Measure Theory and Probability Theory

Stéphane Dupraz

In this chapter, we aim at building a theory of probabilities that extends to any set the theory of probability we have for finite sets (with which you are assumed to be familiar). For a finite set with N elements $\Omega = \{\omega_1, ..., \omega_N\}$, a probability $\mathbb{P}$ takes any n positive numbers $p_1, ..., p_N$ that sum to one, and attributes to any subset S of Ω the number $\mathbb{P}(S) = \sum_{i/\omega_i \in S} p_i$. Extending this definition to infinitely countable sets such as $\mathbb{N}$ poses no difficulty: we can in the same way assign a positive number to each integer $n \in \mathbb{N}$ and require that $\sum_{n=1}^{\infty} p_n = 1$.[1] We can then define the probability of a subset $S \subseteq \mathbb{N}$ as $\mathbb{P}(S) = \sum_{n \in S} p_n$.

Things get more complicated when we move to uncountable sets such as the real line $\mathbb{R}$. To be sure, it is possible to assign a positive number to each real number. But how to get from these positive numbers to the probability of any subset of $\mathbb{R}$?[2] To get a definition of a probability that applies without a hitch to uncountable sets, we give in the strategy we used for finite and countable sets and start from scratch.

The definition of a probability we are going to use was borrowed from measure theory by Kolmogorov in 1933, which explains the title of this chapter. What do probabilities have to do with measurement? Simple: assigning a probability to an event *is* measuring the likeliness of this event. What we mean by *likeliness* actually does not matter much for the mathematics of probabilities, and various interpretations can be used: the objective fraction of times an event occurs if we repeat some experiment an infinity of times, my subjective belief about the outcome of the experiment, etc. From a mathematical perspective, what matters is that a probability is just a particular case of a *measure*, and the mathematical theory of probabilities will at first be quite indifferent to our craving to apply it to the measurement of hazard.

Although our main interest in measure theory is its application to probability theory, we will also be concerned with one other application: the definition of the Lebesgue measure on $\mathbb{R}$ (and $\mathbb{R}^n$), meant to correspond for an interval $[a, b]$ to its length $b - a$.

[1] The infinite sum is to be understood as the limit of the sequence $\sum_{n=1}^{N} p_n$. There is a subtlety because is not obvious that the limit is the same regardless of the ordering of the sequence, but it turns out that the invariance is guaranteed because the p_n are positive.

[2] You may be thinking of using an integral, but we have not defined integrals in the class yet—we will do so in this very chapter *after* defining probabilities on $\mathbb{R}$. There is more to this than my poor organization of chapters: the integral that we could have built before starting this chapter is the *Riemann integral*. But the Riemann integral is only defined on intervals of $\mathbb{R}$, so that it could only have helped us defining the probability of intervals. The integral we will build in this chapter *from the definition of a measure* is the *Lebesgue integral*, a considerable extension of the Riemann integral.

1 Measures, probabilities, and sigma-algebras

We are looking for a notion to tell how big a set is, and whether it is bigger than another set. Note that we already have one such notion: cardinalities. Indeed, for finite sets, we can compare the size of sets through the number of elements they contain, and the notion of countability and uncountability extends the notion to comparing the "size of infinite sets".

Cardinalities have two limitations however. First, it restricts to a specific way of measuring the size of a subset. Instead, maybe we want to put different weights on different elements, for instance in the context of probabilities because some outcomes are deemed more likely than others. We are aiming at a more general notion. Second, even restricting to some "uniform" measure, as soon as we reach the uncountable infinity, cardinality makes very coarse distinctions between sizes: for instance $\mathbb{R}$ and $[0,1]$ have the same "size" according to cardinality since they are in bijection. We would like to be able to say that $\mathbb{R}$ is bigger than $[0,1]$.

So let us build a new notion, that of a measure. In essence, what we want a measure on a set Ω to do is to assign a positive number (or infinity) to subsets of Ω. Really, there is only one property that we wish to impose: that the union of two disjoint subsets be the sum of the measures of the two subsets—additivity. By induction, additivity for 2 disjoint subsets is equivalent to additivity for a finite number of pairwise disjoint subsets. What about infinite unions? Well, why not, but note that we have no clue what the sum of an uncountable infinity of positive numbers is—we have never defined such a notion. So we require additivity for pairwise disjoint *countable* collection of subsets—sigma-additivity.

Definition (provisory). *Let Ω be a set. A **measure** μ on Ω is a function defined on $\mathcal{P}(\Omega)$ which:*

1. *Takes values in $\mathbb{R}_+ \cup \{+\infty\}$, and such that $\mu(\emptyset) = 0$.*
2. *Is **sigma-additive** (**countably-additive**): for any countable collection of subsets $(A_n)_{n\in\mathbb{N}}$ of Ω that are pairwise disjoint ($A_i \cap A_j = \emptyset$ for all $i \neq j$),*

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

Now there is a reservation, which is why the definition has been labeled "provisory". For some sets, such as finite and countable sets, this definition would be quite good. But as it is, it would quickly put us in trouble when dealing with uncountable sets such as the real line $\mathbb{R}$. To understand why, and to understand the little detour that we are going to make before giving the proper definition of a measure—the definition of sigma-algebras—it is useful to consider the problem Lebesgue was trying to solve in 1902. Lebesgue was trying to extend the notion of the length of an interval to all subsets of $\mathbb{R}$. To this end, he asked whether there

exists a positive function μ on the power set of $\mathbb{R}$ that is sigma-additive—a measure according to our provisory definition, invariant by translation (meaning $\mu(S + x) = \mu(S)$ for all subset $S \subset \mathbb{R}$ and vector $x \in \mathbb{R}$), and normalized by $\mu([0, 1]) = 1$. This does not sound like asking for much, but unfortunately, in 1905, Vitali showed that there is no such function.

The way mathematicians reacted to this drawback has been to allow for a measure to be defined on only a collection of subsets smaller than the entire power set, restricting the subsets that can be measured. But not on any collection of subsets of Ω; only on collections that satisfy a few properties: sigma-algebras.

1.1 Sigma-algebras

We are willing to restrict the collection of measurable sets, but there are things on which we are not ready to negotiate. First, if a set is measurable, we want its complement to be measurable too. (In the case of a probability, if we give a probability to an event happening, we want to be able to give a probability to the event not happening). Second, since we want our measures to be sigma-additive, we want the countable union of measurable sets to be measurable. These requirements define a sigma-algebra.

Definition 1.1. *Let Ω be a non-empty set. A collection $\mathcal{A}$ of subset of Ω is a* ***sigma-algebra*** *if:*

1. $\Omega \in \mathcal{A}$.
2. *It is close under complementation:* $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.
3. *It is close under countable union:* $(A_n)_{n\in\mathbb{N}} \in \mathcal{A} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Elements of a sigma-algebra $\mathcal{A}$ are called ***measurable sets****.*
The couple $(\Omega, \mathcal{A})$ is called a ***measurable space****.*

Note that a sigma-algebra also necessarily:

- Contains the empty-set (since $\emptyset = \Omega^c$).
- Is close under countable intersection (using Morgan's law and closeness under countable union and complementation).

