

# 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

<b>1 Basics of probability</b>	<b>2</b>
1.1 Random variables	3
1.2 Functions of random variables	4
1.3 Infinite product spaces	6
1.4 Stochastic processes	7
1.5 Example of random variables	8

## 1 Basics of probability

We start with the probability triplet:  $(\Omega, \mathcal{H}, \mathbb{P})$  Here  $\Omega$  is the set of sample space,  $\mathcal{H}$  is the  $\sigma$ -algebra built upon  $\Omega$  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 consider the elements  $A \in \mathcal{H}$  (which will be subsets of  $\Omega$ ) 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  $\sigma$ -algebra. The probability measure has the following properties:

- $\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\emptyset) = 0$
- **monotonicity of  $\mathbb{P}$** : take 2 events  $H, K \in \mathcal{H}$  such that  $H \subset K$ . Then  $\mathbb{P}(H) \leq \mathbb{P}(K)$ <sup>1</sup>.
- **finite additivity**: take  $H, K \in \mathcal{H}$  such that  $H \cap K = \emptyset$ . Then  $\mathbb{P}(H \cup K) = \mathbb{P}(H) + \mathbb{P}(K)$ ;
- **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  $\mathbb{N} = \{0, 1, 2, 3, \dots\}$  and  $\mathbb{N}^* = \{1, 2, 3, 4, \dots\}$  such that they are disjoint pairwise (except identical pairs). Then

$$\mathbb{P}\left(\bigcup_n H_n\right) = \sum_n \mathbb{P}(H_n)$$

- **Boole inequality (sub-additivity)**: if we have a collection  $(H_n)_{n \in \mathbb{N}} \subset \mathcal{H}$  (not necessarily disjoint) then

$$\mathbb{P}\left(\bigcup_n H_n\right) \leq \sum_n \mathbb{P}(H_n)$$

- **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**.

<sup>1</sup>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.

## 1.1 Random variables

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

### Definition 1.1

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

What does it mean<sup>2</sup>? The inverse image of the set  $A$  through  $X$  ( $X^{-1}A$ ) with  $A \in \mathcal{E}$  is actually the set of the  $\omega$ s 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  $\mathcal{E}$ .

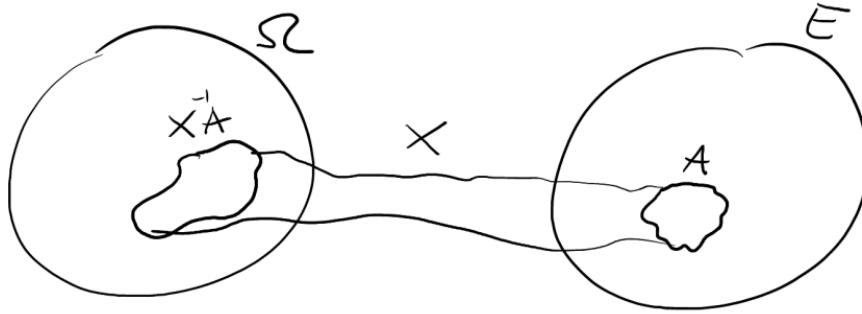


Figure 1: this is an early reminder of the fact that I will take my own life very soon.

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

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

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  $\mathcal{H}$ <sup>3</sup>). If something is not in  $\mathcal{H}$  get it off my fucking face man and kill yourself NOW<sup>4</sup>. 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

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

### Remark

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

<sup>2</sup>who asked

<sup>3</sup>il lettore più arguto avrà notato che, a questo punto, il dio è ormai irrimediabilmente cane.

<sup>4</sup>

<sup>5</sup>Borel  $\sigma$ -algebra. You don't know what a Borel  $\sigma$ -algebra is? [https://en.wikipedia.org/wiki/Borel\\_set](https://en.wikipedia.org/wiki/Borel_set)

### Remark

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

### Definition 1.2

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

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

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

<sup>a</sup>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 evaluate these 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!

### Remark

You should remember (LOL) that when we want to specify a measure on a  $\sigma$ -algebra, it's enough to do it on a *p-system*<sup>a</sup> generating that  $\sigma$ -algebra: by means of the monotone class theorem we are then able to extend the measure to the  $\sigma$ -algebra.

This means that to specify  $\mu$  it is enough to specify it on a *p-system* which generates  $\mathcal{E}$ . For example, consider  $E = \mathbb{R}, \mathcal{E} = \mathcal{B}_{\mathbb{R}}$ . Consider the collection of sets  $[-\infty, x], x \in \mathbb{R}$  which is of course a p-system because it is closed under intersection. Moreover, this shit generates the Borel sigma algebra on  $\mathbb{R}$ .

If we want to define a distribution, that is a measure, it is enough to define it on this p-system. Imagine that we apply our distribution measure to one set of this p-system

$$c(x)^b = \mu([-\infty, x]) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

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

<sup>a</sup>a p-system is a simpler object than a  $\sigma$ -algebra: it is simply a collection of sets closed under intersection

<sup>b</sup>because it is a function of  $x$

<sup>c</sup>I already wanted to kill myself at that time.

<sup>d</sup>I have noticed that my life has not benefited in ANY form since I have been introduced to measure theory.

