}} e^{irx} \mu_n(\mathrm{d}x), \qquad r \in \mathbb{R}.$$

Remember that this is a function of r!

Theorem 5.9

Let $(\mu_n)_n$, our sequence of probability measures, be weakly convergent if and only if

$$\lim_n f_n(r) = f(r) \qquad \forall r \in \mathbb{R}$$

and f is continuous at 0. Moreover, f is the transform of a probability measure μ on $\mathbb{R}$ and μ is the weak limit of $(\mu_n)_n$.

With this notion we study weak convergence by means of the Fourier transform. Why would we want to do this? I am seriously at a loss for words. But dear Professor Polito comes to help! Look back at the equation for the Fourier transform of a measure μ:

$$f(r) = \int_{\mathbb{R}} e^{irx} \mu(\mathrm{d}x), \qquad r \in \mathbb{R}$$

But this is an integral, which is... an expected value! consider x as being the random variable X and write

$$f(r) = \mathbb{E} e^{irX}.$$

This is called the **characteristic function of X** or the **fourier transfrom of its measure**.

Corollary

The sequence $(X_n)_n \xrightarrow{\mathrm{d}} X$ if

$$\lim_n \mathbb{E} e^{irX_n} = \mathbb{E} e^{irX} \qquad r \in \mathbb{R}.$$

But up to now we didn't say anything about the limiting random variables... to do so we need to introduce one of the main results in asymptotic analysis: the law of large numbers.

[45] Archived at https://boards.4chan.org/sci/thread/16313511

5.7 The law of large numbers

"Large numbers" always mean "what happens in the limit". Let's start with the basic theorem.

Theorem 5.10

Consider a sequence of real valued, pairwise independent random variables $X_1, X_2, \dots$ with finite common expectation and variance

$$\mathbb{E}X_n = a, \ \mathbb{V}\text{ar}\, X_n = b.$$

Then:

1. $\overline{x}_n = \dfrac{\sum_{i=1}^n X_i}{n} \xrightarrow{L^2} a$;
2. $\overline{x}_n \xrightarrow{\mathbb{P}} a$;
3. $\overline{x}_n \xrightarrow{\text{a.s.}} a$.

Huh. If you're anything like me this doesn't sound at all like the weak and strong laws of large numbers I've studied in the Economics degree. But again, what do I know. I'm sure we will find some surprising connections. Also we only need to prove the first and third point, because the second is implied by the third. Why does the theorem list it as the *third* point, though? What are you trying to do, *keeping me on the edge*?

Proof

1. Define
$$S_n = n\overline{X}_n$$
and
$$\mathbb{E}S_n = na, \ \mathbb{V}\text{ar}\, S_n = nb$$
because of hypothesis of independence. Then
$$\mathbb{E}\overline{X}_n = a, \ \mathbb{V}\text{ar}\, \overline{X}_n = \frac{b}{n}.$$
So if I look at
$$\mathbb{E}|\overline{X}_n - a|^2 = \mathbb{V}\text{ar}\, \overline{X}_n \xrightarrow[n\to\infty]{} 0$$
because the variance is a constant divided by n. Hence
$$\overline{X}_n \xrightarrow{L^2} a$$

2. This comes from Markov's inequality (which we already proved)...

3. Assume $X_n \geqslant 0$. This doesn't cause loss of generality because we can apply what we are going to prove to the positive and negative part of X_n (which are both positive!). Define $N = (n_k)_{k\in\mathbb{N}^*}$ such that $n_k = k^2$. By Chebyshev's inequality we get
$$\mathbb{P}(|\overline{X}_n - a| > \varepsilon) \leqslant \frac{b}{\varepsilon^2 k^2}$$
since we are working on the subsequence N where the variance of the random variables $\overline{X}_n$ is not $\frac{b}{n}$, but it is $\frac{b}{\text{value of subsequence}}$ which is k^2. Now sum and multiply by ε^2 to get
$$\varepsilon^2 \sum_{n\in N} \mathbb{P}(|\overline{X}_n - a| > \varepsilon) \leqslant \sum_{k=1}^{\infty} \frac{b}{k^2} < \infty \qquad \varepsilon > 0$$
since the latter is a geometric series. But this tells us that $\sum_{n\in N} \mathbb{P}(|\overline{X}_n - a| > \varepsilon)$ is a finite number and therefore... we can apply Borel-Cantelli lemma!

Revise with Kotatsu!

Borel Cantelli lemma:

$$\sum_n \mathbb{P}(H_n) < +\infty \implies \sum_n \mathbb{1}_{H_n} < +\infty \quad \text{a.s.}$$

and in particular if

$$\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < +\infty \qquad \forall \varepsilon > 0.$$

then

$$X_n \xrightarrow{\text{a.s.}} X.$$

So

$$\overline{X}_n \to a$$

along N. Let us call Ω_0 the almost sure set on which the convergence takes place and $\forall \omega \in \Omega_0$ we apply the following lemma (we have already proven)

Revise with Kotatsu!

If we have a sequence of positive numbers $(x_n)_n$ and consider the empirical sum $\overline{x}_n = \frac{\sum_{i=1}^n x_i}{n}$ and let $N = (n_k) \subset \mathbb{N}^*$ with $\lim_k \frac{n_{k+1}}{n_k} = r > 0$ (asymptotic linarity) and $(\overline{x}_n)_n$ converges along N to x then

$$\frac{x}{r} \leqslant \liminf \overline{x}_n \leqslant \limsup \overline{x}_n \leqslant rx.$$

Note that

$$\frac{(k+1)^2}{k^2} \xrightarrow[k\to\infty]{} 1$$

and hence $\forall \omega \in \Omega_0$ we have that

$$\lim_{n\to\infty} \overline{X}_n(\omega) = a \qquad \forall \omega \in \Omega_0$$

with probability 1... which means, almost sure convergence!

□

This one was for sequence of random variables with finite variance and expectations... But what if both are infinite?

Proposition 5.13

Let $(X_n)_n$ be a sequence of positive independent and identically distributed (i.i.d.) random variables with

$$\mathbb{E}X_1 = +\infty.$$

Consider also a further random variable X distributed as X_1 (which means that they also have the same expectation). Then

$$\overline{X}_n \xrightarrow{\text{a.s.}} \infty.$$

This shouldn't be surprising, since we are given a sequence of positive random variables with infinite expectation, that means that there is a lot of the probability mass resides in the tails. This causes the expectation to diverge when doing the integral. To prove this fact we must rely on the previous result.

Proof

We know that this behaviour is caused by the tail of the distribution, so we are going to truncate the distribution.
Fix $b \in \mathbb{R}$ and let $Y_n = X_n \wedge b$. So X_n can have infinite support, while Y_n is definitely bounded with the support finishing at b (**truncated random variable**). Let also

$$\overline{Y}_n = \frac{\sum_{i=1}^n Y_i}{n}$$

and note that

$$\mathbb{E}Y_n = \mathbb{E}(X_n \wedge b) < +\infty.$$

By the previous theorem we know that

$$\overline{Y}_n \xrightarrow{\text{a.s.}} \mathbb{E}(X \wedge n).$$

The second result that we need is an ordering between X_n and Y_n. Which is the largest? For each n we have that $X_n \geqslant Y_n$ and this further implies that $\liminf_n \overline{X}_n \geqslant \lim_n \overline{Y}_n = \mathbb{E}(X \wedge b)$ almost surely. This must be true for any choice of b, even if $b \to \infty$ and thus

$$\mathbb{E}(X \wedge b) \xrightarrow[b\to\infty]{} \mathbb{E}X = +\infty$$

using the monotone covergence theorem. That is,

$$\liminf_n \overline{X}_n = +\infty \text{ a.s.}$$

Which basically means, since $\overline{X}_n$ is a non decreasing sequence, that

$$\overline{X}_n \xrightarrow{\text{a.s.}} +\infty$$

□

There is a general theorem that puts together the last two theorems that we have seen so far:

Theorem 5.11

Law of large numbers:
Let $(X_n)_n$ be a sequence of pairwise independent random variables with the same distribution as X. If $\mathbb{E}X$ exists (infinite values are admitted!) then

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

This is an almost sure result and that's why we often call this theorem the **strong law of large numbers** (SLLN). Here are two useful inequalities:

- **Chebyshev's inequality**: here we are given a sequence of random variables X_n and build up the partial sums S_n. Here we must assume that $(X_n)_n$ are i.i.d. and such that $\mathbb{E}X_1 = 0$.

  $$\varepsilon^2 \mathbb{P}(|S_n| > \varepsilon) \leqslant \mathbb{V}\text{ar}\, S_n = \underbrace{\mathbb{E}S_n^2}_{\text{second moment!}}$$

  so the variance gives us an upper bound of the probability $\mathbb{P}(|S_n| > \varepsilon)$. If we don't have that $\mathbb{E}X_1 = 0$ then $\mathbb{V}\text{ar}\, S_n \leqslant \mathbb{E}S_n^2$ (the second moment is usually larger than the variance... still a good upper bound tho).

- **Kolmogorov's inequality**: let $(X_d)_n$ be a sequence of independent random variables such that $\mathbb{E}X_n = 0\ \forall n$. Then $\forall a \in (0, \infty)$

  $$a^2 \mathbb{P}\left(\max_{k\leqslant n} |S_k| > a\right) \leqslant \mathbb{V}\text{ar}\, S_n$$

  We are interested in the times from 1 to n: the interesting thing is that the maximum value of the random walk is bounded by something that happens in time n.

We will prove this last inequality.

Proof

Fix $a > 0,\ n \geqslant 1$ and define

$$N(\omega) = \inf\{k \geqslant 1 : |S_k(\omega)| > a\} \qquad \forall\, \omega.$$

So for that specific ω we consider the first time for which the random walk exceeds the value a. Of course this time index k depends on ω, so $N(\omega)$ is a random variable. Observe a couple of facts:

1) $N(\omega) = k \iff |S_k(\omega)| > a$ and $S_j(\omega) \leqslant a \quad \forall\, j < k$. Then $\mathbb{1}_{N=k}$ is a function of $(X_1, X_2, \ldots, X_n)$ because I need the knowledge of X up to k to know whether $N = k$.

2) For $k < n$ we define the following random variables:

$$U = S_k \mathbb{1}_{N=k} \qquad \text{and} \qquad V = S_n - S_k.$$

We know that U depends on the first k values of X_n, while V depends only on $X_{k+1}, X_{k+2}, \ldots, X_n$. But this means that U and V are functions of independent random variables, so $U \perp\!\!\!\perp V$. Hence

$$\mathbb{E}UV = \mathbb{E}U \underbrace{\mathbb{E}V}_{=0 \text{ because } \mathbb{E}X_n = 0}.$$

Therefore

$$\mathbb{E}S_k \mathbb{1}_{\{N=k\}}(S_n - S_k) = 0.$$

3) Consider S_n^2:

$$S_n^2 = [S_k + (S_n - S_k)]^2 \geqslant S_k^2 + 2S_k(S_n - S_k)$$

The inequality is what I get if I neglect the negative term. So i am bounding what happens at time n with what happens at time k. Note that $|S_k|^2 > a^2$ when $\{N = k\}$ holds. Then

$$\mathbb{E}S_n^2 \mathbb{1}_{N=k} \leqslant \underbrace{a^2 \mathbb{E}\mathbb{1}_{\{N=k\}}}_{a^2 \mathbb{P}(N=k)} + 2 \underbrace{\mathbb{E}S_k(S_n - S_k)\mathbb{1}_{\{N=k\}}}_{=0}$$

and summing both sides for $k \leqslant n$ we get

$$a^2 \mathbb{P}(N \leqslant n) \leqslant \mathbb{E}S_n^2 \mathbb{1}_{\{N \leqslant n\}} \leqslant \mathbb{E}S_n^2 = \underbrace{\mathbb{V}\mathrm{ar}\, S_n}_{\text{because of independence}}.$$

Hence

$$a^2 \mathbb{P}(N \leqslant n) \leqslant \mathbb{V}\mathrm{ar}\, S_n$$

But we know that

$$\mathbb{P}(N \leqslant n) = \mathbb{P}(\max_{k \leqslant n} |S_k| > a)$$

So our proof is finished. □

Proposition 5.14

Let $(X_n)_n$ be independent and with mean 0. If

$$\sum_n \mathbb{V}\mathrm{ar}\, X_n < +\infty \implies \sum_n X_n < +\infty \text{ a.s.}$$

Proposition 5.15

Let $(X_n)_n$ be a bounded sequence of i.i.d. random variables. If

$$\sum_n (X_n - a_n) < +\infty \text{ a.s.}$$

for some sequence $(a_n)_n \subset \mathbb{R}$ then

$$\sum_n \mathbb{V}\text{ar}\, X_n < +\infty.$$

Theorem 5.12

Kolmogorov's 3-series theorem.
Let $(X_n)_n$ be independent random variables. We have

$$\sum_n X_n < +\infty \text{ a.s.}$$

if and only if the following 3 series are convergent:

1. $\sum_n \mathbb{P}(X_n \neq Y_n)$ where $Y_n = X_n \mathbb{1}_{\{|X_n|<b\}}$;
2. $\sum_n \mathbb{E}Y_n$;
3. $\sum_n \mathbb{V}\text{ar}\, Y_n$.

So: remember that expectation is a number, so the LLN tells us that $\overline{X}_n$ (a random variable) tends to $\mathbb{E}X$ (a degenerate random variable): our variability gets lots as $n \to \infty$. This affects many characteristics of the random variable...

5.8 Central limit theorem

There are different versions of the Central limit theorem and we have probably already studied some of those: the most common in undergraduate courses is the **De Moivre-Lapalce Central limit Theorem**.

Theorem 5.13

Let $(X_i)_i$ be a sequence of i.i.d. real-valued random variables with finite mean $\mathbb{E}X_1 = a$ and finite variance $\mathbb{V}\text{ar}\, X_1 = b$. Then

$$\mathcal{Z}_n = \frac{S_n - na}{\sqrt{nb}} \xrightarrow{\text{d}} Z \sim N(0,1)$$

The usual way to prove this theorem is relying on characteristic function. If $a = 0, b = 1$ then $Z_n = \frac{S_n}{\sqrt{n}}$ where Z_n is the sum of "small" indpendent random variables $X_{n,1} = \frac{X_1}{\sqrt{n}}, X_{n,2} = \frac{X_2}{\sqrt{n}}, \ldots, X_{n,n} = \frac{X_n}{\sqrt{n}}$. All of them are "small" with respect to their original random variable X_n because we are dividing by $\sqrt{n}$: they get smaller and smaller as $n \to \infty$ but overall, for each n, Z_n has a stable mean 0 and variance 1. The general idea behind the CLT is exactly this: we divide all of these variables for the "right amount" so that the limit of our sequence of random variables stays stable and does not become a degenerate limit.

Definition 5.8

A **triangular array** (Professor Polito calls it "a nice object") is an infinite matrix with a special structure:

$$[X_{nj}] = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \ldots \\ X_{21} & X_{22} & X_{23} & \ldots \\ X_{31} & X_{32} & X_{33} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where the entries are real-valued random variables (so it is actually an infinite random metric). Furthermore, for each n there exists an integer k_n such that $X_{nj}, j > k_n$. The sequence of indices $(k_n)_n$ is increasing towards ∞.