It is easy to build sigma-algebras. For instance, in the set $\Omega = \{1, 2, 3\}$, $\{\emptyset, \Omega\}$ is a sigma-algebra, as are $\{\emptyset, \{1\}, \{2, 3\}, \Omega\}$ and $\mathcal{P}(\Omega)$. More generally, on any set Ω, the collection $\{\Omega, \emptyset\}$ is a sigma-algebra—it is the coarsest sigma-algebra since it is the one that allows to measure the fewer subsets. Also, $\mathcal{P}(\Omega)$ is a sigma-algebra—it is the finest sigma-algebra since it allows to measure all subsets. The whole point of defining

sigma-algebras however is to end up with a collection of sets that is smaller than $\mathcal{P}(\Omega)$. On this account, be careful that only *countable* unions (intersections) of measurable sets are required to be measurable: asking for sigma-algebra to be close under any union would considerably restrict the number of sigma-algebras we can define on a set. For instance, were we to require all singletons of a set Ω to me measurable, we would fall back on $\mathcal{P}(\Omega)$.

How to generate sigma-algebras? Let us import a trick we used in linear algebra to create vector subspaces. There, we saw that given any subset S of a vector space, we can always define the vector subspace generated by S as the smallest vector subspace containing S—the intersection of all vector subspaces containing S. What allowed us to do this was that any (possibly infinite) intersection of vector subspaces is a vector subspace. It is easy to check that similarly, any (possibly infinite—possibly uncountably infinite for that matter) intersection of sigma-algebras is a sigma-algebra. So that we can define the sigma-algebra generated by any subset $\mathcal{S}$ of $\mathcal{P}(\Omega)$.

Definition 1.2. *Let $\mathcal{S}$ be a collection of subset of the set Ω.*

*The **sigma-algebra generated by** $\mathcal{S}$, noted $\sigma(\mathcal{S})$, is the smallest sigma-algebra that contains $\mathcal{S}$, or:*

$$\sigma(\mathcal{S}) = \bigcap \{\mathcal{A}, \mathcal{A} \text{ is a sigma-algebra and } \mathcal{S} \subseteq \mathcal{A}\}$$

We are now ready to define the sigma-algebra that we will use in practice to define all our measures on $\mathbb{R}$: the Borel sigma-algebra. The logic behind the definition is simple: we want our measures to be able to measure the open intervals (a, b) of $\mathbb{R}$. However, this is not a sigma-algebra—just consider the union of two disjoint open intervals—so we take the sigma-algebra generated by the open intervals of $\mathbb{R}$. It is easy to check that the sigma-algebra generated by the open intervals of $\mathbb{R}$ is equivalently the sigma-algebra generated by the open set of $\mathbb{R}$ (you are asked to check it in the problem-set), so that the definition of the Borel sigma-algebra is frequently phrased as the sigma-algebra generated by the open sets of $\mathbb{R}$.

Definition 1.3.

*The **Borel sigma-algebra on** $\mathbb{R}$, noted $\mathcal{B}(\mathbb{R})$, is the sigma-algebra generated by the open sets of $\mathbb{R}$.*

Equivalently, it is the sigma-algebra generated by the open intervals (a, b) of $\mathbb{R}$.

*The **Borel sigma-algebra on** $\mathbb{R}^n$, noted $\mathcal{B}(\mathbb{R}^n)$, is the sigma-algebra generated by the open sets of $\mathbb{R}^n$.*

Equivalently, it is the sigma-algebra generated by the sets $\prod_{i=1}^{n}(a_i, b_i)$ of $\mathbb{R}^n$.

*A measurable set of the Borel sigma-algebra is called a **Borel set**.*

(To be clear: we are referring to the open sets for the Euclidian distance in $\mathbb{R}^n$—the absolute value in $\mathbb{R}$). The open sets of $\mathbb{R}$ do not form more of a sigma-algebra than the set of finite open intervals of $\mathbb{R}$—if so, it would

need to include all closed sets, and it does not—so the Borel set is a bigger collection of sets than the collection of open sets of $\mathbb{R}$. Actually, finding a non-measurable set according to the Borel sigma-algebra is rather hard, but Vitali showed such sets exist (the counterexamples he used are now called *Vitali sets*). Simply put, the Borel sigma-algebra is quite huge but is not the whole power set of $\mathbb{R}$, so that it makes it a perfect candidate to define measures on. From now on, anytime we talk of $\mathbb{R}$ and $\mathbb{R}^n$, it is to be understood as the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

1.2 Measures

We are now ready to give the proper definition of a measure: it only generalizes the provisory definition given above to allow probabilities to be defined on a sigma-algebra of Ω that is not necessary the power set of Ω.

Definition 1.4. *Let $(\Omega, \mathcal{A})$ be a measurable set. A **measure** μ on $(\Omega, \mathcal{A})$ is a function defined on $\mathcal{A}$ which:*

1. *Takes values in $\mathbb{R}_+ \cup \{+\infty\}$, and such that $\mu(\emptyset) = 0$.*
2. *Is sigma-additive: for any countable collection $(A_n)_{n\in\mathbb{N}}$ of $\mathcal{A}$ that are pairwise disjoint:*

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

*The triple $(\Omega, \mathcal{A}, \mu)$ is called a **measure space**.*

It is easy to check that on any finite set Ω, the number of elements in a subset is a measure on $(\Omega, \mathcal{P}(\Omega))$. It can be generalized to countably infinite sets: on $\mathbb{N}$, the measure that associates the number of elements in a subset, or ∞ if the set is infinite, is a measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. It is called the **counting measure**.

Below are three essential properties of a measure.

Proposition 1.1. *A measure μ on $(\Omega, \mathcal{A})$ satisfies the following properties:*

1. ***Monotonicity**: Let $A, B \in \mathcal{A}$. $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.*
2. ***Sigma-sub-additivity**: for any countable collection $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}$, $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n)$.*
3. ***"Continuity property"**: If A_n is an increasing sequence for $\subseteq$ (meaning $A_n \subseteq A_{n+1}$ for all n), then $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$.*

Proof. All three proofs consist in re-partitioning the sets so as to end up with disjoint sets, and use sigma-additivity.

- Monotonicity: write $B = A \cup (B - A)$. Since A and $B - A$ are disjoint, $\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A)$.
- Sigma-sub-additivity: define the disjoint sequence of sets $(B_n)_n$ as $B_n = A_n - \bigcup_{k=1}^{n-1} A_k$. We have that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$ and $\mu(B_n) \leq \mu(A_n)$ for all n, so $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$.
- Define the pairwise disjoint collection of sets $B_n = A_n - A_{n-1}$ (and $B_1 = A_1$), so that $A_n = \bigcup_{i=1}^{n} B_i$ and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Then using sigma-additivity twice:

$$\mu(A_n) = \mu\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \mu(B_i) \to \sum_{i=1}^{\infty} \mu(B_i) = \mu\left(\bigcup_{i=1}^{\infty} B_i\right) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right).$$

□

The continuity property is really a particular application of sigma-additivity, but a very useful one to find the measure of a set. To find the measure of a set A, if we can write A as the limit of an increasing sequence of sets A_n the measure of which we know, then we can find $\mu(A)$ as the limit of the real-valued sequence $(\mu(A_n))_n$.

