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).

[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.

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 $\mathcal{E}$.

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 $\mathcal{E}$.

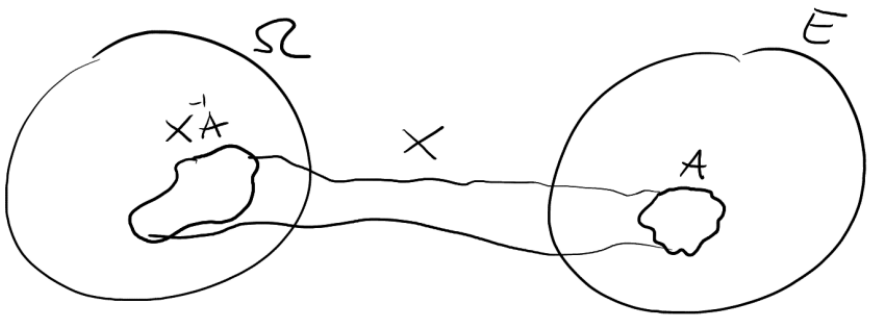


Figure 1: this is an early reminder of the fact that I will take my own life very soon.

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

[2] who asked

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

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

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

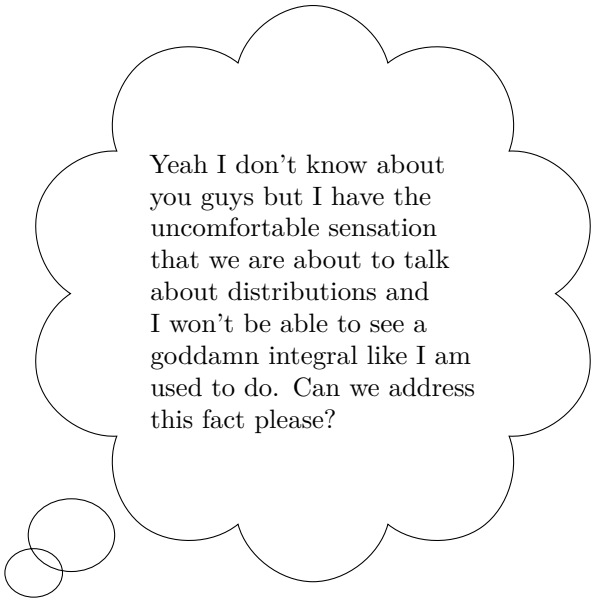
Remark

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...



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

[4] 

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

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 $(E, \mathcal{E})$:

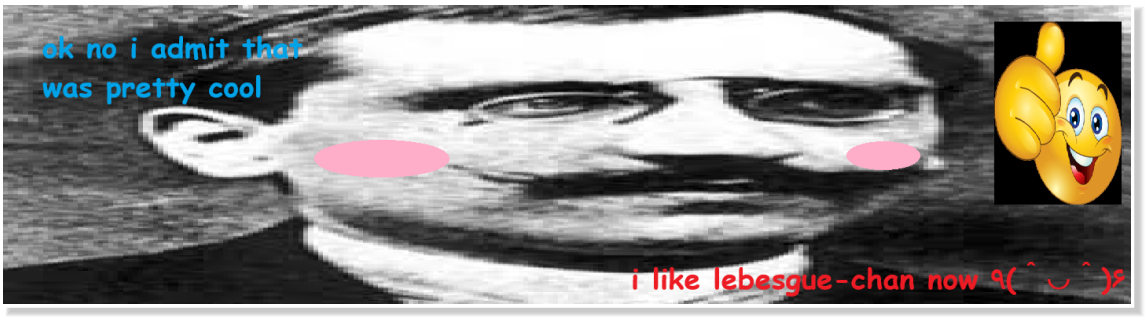
$$\mu f = \int_E f(x)\mu(\mathrm{d}x) \qquad \text{with } f \in \mathcal{E}$$

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:

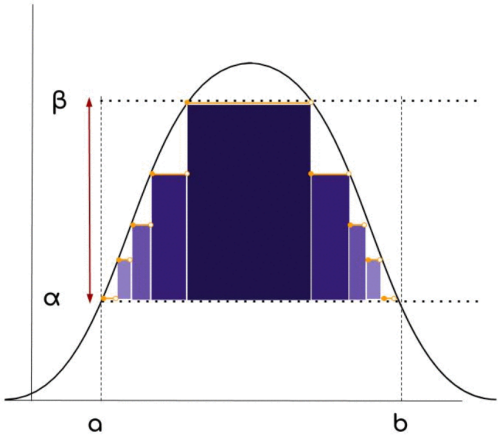


Figure 2: Imagine the steps getting smaller and smaller...