## 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  $f : E \rightarrow F$  be a measurable function relative to  $\mathcal{E}$  and  $\mathcal{F}$ <sup>6</sup>. This function should be measurable by means of  $\mathbb{P}$ , otherwise we couldn't do anything useful with it. Consider the composition

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

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 1.3

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

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

<sup>6</sup>This basically means that this bitch won't do anything evil. The whole point of measure theory,  $\sigma$  algebras and all other shit is to ensure everything behaves.

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  $\sigma$ -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  $\sigma$ -algebra  $\mathcal{E} \otimes \mathcal{F}$ , that is the product  $\sigma$ -algebra generated 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  $\sigma$ -algebra (but it may be generated by other  $\sigma$ -algebras<sup>7</sup>!).

Definition 1.4

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<sup>8</sup>.

Remark

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

On the product space, it is enough to only specify it on this p-system.

Let denote with  $\pi$  the joint distribution of  $X, y$ . It is sufficient to specify

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

We exploited the measurability of  $X$  and  $Y$

Definition 1.5

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

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

and we call **marginal distribution of  $Y$**

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

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

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

and

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

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<sup>9</sup>?

Definition 1.6

Let  $X, Y$  be random variables taking values in  $(E, \mathcal{E})$  and  $(F, \mathcal{F})$  respectively and let  $\mu$  and  $\nu$  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.

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

<sup>7</sup>Repeatedly inflicting painful kicks on my gonads.

<sup>8</sup>it eludes me how anyone could find this interesting. We have to think about the whole vector as being distributed like its components separately

<sup>9</sup>Abandoning myself in the sweet embrace of Death, methinks.

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<sup>10</sup> something more general.

Definition 1.7

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

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

Definition 1.8

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 independency.

We now turn to stochastic processes<sup>11</sup>! 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<sup>12</sup>  $(x_t)_{t \in T}$ . 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  $\sigma$ -algebra...

A **rectangle** 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 say 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<sup>13</sup>.

The  $\sigma$ -algebra generated by the collection of all measurable rectangles is denoted by

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

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

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

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<sup>14</sup>, is not restricted on all the

<sup>10</sup>No.

<sup>11</sup>Please no.

<sup>12</sup>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.



<sup>13</sup>→ my honest reaction.

<sup>14</sup>Oh thank god someone finally said it. I was starting to get scared.

coordinates (like the shape<sup>15</sup>) 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  $\sigma$ -algebra built upon the p-system which is the collection of all possible rectangle defined in the way we saw above<sup>16</sup>.

## 1.4 Stochastic processes

### Definition 1.9

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  $\omega$ . 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  $\omega$  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.

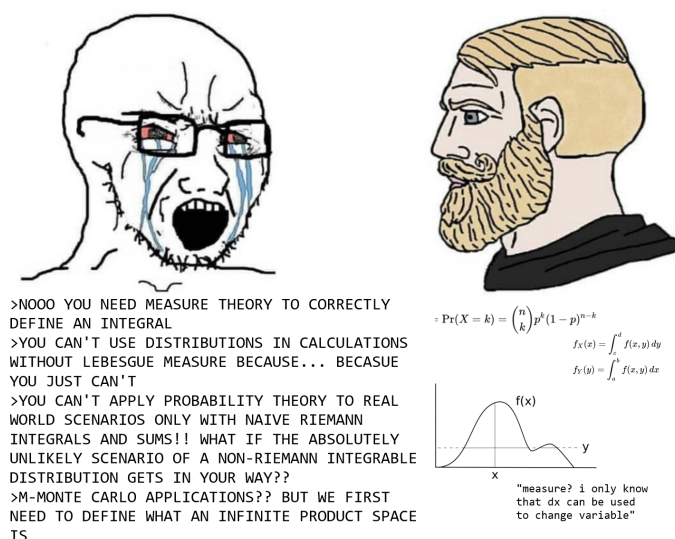


Figure 2: I'm sorry but here you're the soyjack and I'm the chad.

<sup>15</sup>I swear to god.

<sup>16</sup>NO I WON'T USE LABELS AND NUMBERED EQUATIONS.



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 values in  $\mathbb{N}$  (it's a one dimensional random variable). We consider the power set of  $\mathbb{N}$ <sup>a</sup>. We know that power sets *are*  $\sigma$ -algebras that we can use (but we could encounter some trouble with uncountable elements, for which we would need smaller  $\sigma$ -algebras<sup>b</sup>).

What is the distribution of this random variable?

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

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

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 ( $\mathbb{1}_A(n)$ ).

<sup>a</sup>Subset of all subsets of  $\mathbb{N}$ .

<sup>b</sup>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<sup>a</sup>. 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

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

WHY ON EARTH did we write  $dx$  instead of just  $x$ ? That's a good question.  $dx$  stands for the Lebesgue measure of  $dx$ <sup>b</sup>. The second part of the equation ( $ce^{-cx}$ ) is the density part and, being a continuous function, it implies that the distribution is continuous with respect to the Lebesgue measure<sup>c</sup>. So  $p(x) = ce^{-cx}$ ,  $d \in \mathbb{R}_+$  is the density relative to  $\mu$ .

<sup>a</sup>tacci tua.

<sup>b</sup>Why? I will kill my own self.

<sup>c</sup>Really??

It is interesting<sup>17</sup> 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.

---

<sup>17</sup>Debatable claim.