Finally, just a piece of vocabulary. We think of sets of measure zero as negligible. So if a property is true everywhere except on a set of measure zero, we want to say that it is "almost true". To make such statements rigorous, we define the notion of true almost everywhere.

Definition 1.5. *A property is true **almost everywhere**, abbreviated **a.e.**, if the set on which it is false has measure zero.*

1.3 Probabilities

Let us come back to our main interest: probabilities. From a mathematical perspective, a probability is just a particular case of a measure on a set: one such that the whole set has size one. There is nothing in the definition of a probability that stresses that we will apply such measures to measuring the likeliness of events.

Definition 1.6.

- *A measure such that $\mu(\Omega)$ is finite is called a **finite measure**.*

  (If so, using monotonicity, any measurable subset has finite measure).
- *A (finite) measure such that $\mu(\Omega) = 1$ is called a **probability**.*

When dealing with a probability:

- *we call* ***events*** *the measurable sets of the associated sigma-algebra.*
- *we call the measure space a* ***probability space****.*
- *we say that a property is true* ***almost surely (a.s.)*** *if it is true on a set of probability 1.*

The only difference between a finite measure and a probability is the cosmetic additional requirement of the normalization of $\mu(\Omega)$ to 1. There is nothing more complicated, but also nothing to gain, in studying finite measures that are not normalized to one, and so we restrict to probabilities only. Let us just add one definition, which is not nearly as important as finite measure, but will show up as a technical requirement in theorems below.

Definition 1.7. *Let μ be a measure on a measurable set $(X, \mathcal{A})$.*

μ is ***sigma-finite*** *if there exists a countable family of subsets $(A_k)_k \in \mathcal{A}$ of finite measure $\mu(A_k) < \infty$ for all k, such that $X \subseteq \bigcup_{k=1}^{\infty} A_k$.*

A probability—a finite measure—is necessarily sigma-finite.

2 Defining measures by extension on $\mathbb{R}^n$

To define probabilities on a countable set, we usually take the sigma-algebra on Ω to be the entire power set $\mathcal{P}(\Omega)$. On such a measurable space, a probability $\mathbb{P}$ is entirely characterized by the probabilities $p_\omega = \mathbb{P}(\{\omega\})$ that it assigns to each singleton $\{\omega\}$—equivalently to each element ω. Indeed, given positive numbers (p_ω) for all singletons, sigma-additivity allows to recover the probability of all subsets of Ω, since any subset is the countable union of its elements. The only requirement is that $\mathbb{P}(\Omega) = 1$: that the p_ω's sum to one. Thus, our definition of a probability chimes with the one we gave in the introduction for finite and countably infinite sets. The benefit of our new definition is that it also applies to uncountable sets such as the real line $\mathbb{R}$, where we would have no way to extend a probability from singletons to Borel sets.

But our general definition of a probability—and more generally of a measure—is not constructive: in practice, how to define a measure on $\mathbb{R}$ and $\mathbb{R}^n$? The strategy we are going to adopt is not so different from the one we used for countable sets: we are going to define our measures on a simple collection of subsets of $\mathbb{R}$ or $\mathbb{R}^n$, and then extend it to the whole Borel sigma-algebra.

2.1 Carathéodory's extension theorem

Two questions then. First, what simple collection of sets? Simple: remember that we defined the Borel sigma-algebra as the one generated by open intervals (a, b) of $\mathbb{R}$. So we are going to pick intervals as our simple collection of sets. There is one technical subtlety however. For technical reasons, it is more practical to use right-semiclosed intervals $(a, b]$. As is easily checked, they too generate the Borel sigma-algebra. In this course—the notation is not universal—we will note $\mathcal{I}$ the set of all right-semiclosed intervals $(a, b]$, and more generally $\mathcal{I}^n$ for the corresponding set on $\mathbb{R}^n$.

$$\mathcal{I} = \{(a, b], a, b \in \mathbb{R} \cup \pm\infty, a < b\} \cup \emptyset$$
$$\mathcal{I}^n = \left\{ \prod_{i=1}^{n} (a_i, b_i], a_i, b_i \in \mathbb{R} \cup \pm\infty, a_i < b_i \text{ for all } i \right\} \cup \emptyset$$

Note that we impose $a < b$ since $(a, a]$ would not make sense, and that we allow a and b to be $\pm\infty$; in particular $\mathbb{R}^n$ belong to $\mathcal{I}^n$. Also, note that we add the empty set to $\mathcal{I}^n$.

Second, how to extend our measure? We will not dig too much into the details here, and instead admit the following theorem, that gives the existence and uniqueness of an extension of a measure from $\mathcal{I}$ to $\mathcal{B}(\mathbb{R})$.

Theorem 2.1. *(**Carathéodory's extension theorem** on $\mathbb{R}^n$). Let μ be a function from $\mathcal{I}^n$ to $\mathbb{R}$. If:*

1. *μ is a "measure" on $\mathcal{I}^n$ ("measure" is a bit abusive since $\mathcal{I}^n$ is not a sigma-algebra), that is:*
   (a) *μ takes values in $\mathbb{R}_+ \cup \{+\infty\}$, and $\mu(\emptyset) = 0$.*
   (b) *μ is sigma-additive on $\mathcal{I}^n$: for all $(A_k)_{k\in\mathbb{N}} \in \mathcal{I}^n$ pairwise disjoint, if $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{I}^n$,*

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k).$$

2. *μ is sigma-finite.*

Then there exists a unique measure μ^* on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\mu(A) = \mu^*(A)$ on all $A \in \mathcal{I}^n$.

Proof. Admitted. □

The uniqueness part of the Carathéodory theorem tells us that a measure is characterized by its values on $\mathcal{I}^n$: if two measures coincide on $\mathcal{I}^n$, then they are equal on the whole Borel set $\mathcal{B}(\mathbb{R}^n)$. The existence part of the Carathéodory theorem tells us that it is possible to extend a measure from $\mathcal{I}^n$ to $\mathbb{R}^n$. Thus, when defining a measure, it is enough to define the values the measure takes on $\mathcal{I}^n$. We turn to two applications of this: defining probabilities on $\mathbb{R}$, and defining the Lebesgue measure on $\mathbb{R}^n$.

2.2 Application 1: defining probabilities on $\mathbb{R}$ through their CDF

Let us consider the case of probabilities on $\mathbb{R}$. Because a probability is a finite measure, we can simplify the problem further: it is enough to define a probability $\mathbb{P}$ on the set $\mathcal{J} \subseteq \mathcal{I}$ of intervals of the form $(-\infty, x], x \in \mathbb{R}$. Indeed, for $a < b$, $(-\infty, b] = (-\infty, a] \cup (a, b]$, so $\mathbb{P}((a, b]) = \mathbb{P}((-\infty, b]) - \mathbb{P}((-\infty, a])$. Thus, from the knowledge of $\mathbb{P}$ on $\mathcal{J}$, we can back out the values of $\mathbb{P}$ on $\mathcal{I}$.[3] Why is $\mathcal{J}$ an improvement with respect to $\mathcal{I}$? Because, since the intervals $(-\infty, x]$ are indexed by a single number x, we can sum up all the information that we need to define $\mathbb{P}$ through a single-variable function. (For $x = \pm\infty$, we always have $\mathbb{P}((-\infty, -\infty)) = \mathbb{P}(\emptyset) = 0$ and $\mathbb{P}((-\infty, +\infty)) = \mathbb{P}(\mathbb{R}) = 1$). Given a probability $\mathbb{P}$ defined on $\mathcal{B}(\mathbb{R})$, we call this function the cumulative distribution function of $\mathbb{P}$.