So the random variables are equal to 0 to the right hand side of the index k_n for each row. The first row has k_1, the second row is composed by $k_2 > k_1$ (so there are more non-zero random

variables than in row 1) and so on. So it's not really a triangular matrix (there's no rule on *how much* k_n is larger than k_{n-1}) but it is "basically" triangular.
Now define

$$Z_n = \sum_j X_{nj} \qquad (j = 1, 2, \ldots, k_n).$$

Remember that even if we are summing infinite random variables only a finite number of them are non-zero. Now consider the assumption that the random variables on each row are independent among them (there *may* be dependence with random variables in other rows.). With this structure in mind we can formalize the following formulation of the CLT:

Theorem 5.14

Lyapunov Central limit Theorem:
Suppose $\mathbb{E}X_{nj} = 0 \,\forall n, j$ and $\mathbb{V}\mathrm{ar}\, Z_n = 1 \,\forall n$ and $\lim_n \sum_n \mathbb{E}|X_{nj}|^3 = 0$. Then

$$Z_n \xrightarrow{\mathrm{d}} Z \sim N(0, 1).$$

Here we have replaced the conditions on the random variables with the condition $\lim_n \sum_n \mathbb{E}|X_{nj}|^3 = 0$ which allows us to use the structure of the triangular array. To prove this theorem we need the following lemma:

Lemma 5.3

Lindeberg's lemma: Let $(Y_1, Y_2, \ldots, Y_k)$ be independent random variables with mean zero and let $S = \sum_{j=1}^k Y_j$. Let us further assume that $\mathbb{V}\mathrm{ar}\, S = 1$. Let f be a function which can be differentiated 3 times and let f', f'', f''' be bounded and continuous and such that

$$|f'''| \leqslant x, \qquad c \in \mathbb{R}_+.$$

Then for $Z \sim N(0, 1)$

$$|\mathbb{E}f \circ S - \mathbb{E}f \circ Z| \leqslant c \sum_{j=1}^k \mathbb{E}|Y_j|^3$$

This means that if the hypotheses are met then the distance between S and a normal random variable is bounded so we can approximate S by means of a normal random variable. We will prove this lemma.

Proof

Let $Z_1, \ldots, Z_k$ be independent normal random variables with mean $\mathbb{E}Z_j = \mathbb{E}Y_j = 0$ for $j = 1, \ldots, k$ and variance $\mathbb{V}\mathrm{ar}\, Z_j = \mathbb{V}\mathrm{ar}\, Y_j$ for $j = 1, \ldots, k$. Then construct

$$T = \sum_{j=1}^k Z_j \sim N(0, \sum_{j=1}^k \mathbb{V}\mathrm{ar}\, Z_j = \sum_{j=1}^k \mathbb{V}\mathrm{ar}\, Y_j = 1).$$

So we know that T is distributed as Z (they are both $N(0, 1)$) so $T \overset{d}{=} Z$ and since we are using the expectation of Z we can replace $\mathbb{E}f \circ Z$ with $\mathbb{E}f \circ T$. So the Linderberg's lemma becomes

$$|\mathbb{E}f \circ S - \mathbb{E}f \circ T| \leqslant c \sum_{j=1}^k \mathbb{E}|Y_j|^3$$

which we want to prove, to exploit the structure of T. Define now the random variables $V_1, V_2, \ldots, V_k$ as follows:

$$\begin{aligned}
V_1 \quad &\text{s.t.} \quad S = V_1 + Y_1 \\
V_2 \quad &\text{s.t.} \quad V_1 + Z_1 = V_2 + Y_2 \\
&\vdots \\
V_j \quad &\text{s.t.} \quad V_j + Z_j = V_{j+1} + Y_{j+1}, \quad 1 \leqslant j < k \\
&\vdots \\
V_k \quad &\text{s.t.} \quad V_k + Z_k = T.
\end{aligned}$$

Note that

$$\begin{aligned} V_1 &= Y_2 + Y_3 + \ldots + Y_k \\ V_2 &= Z_1 + Y_3 + \ldots + Y_k \\ V_3 &= Z_1 + Z_2 + Y_4 + \ldots + Y_k \end{aligned}$$

so the Y get replaced by the Z one at the time in the V. We can now focus on the following expression:

$$\begin{aligned} f \circ S - f \circ T &= f(V_1 + Y_1) - f(V_k - Z_k) \\ &= f(V_1 + Y_1) + f(V_2 + Y_2) - \underbrace{f(V_2 + Y_2)}_{f(V_1+Z_1)} + f(V_3 + Y_3) - \underbrace{f(V_2 + Y_2)}_{f(V_2+Z_2)} + \\ &+ \ldots + f(V_k + Z_k) \\ &= \sum_{j=1}^{k} f(V_j + Y_j) - \sum_{j=1}^{k} f(V_j + Z_j). \end{aligned}$$

Now take the expectation and the absolute value:

$$\begin{aligned} |\mathbb{E}f \circ s - \mathbb{E}f \circ t| &= \left| \sum_{j=1}^{k} \mathbb{E}f(V_j + Y_j) - \sum_{j=1}^{k} \mathbb{E}f(V_j + Z_j) \right| \\ &\leqslant \sum_{j=1}^{k} |\mathbb{E}f(V_j + Y_j) - \mathbb{E}f(V_j + Z_j)| \end{aligned}$$

and now we only need to prove that

$$|\mathbb{E}f(V_j + Y_j) - \mathbb{E}f(V_j + Z_j)| \leqslant c\mathbb{E}|Y_j|^3.$$

Let's write the Taylor formula for this function:

$$f(v + x) = f(v) + f'(v)x + \frac{1}{2}f''(v)x^2 + R_2(v, x)$$

where

$$\begin{aligned} R_2(v, x) &= \frac{1}{2}\int_v^{v+x} (v + x - t)_2 f''(t)\,\mathrm{d}t \\ &\leqslant \frac{1}{2}c\int_v^{v+x} (v + x - t)_2\,\mathrm{d}t = c\frac{x^3}{6} \end{aligned}$$

so that

$$|R_2(v, x)| \leqslant \frac{c}{6}|x|^3.$$

We now have

$$\begin{aligned} f(V_j + Y_j) &= f(V_j) + f'(V_j)Y_j + \frac{1}{2}f''(V_j)Y_j^2 + R_2(V_j, Y_j) \\ f(V_j + Z_j) &= f(V_j) + f'(V_j)Z_j + \frac{1}{2}f''(V_j)Z_j^2 + R_2(V_j, Z_j) \end{aligned}$$

and subtract side by side:

$$f(V_j + Y_j) - f(V_j + Z_j) = (Y_j - Z_j)f'(V_j) + \frac{1}{2}f''(V_j)(Y_j^2 - Z_j^2) + R_2(V_j, Y_j) - R_2(V_j, Z_j).$$

Now take the expectation

$$\begin{aligned} \mathbb{E}f(V_j + Y_j) - \mathbb{E}f(V_j + Z_j) &= \frac{1}{2}\mathbb{E}f''(V_j)(\underbrace{\mathbb{E}Y_j^2 - \mathbb{E}Z_j^2}_{=0 \text{ since they have same variance}}) + \mathbb{E}\left[R_2(V_j, Y_j) - R_2(V_j, Z_j)\right] \\ &= \mathbb{E}\left[R_2(V_j, Y_j) - R_2(V_j, Z_j)\right]. \end{aligned}$$

