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.

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.

[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.

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 $\mathcal{E}$.

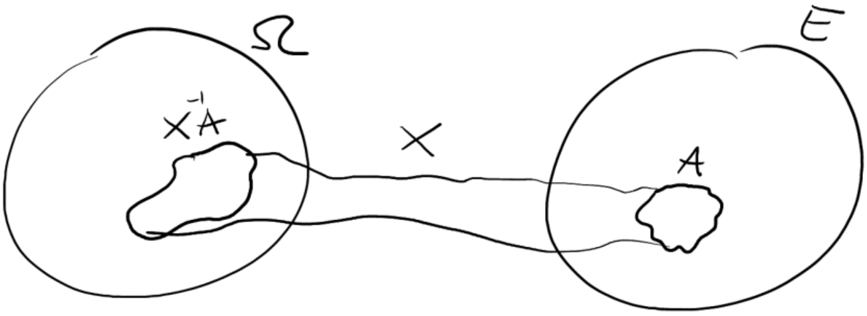


Figure 1: this is an early reminder of the fact that I will take my own life very soon.

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 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 $\mathcal{B}_{\mathbb{R}^d}$.

[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

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...

Yeah I don't know about you guys but I have the uncomfortable sensation that we are about to talk about distributions and I won't be able to see a goddamn integral like I am used to do. Can we address this fact please?

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 $i = 1, 2, \dots, n$.

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 do multiply the height of the function $f(x)$ by calculating the "weight", or "measure" of the infinitesimal part of the domain of the function $\mathrm{d}x$ according to our method of measure of choice. We do this over the set E.

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

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...

Anyway was it really SO DIFFICULT to explain?

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 $f \in \mathcal{M}$.

[6] This is only the first chapter.

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 $\mathcal{M}$ includeds 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]!).

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].

[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

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 $(X_1, X_2, \ldots, X_n)$ 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 $i = 1, \ldots, n$.

Cynlar 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...

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

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

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

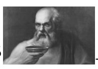
or

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

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

[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]  $\rightarrow$ 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.

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.

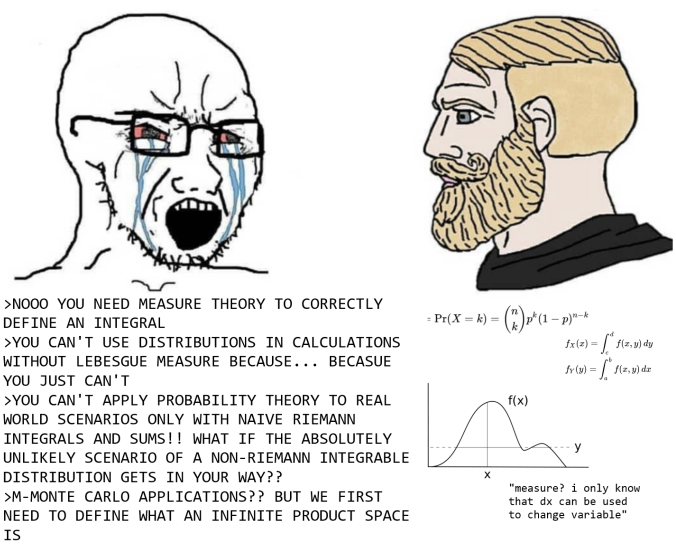


Figure 2: I'm sorry but here you're the soyjack and I'm the chad.

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 dx 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 trackbacking.

Revise with Kotatsu!

First of all: what exactly *means* to be absolutely continuous?

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 lose track of your own identity.

Now everything should make more sense[b]. 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.

[a]tacci tua.
[b]Enviable optimism.

It is interesting[19] 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

[19]Debatable claim.

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.)

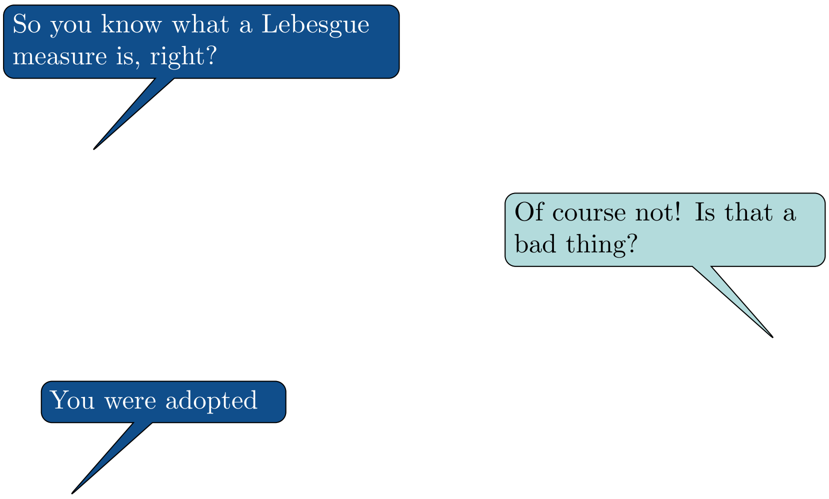


Figure 3: Actual conversation happened between me and Professor Lods.

So we're due for a little refresh on what the hell a Lebesgue measure is. I'm sorry[20] 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

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

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**.

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 $\{I_j\}_{j=1,\ldots,N} \subset \mathcal{C}_0$ 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

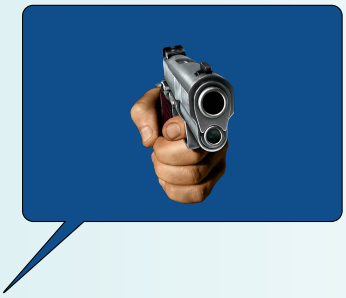
$$\lambda([a, b)) = b - a \qquad \forall\, a < b.$$

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).
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[21].

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:

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

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*

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

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):

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

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:

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

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 certaninly a measure, but what about $K(x, \mathrm{d}y)$?
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**[a]

[a] Kernel? Colonel? I thought we were over with the Metal Gear Solid jokes.

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