Definition 2.1. *The **cumulative distribution function (CDF)** F of a probability $\mathbb{P}$ is the function*

[3] The fact that $\mathbb{P}$ is finite intervenes in excluding the possibility of an indeterminate form $\mathbb{P}((a, b]) = \infty - \infty$.

from $\mathbb{R}$ *to* $[0,1]$ *defined as:*

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}((-\infty, x])$$

As an immediate corollary of the uniqueness part of the Carathéodory theorem, we know that the CDF of a probability $\mathbb{P}$ characterizes $\mathbb{P}$: two probabilities that have the same CDF define the same function on $\mathcal{J}$, hence on $\mathcal{I}$, hence on $\mathcal{B}(\mathbb{R})$, thus are equal. The existence part of the theorem can help us to characterize the functions F that correspond to a probability.

Proposition 2.1. *A function F from $\mathbb{R}$ to $[0,1]$ is a CDF of a probability $\mathbb{P}$ on $\mathbb{R}$ if and only if it is:*

1. *Increasing.*
2. *Right-continuous (everywhere): $\forall x_0, \forall \varepsilon > 0, \exists \delta > 0 / 0 \leq x - x_0 < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$.*
3. *With limit 0 in $-\infty$ and limit 1 in $+\infty$.*

Proof. We are going to skip to proof, but roughly: F increasing parallels the monotonicity of $\mathbb{P}$, F right-continuous parallels the continuity property of $\mathbb{P}$, and the limits in $\pm\infty$ parallel the measure of the empty set and the whole real line. □

Be careful that a CDF does not need to be left-continuous (hence continuous). For instance if we flit a coin and gain one dollar with probability $1/2$, and lose one dollar with probability $1/2$, that associated CDF on $\mathbb{R}$ will be discontinuous at 0 and 1, jumping from 0 to $1/2$, and from $1/2$ to 1.

2.3 Application 2: the Lebesgue measure on $\mathbb{R}^n$

We now turn to the only infinite measure—meaning that it assigns an infinite measure to some sets—that we will be concerned with in this chapter: the Lebesgue measure. The Lebesgue measure is meant to capture the "natural" measure on the real line $\mathbb{R}$. What we mean by natural is that it corresponds for an interval $(a, b]$ to its length $b - a$ (and if the interval is unbounded—$a = -\infty$ or $b = +\infty$—the measure is infinite). It is easy to check that this measure defined on $\mathcal{I}$ is sigma-additive on $\mathcal{I}$ and sigma-finite ($\mathbb{R} = \bigcup_{n \in \mathbb{Z}} (n, n+1]$), so Carathéodory theorem tells us that it extends uniquely to a measure defined on $\mathcal{B}(\mathbb{R})$.

Definition 2.2.

The Lebesgue measure λ on $\mathbb{R}$ is the unique extension to $\mathcal{B}(\mathbb{R})$ of the function defined on $\mathcal{I}$ as:

$$\lambda((a,b]) = b - a \text{ (and } +\infty \text{ if } a = -\infty \text{ or } b = +\infty).$$

The Lebesgue measure λ on $\mathbb{R}^n$ is the unique extension to $\mathcal{B}(\mathbb{R}^n)$ of the function defined on $\mathcal{I}^n$ as:

$$\lambda\left(\prod_{i=1}^{n}(a_i,b_i]\right) = \prod_{i=1}^{n}(b_i - a_i) \text{ (and } +\infty \text{ if } a_i = -\infty \text{ or } b_i = +\infty \text{ for some } i).$$

The notation λ for the Lebesque measure is standard, although be careful that λ if sometimes used to designate other measures as well.

One remark: the Lebesgue measure is actually slightly extended in the following way. Define $\mathcal{N}$ as the collection of all subsets of $\mathbb{R}$ (not necessarily Borel-measurable) that are included in a Borel set of Lebesgue measure zero: $A \subseteq B$, $\lambda(B) = 0$. It is possible to show that $\mathcal{B}(\mathbb{R}) \cup \mathcal{N}$ is a sigma-algebra. We add $\mathcal{N}$ to the collection of sets on which we define the Lebesgue measure, and define the measure of the subsets in $\mathcal{N}$ to be zero.

Another remark, as names may be intriguing you: does the Lebesgue measure on $\mathbb{R}$ solve the Lebesgue problem? Yes: it is sigma-additive and normalized to 1, and can be shown to be invariant by translation (that it is invariant by translation on $\mathcal{I}$ is straightforward). It can also be shown to be the only measure that satisfy these three properties.

Here are some results that are useful to get a better sense of the Lebesgue measure.

Proposition 2.2.

- *Any singleton of $\mathbb{R}$ has Lebesgue-measure zero.*
- *Any countable set of $\mathbb{R}$—in particular $\mathbb{N}$ and $\mathbb{Q}$—has Lebesgue-measure zero.*
- *There are also uncountable sets of $\mathbb{R}$ that have a zero Lebesgue-measure.*

Proof.

- Let $\{a\}$ be a singleton. Since $(a-1,a] = (a-1,a) \cup \{a\}$, which are disjoint sets, if we show that $\lambda((a-1,a)) = 1$ we are done since then $\lambda(\{a\}) = \lambda((a-1,a]) - \lambda((a-1,a)) = 1 - 1 = 0$. Write $(a-1,a) = \bigcup_{n=1}^{\infty}(a-1,a-\frac{1}{n}]$. The sets $(a-1,a-\frac{1}{n}]$ are increasing so by the continuity property, $\lambda((a-1,a)) = \lim_{n\to\infty}\lambda((a-1,a-\frac{1}{n})] = \lim_{n\to\infty} 1 - \frac{1}{n} = 1$. QED.
- It is then a direct consequence of sigma-additivity that countable sets have measure zero.

- You will see an example in the problem-set.

□

Note that the fact that countable sets have zero Lebesgue measure seems reminiscent of the insight from cardinality theory that countable infinity is "negligible" in front of uncountable infinity. But the last item in the proposition warns us against the perils of the analogy between cardinality and the Lebesgue measure: the Lebesgue measure also sees as "negligible"—meaning of measure zero—some uncountable sets.