Now take the absolute value

$$\begin{aligned}|\mathbb{E}\left[R_2(V_j,Y_j)-R_2(V_j,Z_j)\right]| &\leqslant \mathbb{E}|\left[R_2(V_j,Y_j)\right]|+\mathbb{E}\left[|R_2(V_j,Z_j)|\right]\\ &\leqslant \frac{c}{6}\left(\mathbb{E}|Y|^3+\mathbb{E}|Z|^3\right)\end{aligned}$$

so that

$$|\mathbb{E}f(V_j+Y_j)-\mathbb{E}f(V_j+Z_j)|\leqslant \frac{c}{6}\left(\mathbb{E}|Y|^3+\mathbb{E}|Z|^3\right).$$

Recall that $Z_j\sim N(0,b^2)$ where $b^2=\mathbb{E}Y_j^2$. We know that

$$\mathbb{E}|Z_j|^3=b^3\sqrt{\frac{8}{\pi}}\leqslant 2b^3$$

and we also have

$$b=\left(\mathbb{E}Y_j^2\right)^{\frac{1}{2}}\leqslant\left(\mathbb{E}|Y_j|^3\right)^{\frac{1}{3}}$$

Because L^2 norm is less or equal than L^3 norm (revise inclusions in L^p-spaces for different values of p). But this last inequality is equivalent to

$$b^3\leqslant\mathbb{E}|Y_j|^3$$

which leads to

$$\mathbb{E}\left|Z_j\right|^3\leqslant 2b^3\leqslant 2\mathbb{E}\left|Y_j\right|^3.$$

Finally we get

$$\begin{aligned}|\mathbb{E}f(V_j+Y_j)-\mathbb{E}f(V_j+Z_j)| &\leqslant \frac{c}{6}\left(\mathbb{E}\left|Y_j\right|^3\right)\\ &=\frac{c}{6}3\mathbb{E}\left|Y_3\right|^3\\ &=\frac{c}{2}\mathbb{E}\left|Y_3\right|^3\leqslant c\mathbb{E}\left|Y_3\right|^3.\end{aligned}$$

□

This was horrible, horrible. Truly an horrible experience and honestly useless proof. And we still have to prove Lyapunov's theorem.

Proof

Lyapunov's CLT. Recall

$$Z_n=\sum_j X_{nj}\qquad Z\sim N(0,1).$$

We are interested in evaluating the characteristic function.

$$\begin{aligned}e^{irZ_n}&=\cos rZ_n+i\sin rZ_n\\ e^{irZ}&=\cos rZ+i\sin rZ.\end{aligned}$$

Consider now

$$\begin{aligned}\left|\mathbb{E}e^{irZ_n}-\mathbb{E}e^{irZ}\right|&=\left|\left(\mathbb{E}\cos rZ_n-\mathbb{E}\cos rZ\right)+i\left(\mathbb{E}\sin rZ_n-\mathbb{E}\sin rZ\right)\right|\\ &\leqslant\left|\mathbb{E}\cos rZ_n-\mathbb{E}\cos rZ\right|+\underbrace{|i|}_{1}\left|\mathbb{E}\sin rZ_n-\mathbb{E}\sin rZ\right|.\end{aligned}$$

By applying the above lemma we obtain

$$\left|\mathbb{E}e^{irZ_n}-\mathbb{E}e^{irZ}\right|\leqslant\sum_j|r|^3\mathbb{E}|X_{nj}|^3+\sum_j|r|^3\mathbb{E}|X_{nj}|^3$$

and this is possible since both sine and cosine are differentiable three times and they are bounded by 1 (so our $c = 1$). Now, according to the hypotheses of Lyapunov's theorem, we need to take the limit considering the hypothesis that $\lim_{n\to\infty} \sum_j \mathbb{E}|X_{nj}|^3 = c$ and we obtain the claim. This theorem applies to all triangular arrays which include the one in the CLT. □

WHY.

6 Conditional expectations

Are you comfortable with the concept of expectation? Well brace yourself buddy because we are going to fuck it up and reveal that normal expectations are just a particular case of conditional expectations.

Consider our dear old probability space $(\Omega, \mathcal{H}, \mathbb{P})$, a $\overline{\mathbb{R}}$-valued random variable X, a sub-σ-algebra $\mathcal{F} \in \mathcal{H}$ where $\mathcal{F}$ is the space of $\mathcal{F}$-measurablerandom variables taking values in $\overline{\mathbb{R}}$. Remember that $\mathcal{F}$ can be considered the body of information revealed by our random variable. Consider now the event H. Fix $\omega \in \Omega$ and pretend that our only knowledge of ω is that of the $\omega \in H$. First of all, remember that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{\mathbb{P}(B)} \int_A \mathbb{1}_B \, \mathrm{d}\mathbb{P}$$

Evaluate our best estimate for $X(\omega)$, which may very well be the average of X over H:

$$\frac{1}{\mathbb{P}(H)} \int_H X(\omega)\mathbb{P}(\mathrm{d}\omega) = \frac{1}{\mathbb{P}(H)} \mathbb{E}X\mathbb{1}_H = \mathbb{E}_H X.$$

If $\mathbb{P}(H) = 0$ then we allow any value for $\mathbb{E}_H X$.

Note the notation!

We usually denote that quantity as $\mathbb{E}(X|H)$ but Çinlar denotes it as $\mathbb{E}_H X$.

Another interesting way to look at this question is the following: conditional probability basically gives us a new probability measure while leaving the rest of the "ingradients" of the probability triplet unchanged. And it makes quite sense: we change the way we weigh a certain event. So if we want, for instance, to condition our probability measure to the event B our probability space $(\Omega, \mathcal{H}, \mathbb{P})$ becomes $(\Omega, \mathcal{H}, \mathbb{P}(\cdot|B))$. This also means that our expectation $\mathbb{E}X = \int_\Omega X \, \mathrm{d}\mathbb{P}$ becomes $\mathbb{E}_B X = \int_\Omega X \, \mathrm{d}\mathbb{P}(\cdot|B)$.

Remark

This best estimate is "best" in the same sense as the expectation of X is the best estimate of X when we do not know anything about ω. The unconditional case is retrieved if $H = \Omega$.

Remark

The quantity $\mathbb{E}_H X$ is called **conditional expectation of X given the event H**.

Up to now, nothing new. Let's go a step further. Now suppose that $\mathcal{F}$ is generated by a measurable partition $(H_n)_n$ of Ω. Fix $\omega \in \Omega$ and consider the problem of building an estimate of the random variable $X(\omega)$ with only the body of information given by the partition/sub-σ-algebra $\mathcal{F}$: we could construct a "best estimate" of $X(\omega)$ for every event H_n:

$$\overline{X}(\omega) = \sum_n \mathbb{E}_{H_n} X \mathbb{1}_{H_n}(\omega). \tag{𝅘𝅥𝅮}$$

So for every specific ω only one term of the sum will be different from 0: if ω_0 belongs in H_0 then only $\mathbb{E}_{H_0} X$ in the sum will be non-zero and $\overline{X}(\omega_0) = \mathbb{E}_{H_0} X$. So in this sense we can summarize all the information available. So $\overline{X}(\omega)$ is a random variable that we will call **conditional expectation of X given $\mathcal{F}$** and we will write

$$\overline{X} = \mathbb{E}_{\mathcal{F}} X \quad \text{or} \quad \mathbb{E}\left[X|\mathcal{F}\right].$$

So expectation of a random variable is a *number*, while the conditional expectation with respect to a sub-σ-algebra is a *random variable*.