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*:

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

[6] This is only the first chapter.

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]!).

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

[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.

[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

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

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?

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

[11] Abandoning myself in the sweet embrace of Death, methinks.
[12] No.
[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.

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

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

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

[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.

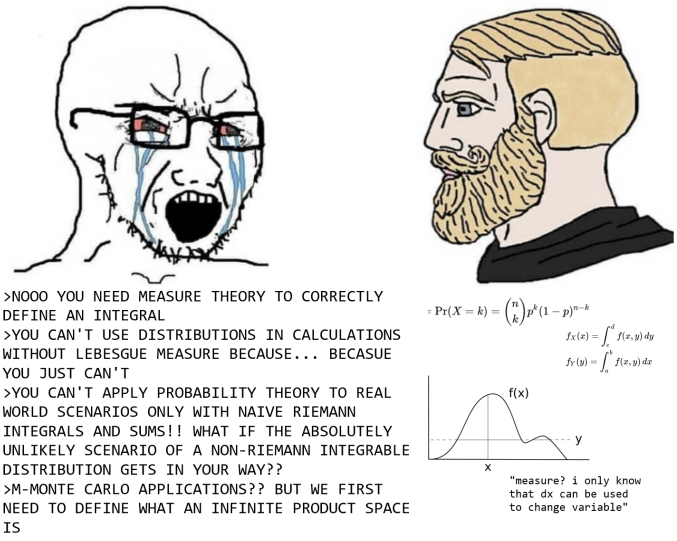


Figure 3: I'm sorry but here you're the soyjack and I'm the chad.

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.

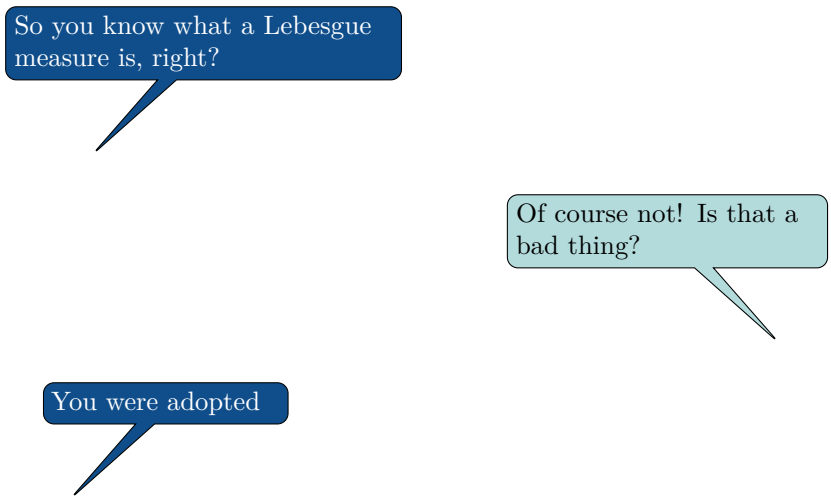


Figure 4: Actual conversation happened between me and Professor Lods.