3 Defining measures by image

An alternative way to define a measure on a measurable space $(Y, \mathcal{B})$ is by importing a measure μ from another measurable space $(X, \mathcal{A})$ where μ is defined. Consider the example of playing (European) roulette. We can model it through a sample space Ω consisting of the 37 outcomes:

$$\Omega = \{\text{the ball falls into 0, the ball falls into 1, ..., the balls falls into 36}\}$$

We can turn Ω into a measurable space by coupling it with the sigma-algebra $\mathcal{P}(\Omega)$, and into a measured space by adding for instance the uniform probability $\mathbb{P}$ that assigns $1/37$ to each outcome. Now Roulette is only fun if there is something at stake, so let us assume we bet \$1 on number 25 so that we win \$35 if the ball falls into 25, and lose \$1 if it falls into anything else. It is natural to define the payoff function f from Ω to $\mathbb{R}$:

$$f : \Omega \to \mathbb{R}$$
$$\{\text{the ball falls into 25}\} \mapsto 35$$
$$\text{any other outcome} \mapsto -1$$

You may have already calculated that our probability of getting \$35 is $1/37$, and our probability of getting -\$1 is $36/37$. But do note how we come up with these numbers. To come up with $1/37$, we sum up the probabilities of all the outcomes that lead to a payoff of 35—there is only one. To come up with $36/37$, we sum up the probabilities of all the outcomes that lead to a payoff of -\$1. We could also calculate the probability of getting any other real number x as a payoff: there is no outcome that leads to such payoff, so the probability would be $\mathbb{P}(\emptyset) = 0$. We could also get the probability of getting less than \$50 dollar, or of any event—any measurable subset S of $\mathbb{R}$. In all cases, the logic would be to sum the probabilities of all the outcomes of Ω such that the payoff is in S. In other words, we define a probability $\mathbb{Q}$ on the real line as the probability of its inverse image.

$$\forall B \in \mathcal{B}(\mathbb{R}), \mathbb{Q}(B) = \mathbb{P}(f^{-1}(B))$$

This way of defining a measure is very general: given any two measurable sets $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, we can define a new measure on $(Y, \mathcal{B})$ from a measure μ on $(X, \mathcal{A})$ as long as we have a function f such that the inverse image of any measurable set of Y is a measurable set of X. This leads us to define measurable functions.

3.1 Measurable functions

Definition 3.1. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces, and $f : X \to Y$ a function.*

*f is a **measurable function** if the inverse images by f of a measurable set of Y is a measurable set of X:*

$$\forall B \in \mathcal{B}, f^{-1}(B) \in \mathcal{A}$$

Note that the definition of a measurable function makes no reference to the existence of a measure on either set.

3.2 Image measure

So now, here is our general result:

Proposition 3.1.

Let $(X, \mathcal{A}, \mu)$ be a measure space, $(Y, \mathcal{B})$ be a measurable space, and $f : X \to Y$ a measurable function.

The function μ_f on $(Y, \mathcal{B})$ defined by:

$$\mu_f(B) = \mu(f^{-1}(B)) \text{ for all } B \in \mathcal{B}$$

*is a measure on $(Y, \mathcal{B})$, called **image measure of μ under f**.*

Proof. The function μ_f is positive and $\mu_f(\emptyset)$=$\mu(f^{-1}(\emptyset)) = \mu(\emptyset) = 0$. As for sigma-additivity, let (B_n) be a countable family of pairwise disjoint measurable sets:

$$\mu_f\left(\bigcup_{n=1}^{\infty} B_n\right) = \mu\left(f^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right)\right) = \mu\left(\bigcup_{n=1}^{\infty} f^{-1}(B_n)\right) = \sum_{n=1}^{\infty} \mu(f^{-1}(B_n)) = \sum_{n=1}^{\infty} \mu_f(B_n)).$$

□

3.3 Random variables

As always, probability theory is ashamed of just being a particular case of measure theory, and insists on having its own vocabulary.

Definition 3.2. *Let $(\Omega, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces.*

When the measures we deal with on $(\Omega, \mathcal{A})$ and $(Y, \mathcal{B})$ are probabilities:

- *We call a measurable function a **random variable** and usually note it X.*

- *When $(Y, \mathcal{B})$ is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we call X a **real random variable**.*
- *When $(Y, \mathcal{B})$ is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, we call X a **random vector**.*
- *We call the image probability of a probability $\mathbb{P}$ on Ω under X the **probability of the random variable** X, and note it $\mathbb{P}_X$.*

Remember that a probability on $\mathbb{R}$ is characterized by its CDF. Noting F_X the CDF of $\mathbb{P}_X$, we have:

$$\forall x \in \mathbb{R}, F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X^{-1}((-\infty, x]) = \mathbb{P}(\omega \in \Omega / X(\omega) \in (-\infty, x]) = \mathbb{P}(\omega \in \Omega / X(\omega) \leq x)$$
$$F_X(x) = \mathbb{P}(X \leq x)$$

So the vocabulary is different when we deal with probabilities. But also, the vocabulary is weird: the term *variable* is not very descriptive of what it names—a function. And the practice of probability theory adds to the confusion: theorems and exercises often start with *"Let X be a random variable with CDF given by..."*, while remaining silent on what the domain Ω of the function X is, and not really using at any point that X is a function.

Here is why we *can* do so: very often, we do not really care about the set Ω and the function X. They are only instrumental objects in defining a probability on $(Y, \mathcal{B})$. Once we have used them to define $\mathbb{P}_X$, we can just as well toss them aside. Actually, we could as well start by defining a probability on $(Y, \mathcal{B})$ directly through its CDF, without any reference to a set Ω or a function X. For instance, in the roulette example above, we could have started by directly defining the probability $\mathbb{Q}$ on the real line.

Why we *do* keep referring to our probability $\mathbb{Q}$ as the one associated to a function X defined of an implicit set Ω is just because it is a convenient way of thinking about it in our application of probability theory to hazard. We like to think of our probability $\mathbb{Q}$ on the real line as resulting—through a function—from some source of randomness in some set Ω. What is Ω? The roulette, the world, the universe: it is not very clear—but also it does not need to be very clear since only the image probability $\mathbb{Q}$ matters. And it is why we do not bother stating what Ω is.

4 Lebesgue Integral and expectations

You have seen a theory of integration—the Riemann integral—defined for continuous functions on a segment $[a, b]$ of the real line. Measure theory allows to build a theory of integration—the Lebesgue integral—that fits the Riemann integral when integrating a continuous function on a segment of the real line with respect to the Lebesgue measure. But the Lebesgue integral is a considerable generalization: it is defined for (almost) any measurable function from a measurable subset of a measure space to $\mathbb{R}$, and can be defined for different measures, not only the Lebesgue measure. (Don't be confused by vocabulary: the Lebesgue integral does not need to be taken with respect to the Lebesgue measure).

4.1 Sketch of the construction of the integral

We will note enter the details of the construction of the Lebesgue integral, but simply sketch the steps. Fist, some definitions:

Definition 4.1. *Let $(X, \mathcal{A})$ be a measurable space.*

- *The **indicator function** of a measurable set $A \in \mathcal{A}$, noted $\mathbb{1}_A$ is the real function defined on X as $\mathbb{1}_A(x) = 1$ if $x \in A$ and $f(x) = 0$ if $x \notin A$.*
- *A **simple function** is a finite linear combination of indicator functions of measurable sets:*

$$f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}, a_i \in \mathbb{R}, A_i \in \mathcal{A}, \forall i = 1, ..., n.$$

Consider a measure space $(X, \mathcal{A}, \mu)$. We will note the integral of a function f, $\int f d\mu$, or $\int f(x) d\mu(x)$, which makes explicit the dependence of the integral in the measure with respect to which the integral is taken. The construction of the Lebesgue integral is in three steps.