Example 6.1

Imagine we have $\Omega = (0,1), \mathcal{H} = \mathcal{B}_{(0,1)}, \mathbb{P} = \lambda$. Consider the partition of Ω given by $\mathcal{F} = \sigma(\{H_1, H_2, H_3\})$. The wiggly line is the realization of our process $X(\omega)$.

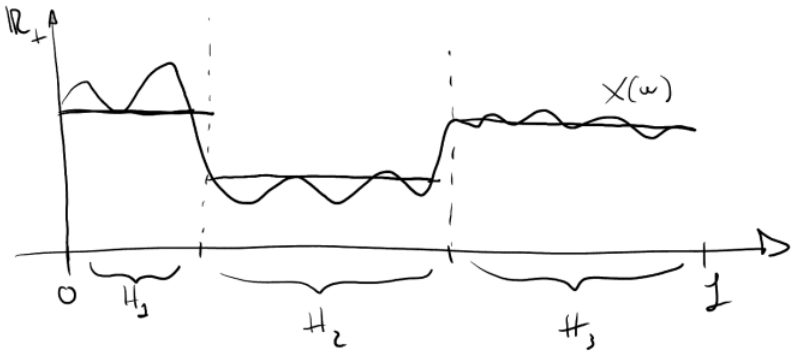


Figure 13: 𝅘𝅥𝅯 I think the apple is rotten right to the cooore

So according to our formula (𝅘𝅥𝅯) in our first partition H_1 we have a certain expected value (which is an average[46]: expectation of a random variable is a number...) which is different from the value in the second partition an in the third.

6.1 Conditional expectation conditional on arbitrary σ-algebras

Consider the following facts:

1. $\overline{X} \in \mathcal{F}$: it is measurable with respect to $\mathcal{F}$ and this is clear from the definition 𝅘𝅥𝅯 of the random variable $\overline{X}$;
2. we know that for each $V \in \mathcal{F}_+$ we have that

$$\mathbb{E}VX = \mathbb{E}\overline{X}$$

which is called **projection property**.

Proof

We start by letting $V = \mathbb{1}_{H_n}$ for some fixed n (and consider $\overline{X} = \mathbb{E}_{H_n} X$). The statement becomes

$$\mathbb{E}\mathbb{1}_{H_n} X = \mathbb{E}\mathbb{1}_{H_n} \overline{X}.$$

The left hand side reads

$$\mathbb{E}\mathbb{1}_{H_n} X = \int_\Omega \mathbb{1}_{H_n}(\omega) X(\omega) \mathbb{P}(\mathrm{d}\omega) = \int_{H_n} X(\omega) \mathbb{P}(\mathrm{d}\omega).$$

The right hand side reads

$$\mathbb{E}\left[\mathbb{1}_{H_n} \cdot \frac{1}{\mathbb{P}(H_n)} \int_{H_n} X(\omega)\mathbb{P}(\mathrm{d}\omega)\right] = \frac{1}{\cancel{\mathbb{P}(H_n)}} \int_{H_n} X(\omega)\mathbb{P}(\mathrm{d}\omega)\cancel{\mathbb{P}(H_n)}$$
$$= \int_{H_n} X(\omega)\mathbb{P}(\mathrm{d}\omega).$$

Now consider a general random variable $V \in \mathcal{F}_+$. Of course we have (for construction)

$$V = \sum_n a_n \mathbb{1}_{H_n}$$

[46] It's funny how this shit makes you really forget that the expected value is just the fucking average.

so V is a simple function. Here,

$$\begin{aligned}
\mathbb{E}V\overline{X} &= \mathbb{E}\left[\sum_n a_n \mathbb{1}_{H_n} \sum_n (\mathbb{E}_{H_n} X \mathbb{1}_{H_n})\right] \\
&= \mathbb{E}\left[\sum_n a_n \mathbb{E}_{H_n} X \mathbb{1}_{H_n}\right] \\
&= \sum_n a_n \mathbb{E}\left[\mathbb{1}_{H_n} \mathbb{E}_{H_n} X\right] \\
&= \sum_n a_n \mathbb{E}\left[\mathbb{1}_{H_n} X\right] \qquad \text{for the previous result} \\
&= \mathbb{E}\left[\sum_n a_n \mathbb{1}_{H_n} X\right] \\
&= \mathbb{E}V X.
\end{aligned}$$

□

These two properties are the basis for the extended definition of conditional expectation.

Definition 6.1

Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. The conditional expectation $\mathbb{E}_{\mathcal{F}} X$ of X given $\mathcal{F}$ is defined in two steps:

(a) for $X \in \mathcal{H}_+$ (positive random variables) it is any random variable $\overline{X}$ satisfying:

  (a) measurability ($\overline{X} \in \mathcal{F}_+$);
  (b) projection property ($\mathbb{E}VX = \mathbb{E}V\overline{X} \qquad \forall V \in \mathcal{F}_+$).

(b) for arbitrary $X \in \mathcal{H}$, if $\mathbb{E}X$ exists, we define

$$\mathbb{E}_{\mathcal{F}} X = \mathbb{E}_{\mathcal{F}} X^+ - \mathbb{E}_{\mathcal{F}} X^-.$$

Otherwise, if $\mathbb{E}X^+ = \mathbb{E}X^- = \infty$, then $\mathbb{E}_{\mathcal{F}}$ is left undefined.

Any random variable satisfying the first property is eligible as conditional expectation: that's why we call it a **version** of the conditional expectation of the random variable.

Revise with Kotatsu!

If Y and Z are positive and $\mathcal{F}$-measurable and if $\mathbb{E}VY \leqslant \mathbb{E}VZ$ for each $V \in \mathcal{F}_+$ then $Y \leqslant Z$ almost surely. Moreover, if $\mathbb{E}VY = \mathbb{E}VZ, \quad \forall V \in \mathcal{F}_+$ (or $\forall \mathbb{1}_H \in \mathcal{F}_+$), then $Y = Z$ a.s.

6.2 Uniqueness of conditional expectation

This is a simple matter. Let $\overline{X}$ and $\overline{\overline{X}}$ be versions of $\mathbb{E}_{\mathcal{F}} X$, $X \geqslant 0$.

1. Both $\overline{X}$ and $\overline{\overline{X}}$ are $\mathcal{F}^+$-measurable;
2. $\mathbb{E}VX = \mathbb{E}V\overline{X} = \mathbb{E}V\overline{\overline{X}}$ for every $V \in \mathcal{F}_+$. Hence

$$\overline{X} = \overline{\overline{X}} \text{ a.s.}$$

Conversely, if $\mathbb{E}_{\mathcal{F}} X - \overline{X}$ and $\overline{\overline{X}} \in \mathcal{F}_+$ and $\overline{X} = \overline{\overline{X}}$ a.s. then $\overline{\overline{X}}$ satisfies the projection property $\mathbb{E}XV = \mathbb{E}\overline{\overline{X}}V$ (i.e. $\overline{\overline{X}}$ is a version of $\mathbb{E}_{\mathcal{F}} X$).

Remark

Tower rule: this is the projection property for $V = 1, X \in \mathcal{H}_+$. We have

$$\mathbb{E}X = \mathbb{E}\overline{X} = \mathbb{E}\mathbb{E}_{\mathcal{F}} X.$$

Moreover, if X is integrable then $\mathbb{E}_{\mathcal{F}}X$ is integrable.