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

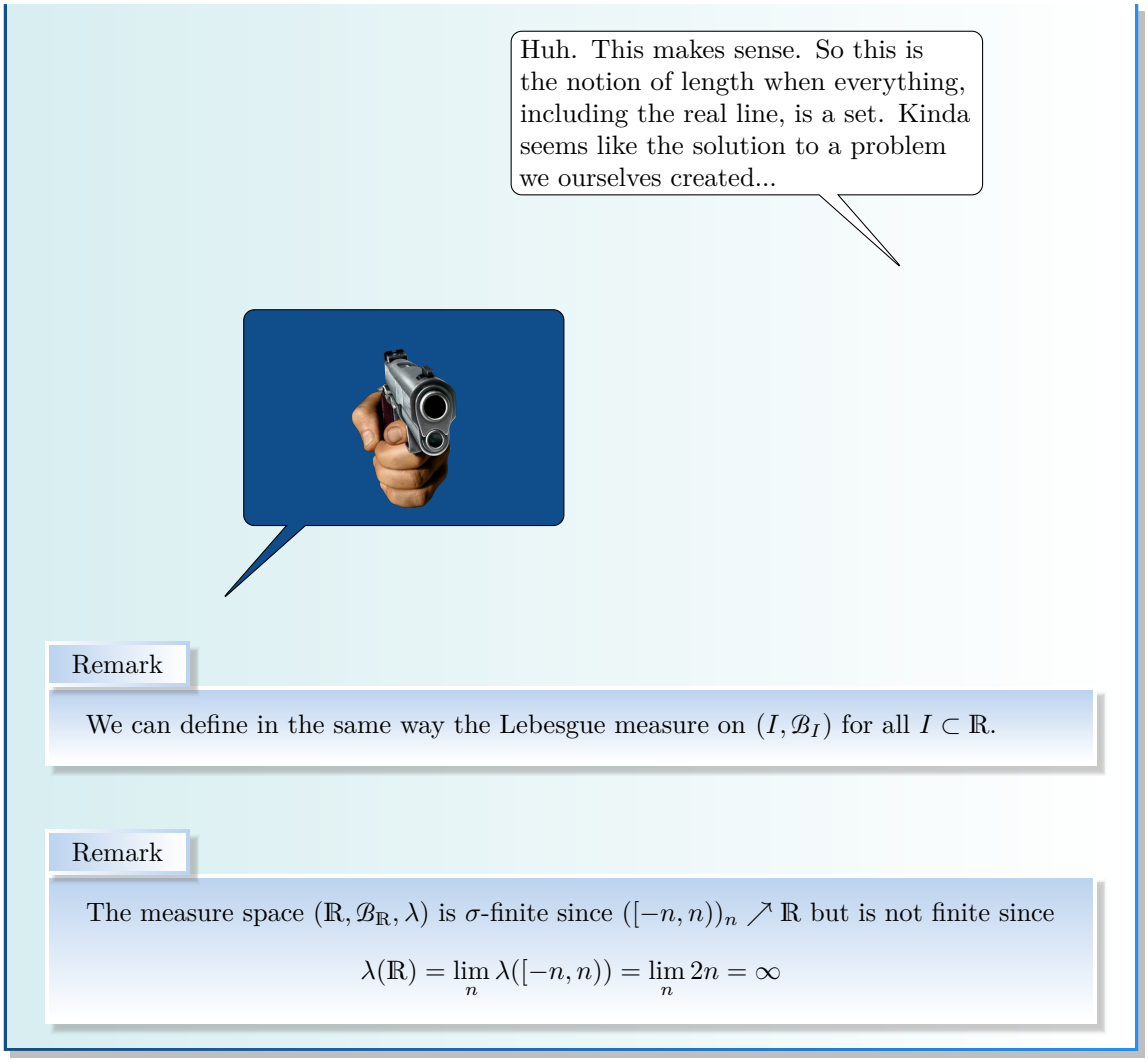
There exists a unique measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that we denote λ (or $\mathfrak{m}$) and such that

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

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.



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:

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

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

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

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

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:

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

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].

[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.

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

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

[26] This is not, in fact, the last Metal Gear Solid joke.
[27] OK I GET 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 $Kf \in \mathcal{E}_+ \ \forall\, f \in \mathcal{F}_+$.

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.

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:

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

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.

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

2.1 Products of kernels

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

Definition 2.3

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

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

Definition 2.4

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

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

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

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.

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 2.3

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

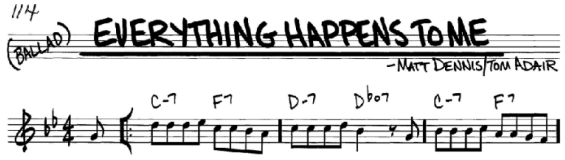
$$\pi f = \int_E \mu(\mathrm{d}x) \int_F K(x, \mathrm{d}y) f(x, y)$$

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?



29

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

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

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

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

[30] I crave to be released from this prison of flesh.
[31] I have an idea.
[32] Why. Stop.
[33] We have a great time ahead, I see.

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.

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 finite 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**.

[34] No. Happy?

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

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

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

- **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
$$(X_n)_n \implies \mathbb{E}\liminf X_n \leqslant \liminf \mathbb{E}X)n.$$
- **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
$$\mathbb{E}\lim_n X_n = \lim_n \mathbb{E}X_n.$$
- **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:

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

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

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

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

Remark

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

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

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

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**

[36] Proceeds *not* to do it.
[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.

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!

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

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

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

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**:

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

- **Hölder inequality**:

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

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

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

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

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
$$k(b) = \sup_{X\in K} \big(\mathbb{E}|X|\mathbb{1}_{(|X|>b)}\big)$$
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 is is uniformly integrable. 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