1. Define the integral of a positive simple function $f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$, $a_i \geq 0$ for all i, with respect to μ as:

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

2. Extend the definition to positive measurable functions. To do so, we show that any positive measurable function can be approximated as the (pointwise) limit of a (pointwise) increasing sequence (f_n) of positive simple functions. We show that the sequence $(\int f_n d\mu)$ converges in $\mathbb{R} \cup \{+\infty\}$, and that the limit is

independent of the choice of the approximating sequence (f_n). We define the integral of f as the common limit. We say that f is **integrable** if its integral is finite: $\int f d\mu < +\infty$.

3. Extend the definition to (almost) all measurable function, not necessarily positive. To do so, we define two measurable positive functions f^+ and f^- as: $f^+(x) = f(x)$ if $f(x) \geq 0$ and $f(x) = 0$ otherwise; $f^-(x) = -f(x)$ if $f(x) \leq 0$ and $f(x) = 0$ otherwise. We can therefore decompose $f = f^+ - f^-$. When $\int f^+ d\mu < \infty$ or $\int f^- d\mu < \infty$, we define the integral of f as:

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

(We cannot define the integral when both $\int f^+ d\mu = \infty$ and $\int f^- d\mu = \infty$). We say that f is **integrable** if both f^+ and f^- are finite, or equivalently if $|f|$ is integrable ($\int |f| d\mu < \infty$).

The Lebesgue integral generalizes the Riemann integral when the measure space is the real line endowed with the Borel sigma-algebra and the Lebesgue measure. We therefore often note $\int f(x)dx$ for the integral with respect to the Lebesgue measure $\int f(x)d\lambda(x)$. But the Lebesgue integral is not restricted to the Lebesgue measure, not even the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For instance, consider the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where μ is the counting measure. Then any function from $\mathbb{N}$ to $\mathbb{R}$—any real sequence—is measurable, and the Lebesgue integral of an integrable sequence (x_n) is $\sum_{n=1}^{\infty} x_n$. Some basic properties of the Lebesgue integral:

Proposition 4.1. *Let $(X, \mathcal{A}, \mu)$ be a measure space.*

- *The integral is linear: if f and g are integrable, and $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g$ is integrable and:*

$$\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu.$$

- *The integral is monotonic: for f and g integrable, if $f \leq g$, then $\int f d\mu \leq \int g d\mu$.*
- *If two integrable functions f and g are equal almost everywhere, then $\int f d\mu = \int g d\mu$.*

Proof. Admitted. □

4.2 Monotone convergence theorem and Dominated convergence theorem

In this section, we gather results that allow to "permute the integral and the limit", or "permute the integral and differentiation". Let us start with the limit. In parallel with the construction of the integral, there are two

results: one for positive measurable functions—the monotone convergence theorem—and one for any integrable function—the dominated convergence theorem.

Theorem 4.1. ***Monotone Convergence Theorem***

If (f_n) is an increasing sequence of positive measurable functions that tends almost everywhere to f (which may be defined as $+\infty$ in some points), then f is measurable and $\lim_{n\to\infty} \int f_n d\mu = \int f d\mu$.

Proof. Admitted. □

Theorem 4.2. ***Dominated Convergence Theorem***

If (f_n) is a sequence of measurable functions that tends almost everywhere to f, and there exists an integrable function g such that $|f_n| \leq g$ for all n, then f is integrable and $\lim_{n\to\infty} \int f_n d\mu = \int f d\mu$.

(Note that the f_n are guaranteed to be integrable by the monotonicity of the integral).

Proof. Admitted. □

Given that the derivative of a function from $\mathbb{R}$ to $\mathbb{R}$ is defined as a limit, it may not be too surprising that we can deduce from the dominated convergence theorem a theorem that allows us to permute integration and differentiation. To state it, consider the Borel sigma-algebra and the Lebesgue measure on the real line, and consider a function f of two real variables: one with respect to which we integrate—t—and one with respect to which we differentiate—x. When integrating over t, we obtain a function $h(x)$ of the single variable x.

Proposition 4.2. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ such that $t \mapsto f(t,x)$ is integrable for all $x \in \mathbb{R}$.*
Note $h(x) = \int f(t,x)dt$.
If $x \mapsto f(t,x)$ is differentiable for all t, and there exists an integrable function g such that $|\frac{\partial}{\partial x} f(t,x)| \leq g(t)$, then h is differentiable and:

$$h'(x) = \frac{d}{dx} \int f(t,x)dt = \int f'_x(t,x)dt$$

Proof. Admitted. □

4.3 Double integrals and Fubini theorem in $\mathbb{R}^2$

One way in which the Lebesgue integral is more general than the Riemann integral is that it painlessly defines an integral on $\mathbb{R}^n$, with respect to the Lebesgue measure of $\mathbb{R}^n$. For instance in $\mathbb{R}^2$, we can consider a function

$f(x_1, x_2)$ of two variables, and take its integral with respect to the Lebesgue measure of $\mathbb{R}^2$. Now, you might also want to consider the real number obtained by first, for each value of x_2, integrating f with respect to its first variable x_1 (with respect to the Lebesgue measure in $\mathbb{R}$), and second integrating the resulting function of x_2 with respect to x_2 (with respect to the Lebesgue measure in $\mathbb{R}$). Or you might want to do the same thing permuting the order of x_1 and x_2. Thus we can calculate three real numbers. Fubini's theorem guarantees that these three numbers are equal (if f is integrable). The result is twofold. First it gives us a way to calculate the integral of a function of two variables in practice, integrating along each variable successively. Second, it allows us to permute the order of integration when facing double integrals.

Theorem 4.3. ***(Fubini's theorem)*** *Let $f : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function.*
If f is integrable (with respect to the Lebesgue measure in $\mathbb{R}^2$), then:

$$\begin{aligned}\int_{\mathbb{R}^2} f(x_1, x_2) d\lambda(x_1, x_2) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x_1, x_2) d\lambda(x_1) \right) d\lambda(x_2) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x_1, x_2) d\lambda(x_2) \right) d\lambda(x_1).\end{aligned}$$

(The theorem implicitly guarantees that the one-variable functions on the right-hand sides are integrable).

Proof. Admitted. □

Fubini theorem is easily extended to integration in $\mathbb{R}^n$, except that the notations get messier. In particular, we can permute the order of integration in any way.

4.4 Integral wrt. an image measure

Consider the case of a measure ν on a measurable space $(Y, \mathcal{B})$ defined as the image measure of a measure μ on a measurable space $(X, \mathcal{A})$ under the measurable function $\phi : X \to Y$. The integral with respect to ν can be expressed as an integral with respect to μ.