Remark

If X is integrable we can rewrite the projection property as follows:

$$\mathbb{E}V(X-\overline{X})=0 \qquad \forall\, V\in\mathcal{F}_b.$$

There is an interesting fact linked to this last representation. Let X be integrable: then $(X-\overline{X})$ is integrable. Define

$$\widetilde{X}=X-\overline{X}$$

and write the following decomposition:

$$X=\overline{X}+\widetilde{X}.$$

Here $\overline{X}$ is determined by the information given by $\mathcal{F}$ while $\widetilde{X}$ is orthogonal to $\mathcal{F}$, that is to say that $\mathbb{E}\mathbb{1}_H\widetilde{X}=0$, $\forall\, H\in\mathcal{F}$. We call $\overline{X}$ the **orthogonnal projection of X on $\mathcal{F}$**.

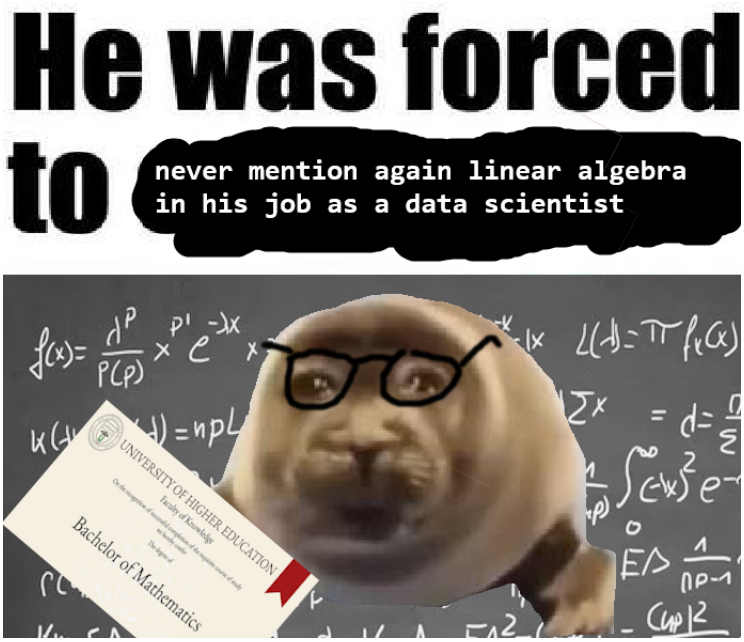


Figure 14: Absolutely honey you *need* all of this stuff to be a data scientist.

We are still to prove the existence of conditional expectation, but to do so we need some measure theory notions... Are you happy?

Revise with Kotatsu!

Let $(E,\mathcal{E},\mu)$ be a measurable space and let g ve $\mathcal{E}_+$-measurable function. Define

$$\nu(A)=\mu(g\mathbb{1}_A)=\mathbb{1}_A g(x)\mu(\mathrm{d}x) \qquad \forall\, A\in\mathcal{E}.$$

It can be proved that ν is a measure on $(E,\mathcal{E})$. So putting togheter $(E,\mathcal{E})$ with out new measure ν we can get a new measurable space.

Proposition 6.1

For every $f\in\mathcal{E}_+$ we have that (for such measures μ and ν)

$$\nu f=\mu(gf)=\int_E f(x)\nu(\mathrm{d}x)=\int_E g(x)f(x)\mu(\mathrm{d}x).$$

Definition 6.2

Consider a measurable space $(E, \mathcal{E})$ and two measures on it, say μ and ν (two general measures). Then ν is said to be **absolutely continuous** with respect to the other measure μ if

$$\mu(A) = 0 \implies \nu(A) = 0 \qquad \forall A \in \mathcal{E}.$$

Note that if $\nu(A) = \int_A g(x)\mu(\mathrm{d}x)$ then ν is absolutely continuous with respect to μ, of course. This is a good way to construct an absolutely continuous measure!

Theorem 6.1

Radon-Nikodyn theorem.
Suppose that μ is σ-finite and ν is absolutely continuous with respect μ then there exists a positive $\mathcal{E}$-measurable function g such that

$$\int_E f(x)\nu(\mathrm{d}x) = \int_E f(x)g(x)\mu(\mathrm{d}x) \qquad (\Delta)$$

Moreover, g is unique up to equivalence (that is, if (Δ) holds for another $\hat{g} \in \mathcal{E}_+$ then $g = \hat{g}$ μ-almost everywhere).

The function g is called **density** or **Radon-Nikodyn derivative of measure ν with respect to measure μ**.

Theorem 6.2

Let $X \in \mathcal{H}$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then the conditional expectation $\mathbb{E}_{\mathcal{F}}X$ exists and it is unique up to equivalence.

Let's prove this theorem for $\mathcal{H}$-positive random variables.

Proof

$\forall H \in \mathcal{F}$ on the (reduced) measurable space $(\Omega, \mathcal{F})$. Now on the (reduced) measurable space we consider the restriction of $\mathbb{P}$ on $\mathcal{F}$. Further consider

$$\mathbb{Q}(H) = \int_H \mathbb{P}(\mathrm{d}\omega)X(\omega)$$

Where $\mathbb{P}$ is a probability measure and $\mathbb{Q}$ is a measure which is absolutely continuous with respect to $\mathbb{P}$. Remember that in this measurable space random variables are functions, so we can apply Radon-Nikodyn theorem: this tells us that there exists a random variable $\overline{X} \in \mathcal{F}_+$ such that

$$\int_\Omega \mathbb{Q}(\mathrm{d}\omega)V(\omega) = \int_\Omega \mathbb{P}(\mathrm{d}\omega)\overline{X}(\omega)V(\omega) \qquad \forall V \in \mathcal{F}_+$$

So the random variable $\overline{X}$ is a *version* of $\mathbb{E}_{\mathcal{F}}X$. □

To better understand this proof, set $V = \mathbb{1}_H, H \in \mathcal{F}$. We have

$$\begin{aligned}\int_\Omega \mathbb{Q}(\mathrm{d}\omega)\mathbb{1}_H(\omega) = &\mathbb{Q} = \int_H X(\omega)\mathbb{P}(\mathrm{d}\omega) \\ &= \int_H \overline{X}(\omega)\mathbb{P}(\mathrm{d}\omega).\end{aligned}$$

6.3 Properties of conditional expectation

. We assume positivity and/or integrability of the random variables involved.

1. **monotonicity**:

$$X \leqslant Y \implies \mathbb{E}_{\mathcal{F}}X \leqslant \mathbb{E}_{\mathcal{F}}Y;$$

2. **linearity**:
$$\mathbb{E}_{\mathcal{F}}(aX + bY + x) = a\mathbb{E}_{\mathcal{F}}X + b\mathbb{E}_{\mathcal{F}}Y + c;$$

3. **monotone convergence theorem**:
$$(X_n)_n \text{ s.t. } X_n \geqslant 0 \forall\, n, X_n \nearrow X \implies \mathbb{E}_{\mathcal{F}}X_n \nearrow \mathbb{E}_{\mathcal{F}}X;$$

4. **Fatou's lemma**:
$$X \geqslant 0 \implies \mathbb{E}_{\mathcal{F}} \liminf X_n \leqslant \mathbb{E}_{\mathcal{F}}X_n;$$

5. **Dominated convergence theorem**:
$$(X_n)_n \text{ a.s. } X_n \to X, |X_n| \leqslant Y, Y \text{ integrable} \implies \mathbb{E}_{\mathcal{F}}X_n \to \mathbb{E}_{\mathcal{F}}X;$$

6. **Jensen's inequality**:
$$f \text{ convex} \implies \mathbb{E}_{\mathcal{F}}f(x) \leqslant f(\mathbb{E}_{\mathcal{F}}X).$$

Figure 15: The name Fatou is hard as fuck, if you ask me. Apparently he teamed up with a dude with no nose, look it up.

Remember that these results hold only almost surely, not up to equivalence!