Proposition 4.3. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces.*
Let ϕ be a measurable function from X to Y, μ a measure on $(X, \mathcal{A})$, ν the image measure of μ under ϕ.
Then for any integrable function $f : Y \to \mathbb{R}$,

$$\int_Y f d\nu = \int_X (f \circ \phi) d\mu.$$

Proof. This is easy to check for indicator function: $\int_Y \mathbb{1}_B d\nu = \nu(B) = \mu(\phi^{-1}(B) = \int_X \mathbb{1}_{\phi^{-1}(B)} d\mu = \int_X (\mathbb{1} \circ \phi) d\mu$. The proof—that we admit—then generalizes the result to any measurable function. □

4.5 Expectations

Once again, probability theory has its own vocabulary: if $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, we call the integral of a real random variable X from Ω to $\mathbb{R}$ the expectation of X, noted $\mathbb{E}(X)$.

Definition 4.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and X a real random variable from Ω to $\mathbb{R}$. If X is integrable, we call its integral the **expectation** of X, and note it $\mathbb{E}(X)$:*

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

Now remember that from the random variable X we can define the probability $\mathbb{P}_X$ of X on $\mathbb{R}$ as the measure image of $\mathbb{P}$ under X. Then, simply applying proposition 4.3 for $f = Id$, we get that:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x d\mathbb{P}_X(x)$$

It follows that the expectation of a random variable depends only on the probability $\mathbb{P}_X$. This is no surprise: we have seen that once we have defined $\mathbb{P}_X$ from $\mathbb{P}$ and X, we could just as well get rid of $\mathbb{P}$, X and Ω. More generally, for any function $f : \mathbb{R} \to \mathbb{R}$, $f \circ X$ is integrable wrt. $\mathbb{P}$ if and only if f is integrable wrt. $\mathbb{P}_X$, and:

$$\mathbb{E}(f(X)) = \int_{\Omega} f \circ X(\omega) d\mathbb{P}(\omega) = \int_{\mathbb{R}} f(x) d\mathbb{P}_X(x).$$

The expectation of $f(X)$ when f is a power function plays an important role:

Definition 4.3. *Provided they exist, we define for a real random variable X:*

- *The **raw moment** of order p: $\mathbb{E}(X^p)$.*
- *The **central moment** of order p: $\mathbb{E}((X - \mathbb{E}(X))^p)$.*

We call the central moment of order 2 the **variance**: $V(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$. Note the two following results about the variance:

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

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

- $V(X) = 0$ if and only if X is almost surely constant.

  (Admitted).

5 Densities

5.1 Defining measures trough densities

From any measure μ on a measurable set $(X, \mathcal{A})$, we can define a new measure for any positive measurable function f.

Definition 5.1. *Let $(X, \mathcal{A}, \mu)$ be a measure space. The function:*

$$\nu : \mathcal{A} \to \mathbb{R}_+ \cup \{+\infty\}$$
$$A \mapsto \int \mathbb{1}_A f d\mu$$

*is a measure on $(X, \mathcal{A})$, called measure induced by the **density** f with respect to μ.*

*If $\int f d\mu = 1$, f induces a probability; f is then also called a **probability density function (PDF)**.*

Proof. We admit that this defines a measure. □

The integral of a function g with respect to ν can be expressed as an integral with respect to μ:

Proposition 5.1. *Let ν be the measure induced by the density f wrt. μ.*

For any measurable function g from $(X, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$,

$$\int g d\nu = \int g \times f d\mu.$$

Proof. Admitted. □

For this reason, the density f is sometimes noted $\frac{d\nu}{d\mu}$.

5.2 The Radon-Nikodyn theorem

Given two measures μ and ν on $(X, \mathcal{A})$, can we always find a measurable positive function f such that ν has density f with respect to μ? The following reasoning shows that we cannot: if there exists a measurable set A such that $\nu(A) > 0$ but $\mu(A) = 0$, then no f will do. Indeed, for all f, $\int \mathbb{1}_A f d\mu = 0 \neq \nu(A)$. However, the Radon-Nikodyn theorem guarantees that except for these measures ν that attribute measure zero to sets of positive μ-measure, we can. Let us first give a name to the measures that will work.

Definition 5.2. *Let $(X, \mathcal{A}, \mu)$ be a measure space.*

A second measure ν on $(X, \mathcal{A})$ is ***absolutely continuous*** *wrt. μ if:*

$$\forall A \in \mathcal{A}, \mu(A) = 0 \Rightarrow \nu(A) = 0.$$

Theorem 5.1. ***(Radon-Nikodyn theorem)***

Let $(X, \mathcal{A})$ be a measurable space and μ and ν two measures on $(X, \mathcal{A})$.

If μ and ν are both sigma-finite, and ν is absolutely continuous wrt. μ, then ν has a density wrt. μ.

The density is unique up to a μ-null set, i.e. any two densities are equal except on a set of measure zero.

Proof. Admitted. □

5.3 Application to probabilities on $\mathbb{R}$

Consider the application of densities to probabilities on the real line (endowed with the Borel sigma-algebra). We take the Lebesgue measure as the reference measure, and a probability $\mathbb{P}$, with CDF F, as the second measure. Relying on section 2.2, we know that for f to be a density of $\mathbb{P}$ wrt. to the Lebesgue measure, it suffices that f satisfy:

$$\forall x \in \mathbb{R}, F(x) = \int_{(-\infty, x]} f(t)dt$$

Relying on the fundamental theorem of calculus (that you proved using the Riemann theory of integration):

Theorem 5.2. ***Fundamental theorem of calculus***

Let $f : [a, b] \to \mathbb{R}$ be a continuous function.

For any function F s.t. $F' = f$ —a ***indefinite integral***, ***primitive integral***, *or* ***antiderivative*** *of f,*

$$\int_a^b f(t)dt = F(b) - F(a).$$

we deduce that whenever the CDF F of $\mathbb{P}$ is $\mathcal{C}^1$, the density of $\mathbb{P}$ wrt. the Lebesgue measure is the derivative $f = F'$ of the CDF.

Proposition 5.2. *Let $\mathbb{P}$ be a probability on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$; note F its CDF.*

If F is $\mathcal{C}^1$, then it has density $f = F'$ with respect to the Lebesgue measure.

Therefore, for a real random variables X whose probabilities $\mathbb{P}_X$ has a $\mathcal{C}^1$ CDF F and PDF $f = F'$, we can calculate the expectation of any measurable function $g(X)$ as:

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) d\mathbb{P}_X(x) = \int_{\mathbb{R}} g(x) f(x) dx.$$

This boils down the calculation of expectations to the calculation of integrals with respect to the Lebesgue measure.

Another example of a density is the case of a finite probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, as presented in the introduction of this chapter. There, we have seen that the probability $\mathbb{P}$ is characterized by n positive numbers p_i that sum to 1. We can now look at the p_i under a new light: they are the density of $\mathbb{P}$ with respect to the counting measure. So in a sense, the density is the generalization of the p_i to probabilities on uncountably infinite sets. Be careful however about a common misconception concerning densities: when the probability is defined on the real line, with respect to the Lebesgue measure, the value $f(x)$ is not to be interpreted as the probability of the event $\{x\}$ (to be convinced, notice that whenever the probability is absolutely continuous wrt. the Lebesgue measure, all singletons have actually measure zero). In particular, the value $f(x)$ needs be positive but not necessarily less than one.