Theorem 6.3

Let $\mathcal{F}, \mathcal{G}$ be two sub-σ-algebras of $\mathcal{H}$. Let W and X be $\mathcal{H}$-measurable random variables such that $\mathbb{E}X, \mathbb{E}WX$ exist. Then

a) $W \in \mathcal{F} \implies \mathbb{E}_{\mathcal{F}}WX = W\mathbb{E}_{\mathcal{F}}X$. This is called **conditional determinism**: we can basically treat W as a constant!

b) $\mathcal{F} \subset \mathcal{G} \implies \mathbb{E}_{\mathcal{F}}\mathbb{E}_{\mathcal{G}}X = \mathbb{E}_{\mathcal{G}}\mathbb{E}_{\mathcal{F}}X = \mathbb{E}_{\mathcal{F}}X$. So the smaller σ-algebra "wins" over the larger: this is called **repeated conditioning**.

Proof

a) Let $X \in \mathcal{H}_+, W \in \mathcal{F}_+$. Then we already know that $\overline{X} = \mathbb{E}_{\mathcal{F}} \in \mathcal{F}_+$. Consider now
$$\mathbb{E}V(WX) = \mathbb{E}\underbrace{(VW)}_{\in\mathcal{F}_+}X \quad \underbrace{= \mathbb{E}(VW)\overline{X}}_{\text{by projection property}} \quad = \mathbb{E}V(W\overline{X}) \quad \forall V \in \mathcal{F}_+$$

and hence $W\overline{X} = W\mathbb{E}_{\mathcal{F}}X$ is a version of $\mathbb{E}_{\mathcal{F}}(WX)$.

b) Let $\mathcal{F} \subset \mathcal{G}, X \in \mathcal{H}_+$. By definition we know that $E = \mathbb{E}_{\mathcal{F}}X \in \mathcal{F}_+$ so $W \in \mathcal{G}_+$. But due to conditional determinism we have that $\mathbb{E}_{\mathcal{G}}W = W$. Hence

$$\mathbb{E}_{\mathcal{G}}\mathbb{E}_{\mathcal{F}}X = \mathbb{E}_{\mathcal{F}}X.$$

Now we have to prove that

$$\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\mathcal{G}}X = \mathbb{E}_{\mathcal{F}}X.$$

Call $\mathbb{E}_{\mathcal{G}}X = Y$ and $\mathbb{E}_{\mathcal{F}}X = \overline{X}$, i.e.

$$\mathbb{E}_{\mathcal{F}}Y = \overline{X}.$$

Clearly $\overline{X} \in \mathcal{F}_+$ (✓measurability) and we still have to prove projection property. By the definition of $\overline{X}$ (as a version of $\mathbb{E}_{\mathcal{F}}X$) we have that

$$\mathbb{E}V\overline{X} = \mathbb{E}VX \qquad \forall V \in \mathcal{F}_+.$$

Recall that $Y = \mathbb{E}_{\mathcal{G}}X \in \mathcal{G}_+$ and nodice that V is also $\mathcal{G}$-measurable since $V \in \mathcal{F}_+ \subset \mathcal{G}_+$. For Y we have that $\mathbb{E}VY = \mathbb{E}VX$ for every $V \in \mathcal{G}_+$ and especially for those that are also $\mathcal{F}_+$-measurable. Hence

$$\mathbb{E}V\overline{X} = \mathbb{E}VY$$

that is to say $\overline{X}$ is also a version of the conditional epectation $\mathbb{E}_{\mathcal{F}}Y = \mathbb{E}_f\left[\mathbb{E}_{\mathcal{G}}X\right]$. □

6.4 Conditioning as projection

Note the notation!

Consider the space L^2. If we are talking about $\mathcal{H}$-measurable functions that belong to L^2 we write

$$L^2(\mathcal{H}) \ni X$$

And if we have a sub-σ-algebra $\mathcal{F}$ of $\mathcal{H}$ we will know that $L^2(\mathcal{F}) \subset L^2(\mathcal{H})$.

Theorem 6.4

$\forall X \in L^2(\mathcal{H})$ there exists a unique (up to equivalence) $\overline{X} \in L^2(\mathcal{H})$ such that

$$\mathbb{E}|X - \overline{X}|^2 = \inf_{Y \in L^2(\mathcal{F})} \mathbb{E}|X - Y|^2.$$

Furthermore, $X - \overline{X}$ is orthogonal to $L^2(\mathcal{F})$, i.e.

$$\mathbb{E}V(X - \overline{X}) = 0 \qquad \forall V \in L^2(\mathcal{F})$$

Note that $L^2(\mathcal{H})$ is a complete Hilbert space in which the inner product of X and Y is given by $\mathbb{E}XY$. $\overline{X}$ is the **orthogonal projection of the vector X** onto the subspace $L^2(\mathcal{F})$ and the decomposition

$$X = \overline{X} + \widetilde{X}$$

holds.

Proof

Let's write the L^2-norm of X calling it $\|X\|$:

$$\|X\| = \|X\|_2 = \sqrt{\mathbb{E}X^2}.$$

Fix $X \in L^2(\mathcal{H})$. Define

$$\delta = \inf_{Y \in L^2(\mathcal{F})} \|X - Y\|.$$

Let $(Y_n)_n \subset L^2(\mathcal{F})$ such that $\delta_n = \|X - Y_n\| \xrightarrow[x\to\infty]{} 0$. Let us prove that $(Y_n)_n$ is a Cauchy sequence for the $L^2(\mathcal{F})$-convergence.

$$|Y_n - Y_m|^2 = 2|X - Y_m|^2 - 4|x - \underbrace{\frac{1}{2}(Y_n + Y_m)}_{\in L^2(\mathcal{F})}|^2.$$

Take the expectation on both sides:

$$\mathbb{E}|Y_n - Y_m|^2 \leqslant 2\delta_m^2 + 2\delta_n^2 - 4\delta^2.$$

Now we take the limit for n and m and what we get is

$$\lim_{m,n\to\infty} \mathbb{E}|Y_n - Y_m|^2 \leqslant 0.$$

Hence it is true that $(Y_n)_n$ is Cauchy and this means that there exists a $\overline{X} \in L^2(\mathcal{F})$ such that $\|Y_n - \overline{X}\| \xrightarrow[n\to\infty]{} 0$. Note that $\overline{X}$ is unique up to equivalence (by definition of L^2-norm). Note also

$$\overline{X} \in L^2(\mathcal{F}) \implies \|X - \overline{X}\| \geqslant \delta.$$

Now, by Minkowski's inequality we can write that

$$\|X - \overline{X}\| \leqslant \|X - Y_n\| + \|Y_n - \overline{X}\| \xrightarrow[n\to\infty]{} \delta + 0 = \delta.$$

We have thus

$$\|X - \overline{X}\| = \delta.$$

For $V \in L^2(\mathcal{F})$ and $a \in \mathbb{R}$, since $\mathbb{E}|X - \overline{X}|^2 = \delta$ then we have that

$$\begin{aligned} a^2\mathbb{E}V^2 - 2a\mathbb{E}V(X - \overline{X}) + \delta^2 &= \|aV - (X - \overline{X})\|^2 \\ &= \|X - (\underbrace{aV + \overline{X}}_{\in L^2(\mathcal{F})})\|^2 = \delta^2 \end{aligned}$$

And therefore

$$a^2\mathbb{E}V^2 - 2a\mathbb{E}V(X - \overline{X}) \leqslant 0 \qquad \forall\, a \in \mathbb{R}$$

which is impossible unless

$$\mathbb{E}V(X - \overline{X}) = 0.$$

□

6.5 Conditional expectations given random variables

Consider Y, a random variable on a measurable space. We have σY that denotes the σ-algebra generated by Y. But σY is also the set of numerical random variables of the form $f \circ Y$ for a deterministic function f.