5.4 Change of Variables

You have seen with the Riemann integral on the real line that it is often useful to rely on a **change of variable** (or **integration by substitution**) to calculate an integral. If $g : [u_a, u_b] \to \mathbb{R}$ is $\mathcal{C}^1$ and such that $g(u_b) = b$ and $g(u_a) = a$, we can use the change of variable $x = g(u)$:

$$\int_a^b f(x) dx = \int_{u_a}^{u_b} f(g(u)) g'(u) du.$$

To memorize it: $\frac{dt}{du} = g'(u) \Leftrightarrow dt = g'(u) du$.

Lebesgue integration allows a generalization of this result to integrals in $\mathbb{R}^n$ with respect to the Lebesgue measure. Here we state the result for $n = 2$.

Proposition 5.3. *Consider the integral $\int_A f(X) d\lambda(X)$, where A is an open set of $\mathbb{R}^2$, and f is integrable.*

Suppose there exists a bijection ϕ:

$$\begin{aligned}\phi : \phi^{-1}(A) &\to A \\ U &\mapsto X = \phi(U)\end{aligned}$$

such that both ϕ and ϕ^{-1} are $\mathcal{C}^1$ (ϕ is then called a $\mathcal{C}^1$-diffeomorphism). Then:

$$\int_A f(X)d\lambda(X) = \int_{\phi^{-1}(A)} f(\phi(U)) \times |det(\phi'(U))|d\lambda(U),$$

where $|det(\phi'(U))|$ is the absolute value *of the determinant of the derivative of ϕ.*

Proof. The proof, which we admit, relies on showing that the inverse of $|det(\phi'(U))|$ is the density of some measure with respect to the Lebesgue measure, which explains why the proposition is in the section. □

Example. *We want to calculate $I = \int_{-\infty}^{+\infty} e^{-x^2}dx$. Notice that I^2 can be written as the double integral $I^2 = \iint e^{-(x^2+y^2)}dxdy$. We calculate I^2 using the change of variables to polar coordinates:*

$$\begin{aligned}\phi : \mathbb{R}_+^* \times [0, 2\pi) &\to (\mathbb{R}^2)^* \\ (r, \theta) &\mapsto (r\cos(\theta), r\sin(\theta))\end{aligned}$$

It is a $\mathcal{C}^1$-diffeomorphism. The gradient of ϕ is:

$$\phi'(r, \theta) = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}$$

so that $|det(\phi'(r, \theta))| = |r\cos^2(\theta) + r\sin^2(\theta)| = r$. Therefore:

$$I^2 = \int_0^{2\pi} \left(\int_0^{+\infty} e^{-r^2} r dr \right) d\theta = \int_0^{2\pi} \left[-\frac{1}{2} e^{-r^2} \right]_0^{+\infty} d\theta = \int_0^{2\pi} \frac{1}{2} d\theta = \pi.$$

Hence $I = \sqrt{\pi}$. Note that this implies that $f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$ is the density of a probability with respect to the Lebesgue measure. The probability is known as the normal distribution of mean 0 and variance $1/\sqrt{2}$.

6 Conditional probability and conditional expectations

So far, nothing that we have done about probabilities has relied on interpreting probabilities as measures of hazard. In contrast, the notions of conditional probabilities and conditional expectations that we now formalize only make sense within this interpretation.

6.1 Conditional probability

Definition 6.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $B \in \mathcal{A}$ an event with non-nil probability $\mathbb{P}(B) \neq 0$. The function $\mathbb{P}(.|B)$ defined on $\mathcal{A}$ by:*

$$P(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

is a probability. It is called the ***conditional probability given*** *B.*

Proof. This defines a positive function and $\mathbb{P}(\emptyset|B) = 0$. As for sigma-additivity, let $(A_n)_{n\in\mathbb{N}}$ be pairwise disjoint measurable sets. Then:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n|B\right) = \frac{\mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{n=1}^{\infty}(A_n \cap B)\right)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n|B).$$

□

6.2 Conditional expectations

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a real random variable Y. Calculating the expectation of Y conditional on some event B poses no difficulty: it is just the integral with respect to the particular measure $\mathbb{P}(.|B)$. While the (unconditional) expectation $\mathbb{E}(Y)$ is meant to capture the average value that we can expect Y to take, the conditional expectation $\mathbb{E}(Y|B)$ is meant to capture the average value that we can expect Y to take when we know that the event B occurs.

We are however also interested in defining a different object, the expectation of Y conditional on *another random variable* X (not an event B), meant to represent the value we can expect Y to take when we know the value x that X takes (this way defining a function of x). We could think of doing it in exactly the same way, defining for each $x \in \mathbb{R}$, $\mathbb{E}(Y|X = x)$. However, this would only work for the x such that $\mathbb{P}(X = x) > 0$. And most of the real random variables we deal with are absolutely continuous wrt. the Lebesgue measure, so that they put a zero probability on singletons. So we need a different approach.

Restrict to real random variables whose square is integrable—random variables for which the variance is defined. It is possible to check that all the set of square-integrable function is a vector space; we note it L^2.[4] We turn it into a inner product space by defining the inner product:

$$< X, Y >= \mathbb{E}(XY),$$

which can be shown to be well defined and to satisfy the three axioms of an inner product. Remember that an inner product defines a norm, hence a distance between real random variables. Now the idea is to define the conditional expectation $\mathbb{E}(Y|X)$ as the (measurable) function of X that best approximates Y, meaning that minimizes the distance to Y:

$$\mathbb{E}(Y|X) = argmin_{f(X)\in L^2}||Y - f(X)|| = (\mathbb{E}(Y - f(X))^2)^{\frac{1}{2}}$$

A famous theorem in analysis—Hilbert projection theorem— guarantees the existence and uniqueness of the solution to this minimization problem, and therefore that the conditional expectation is well defined.

The conditional expectation conditional on the random variable X is a very different object from the conditional expectation conditional on the event B: the second is a number between 0 and 1, while the first is a random variable. As such (and because $\mathbb{E}(Y|X)$ is by definition in L^2), we can calculate its first and second moments. First, its expectation:

Proposition 6.1. ***(Law of Iterated Expectations)***

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y).$$

Proof. Admitted. □

We add a result for the variance. First, we define the conditional variance of Y conditional on X as:

$$V(Y|X) = \mathbb{E}\big[(Y - \mathbb{E}(Y|X))^2|X\big].$$

Just as the conditional expectation, the conditional variance is a random variable. Now:

[4] There is only one subtlety: we actually identify as a single function all the real random variables that are almost surely equal. This avoids some difficulty: for instance this way the positive definiteness axiom in the definition of an inner product will be satisfied, whereas all the functions almost surely nil are such that $\mathbb{E}(X^2) = 0$.

Proposition 6.2.

$$V(\mathbb{E}(Y|X)) = V(Y) - \mathbb{E}(V(Y|X))$$

Proof. Admitted. □

This decomposes the variance of Y between the variance of its conditional expectation and the expectation of its conditional variance (this decomposition can be made for any random variable X). Note that the decomposition implies $V(Y) \geq V(\mathbb{E}(Y|X))$: we are less uncertain about Y when we know X.