Definition 6.3

The conditional expectation of X given Y is

$$\mathbb{E}_{\sigma Y} X$$

If $(Y_t)_{t\in T}$ is a collection of random variables taking values in some measurable space. the conditional expectation of X given $(Y_t)_{t\in T}$ is

$$\mathbb{E}_{\mathcal{F}} X, \quad \text{where } \mathcal{F} = \sigma\left((Y_t)_{t\in T}\right).$$

Theorem 6.5

Let $X \in \mathcal{H}_+$. Let Y be a random variable taking values in the measurable space $(E, \mathcal{E})$. Then every version of $\mathbb{E}_{\sigma Y} X$ has the form $f \circ Y$ for some $f \in \mathcal{E}_+$.
Conversely, $f \circ Y$ is a version of $\mathbb{E}_{\sigma Y} X$ if and only if

$$\mathbb{E} f \circ Y \cdot h \circ Y = \mathbb{E} X \cdot h \circ Y \qquad \forall h \in \mathcal{E}_+.$$

Actually the last equality is just the projection property, if we think about it hard enough. The notation is once more problematic, though.

Note the notation!

Çinlar has a specific notation for conditional expectation, which is $\mathbb{E}_{\sigma Y} X$. He is stupid and autistic because the whole world uses $\mathbb{E}(X|Y)$. Professor Polito says that "Oh well but $\mathbb{E}_{\sigma Y} X$ is more precise b-because it's more rigorous" well SHUT UP. SHUT UP NO ONE USES IT NO ONE EVEN KNOWS WHAT A FUCKING σ-ALGEBRA EVEN IS AND YOU KNOW WHY? BECAUSE IT IS USELESS BECAUSE ALL OF THIS SHIT IS USELESS IT IS A WASTE OF TIME AND IT IS NOT EVEN REMOTELY INTERESTING. LIKE I HAVE A FRIEND WHO STUDIES GEOLOGY THEY *LITERALLY STUDY FUCKING ROCKS* THEY STUDY ALL THE MINUTE CHEMICALS ON THE ROCKS YOU FIND ON THE GROUND THEY STUDY HOW MANY FUCKING DINOSAURS SHITTED AND PISSED AND CUMMED ON THAT FUCKING ROCK AFTER IT GOT EJECTED BY A STUPID VOLCANO A MILLION YEARS AGO AND IT IS MORE INTERESTING AND USEFUL THAN THIS LIKE IT IS LITERALLY MORE USEFUL KNOWLEDGE IN ANY REAL DATA SCIENCE JOB WHERE ANYONE LITERALLY SAYS $\mathbb{E}(X|Y)$ BECAUSE THAT'S WHAT YOU NEED TO KNOW YOU ONLY NEED TO KNOW THAT WE KNOW THAT Y HAPPENED WHEN WE EVALUATE THE PROBABILITY OF X YOU DON'T NEED TO KNOW ANYTHING ELSE I DON'T CARE AND NO ONE CARES.

The two notations are interchangeable but $\mathbb{E}_{\sigma Y} X$ is preferrable.

6.6 Conditional probabilities and distributions

Definition 6.4

For every event $H \in \mathcal{H}$

$$\mathbb{P}_{\mathcal{F}}(H) = \mathbb{E}_{\mathcal{F}} \mathbb{1}_H$$

is called **conditional probability of H given $\mathcal{F}$** (with $\mathcal{F}$ beinf a sub-σ-algebra of $\mathcal{H}$).

Lol what the fuck? Why are we defining probability through expectation? I don't know but remember that conditional expectation is a full-fledged random variable!

Definition 6.5

Consider two events, G and H. The conditional probability of H given G is the number $\mathbb{P}_G(H) \in [0, 1]$ satisfying

$$\mathbb{P}(G \cup H) = \mathbb{P}(G)\mathbb{P}_G(H)$$

that is unique if $\mathbb{P}(G) > 0$.

So in this case is not a random variable anymore but only a number.
Let us consider the conditional probability $\mathbb{P}_{\mathcal{F}} H \forall\, H \in \mathcal{H}$ (we will remove parentheses from now on because you suck and you should die). Let $Q(H)$ be a version of $\mathbb{P}_{\mathcal{F}} H$ (remember... conditional probabilities are now expectations!). Assume $Q(\varnothing) = 0$ and $Q(\Omega) = 1$.

$Q(H)$ is a random variable and it is $\mathcal{F}$-measurable by definition. Let $Q_\omega(H)$ be the value at point $\omega \in \Omega$.

Note the notation!

If we called $Q(H) = Z$ then $Q_\omega(H)$ would be $Z(\omega)$.

So what is Q actually? It is a function $Q : (\omega, H) \mapsto Q_\omega(H)$ whith $\omega \in (\Omega, \mathcal{F})$ (a subset of $(\Omega, \mathcal{H})$) and $H \in (\Omega, \mathcal{H})$. This is similar to a transition kernel... A transition kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$. But it is really? Do we care? Of course not, but we need to check because we need more reasons to end our life by our hand. For something to be a transition kernel it should verify:

1. the mapping $\omega \mapsto Q_\omega(H)$ should be $\mathcal{F}$-measurable (in this case it is ok for every $H \in \mathcal{H}\checkmark$);
2. for every disjoint sequence $(H_n)_n \in \mathcal{H}$ it should hold $\forall\, \omega \in \Omega$ the following:

$$Q_\omega\left(\bigcup_n H_n\right) = \sum_n Q_\omega(H_n). \tag{$\star$}$$

This last requirement arises a problem: we could apply monotone convergence theorem for conditional expectations, but we know that those property hold only almost surely! I can already hear the reader say "oh but why should we care isn't it supposed to be like the same of saying "equal to" and things like this and so on" but well, my dear idiotic little shit, when something is almost surely it is actually *almost* surely. It means that you're *almost* guaranteed that something will not fuck you in the ass unless something is small enough to creep in your asshole whithout you even noticing. Thank measure theory enthusiasts for this.

Figure 16: Snake has had enough and honestly so do I.

The fact is that we could have some ω's with measure zero so technically there are some exception points for which it doesn't hold. These points depends on the choice of the sequence H_n. Let Ω_0 be the set for which ($\star$) holds:

$$\Omega_0 = \bigcap_h \Omega_h$$

where Ω_h is the almost surely set including all ω for the sequence $h = (H_n)_n$. But we may have uncountably many problems... Why are there so many problems with conditional expectation? The root of the problem is that conditional expectation is not an exactly well-defined object.

Definition 6.6

Let $Q(H)$ be a version of $\mathbb{P}_{\mathcal{F}}(H), \forall\, H \in \mathcal{H}$. Then $Q : (\omega, H) \mapsto Q_\omega(H)$ is said to be a **regular version** of the conditional probability $\mathbb{P}_{\mathcal{F}}$ provided that Q is a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$.

Proposition 6.2

Let $\mathbb{P}_{\mathcal{F}}$ have a regular version Q. Then

$$QX : \omega \mapsto Q_\omega X = \int_\Omega Q_\omega(\mathrm{d}\omega')X(\omega')$$

is a version of $\mathbb{E}_{\mathcal{F}}X \forall$ random variable X whose expectation exists.

So we are integrating over Ω our function X against our regualar version of the conditional probability Q.

Theorem 6.6

If $(\Omega, \mathcal{H})$ is a standard measurable space (which means that it is isomorphic to $(F, \mathcal{B}_F)$ for some Borel subset F of $\mathbb{R}$). Then $\mathbb{P}_{\mathcal{F}}$ has a regular version.

List of Figures