Notes on Probability Theory and Statistics*

Joël Terschuur

*All material presented here can be traced down to the course Probability Foundation for Electrical Engineers by Dr. Krishna Jagannathan,Department of Electrical Engineering,IIT Madras in the Youtube Channel nptelhrd, Econometrics I lectures by Juan Carlos Escanciano at UC3M and other sources in the bibliography.

1 Introduction

In real life we encounter many events which we cannot predict perfectly or for which we have imperfect knowledge. Examples range from tossing of a coin, to the state of the weather in the next three days or the extent to which GDP will decrease after a pandemic. However, while it might seem we are clueless about many of these events, there are patterns we can study. We know that if we toss a coin a thousand times, we should be very close to half heads and half tails. We are able to forecast weather and GDP growth to some extent. In essence, probability theory is a science of randomness which allows us to make some reasonable predictions. It provides a foundation to the patterns we observe with more regularity in real life.

We have been playing games of chance and computing probabilities for centuries. In the 17^{th} century a gambler called Chevallier de Mere asked Laplace and Fermat for help. He was playing two betting games he considered equivalent but was consistently losing with one of them and not with the other. Pascal and Fermat showed why these two games were not the same and by doing this paved the way for computation of probabilities. However, modern probability theory is roughly 100 years old. The main founder of the axiomatic approach to probability which we encounter today is the Russian mathematician Andréi Kolmogórov (1903-1987). He noticed that probability theory is just a special case of measure theory which was developed by French mathematicians Émile Borel (1871-1956) and Henri Léon Lebesgue (1875-1941). While people already knew how to compute many probabilities before Kolmogórov, it was usual to run into puzzling results and paradoxes. With a strong axiomatic foundation, the answer to these apparent puzzles and paradoxes was resolved by unifying all that was known into one single logical framework.

We will first do a short review of set theory and cardinality of sets and then we will start diving into probability theory.

2 Mathematical Review

This section follows closely Kolmogorov and Fomin (1975) and De la Fuente (2000). A set is a collection of objects we call elements. For instance the set of all integers or the set of all even numbers. It started to be developed at the end of the 19th century with the paper Cantor (1874). It is a subject within mathematics in its own right and is crucial in many fields in modern mathematics. Specifically, a basic knowledge of set theory is essential to understand probability theory.

We will denote sets with capital letters like $A, B, ...$ and their elements with lower case letters such as $a, b, ...$ A set with elements $a, b, c, ...$ is often denoted by $\{a, b, c, ...\}$. For instance, the set of all positive integers is {1,2,3,...}. The singleton set containing only one element, for instance the number one, is $\{1\}$. If a is an element of a set A, "a belongs in A", we write $a \in A$. If a does not belong to a set A we write $a \notin A$. For instance, $2 \in \{1, 2, 3\}$ but $4 \notin \{1, 2, 3\}$. If every element of a set A belongs to a set B, we say that A is a subset of

B, denoted $A \subseteq B$ or $B \supseteq A$, the latter meaning "B contains A". Formally,

$$A \subseteq B \iff (x \in A \implies x \in B)$$

So from now on, if we want to show that a set A is a subset of a set B we show that all elements belonging to A also belong to B. For instance $\{1,2\} \subseteq \{1,2,3\}$. We say that two sets A and B are equal, $A = B$, if all elements belonging in A also belong in B and vice versa. That is, every time we want to show two sets are equal we need to show that $A \subseteq B$ and $B \subseteq A$. Also, if $A \subseteq B$ and $A \neq B$, then A is a proper subset of B, $A \subset B$. Sometimes we might not know whether a set contains any elements at all. For instance, the set of roots of a given equation. A set containing no elements is called the empty set and is denoted by $\emptyset$.

A collection or family or class of sets is a set whose elements are sets themselves. For instance if we have sets A, B, and C. A collection of sets containing these is $\mathcal{D} = \{A, B, C\}$. Collections of sets are usually denoted by calligraphic capital letters. It is important to see that now the elements of $\mathcal{D}$ are sets, that is, we write $A \in \mathcal{D}$. Hence, we do not say A is a subset of $\mathcal{D}$, it is an element belonging to it in the same way that an element a belongs in A, $a \in A$. Hence it is important to always have in mind what is the typical element of a set. Given a set A, an important collection of subsets is the collection of all subsets of A. This is called the power set of A, denoted by 2^A. For instance, if $A = \{1,2,3\}$ we have that

$$2^A = \Big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\Big\}$$

Again note that 2^A is a set of sets[1]. From now own suppose there is some universal set X and that there is nothing outside of it. We will work only with subsets of X.There are two operations on sets which will be used often. The union and the intersection. If we have two sets A and B, $A, B \subseteq X$, we define their union, $A \cup B$, as the set

$$A \cup B = \{x \in X : x \in A \text{ or } x \in B\}$$

The right hand side (RHS) above reads as: "all x belonging in X such that x belongs in A or x belongs in B". It means all those x which belong *at least* in either A or B. Hence if x belongs to A, to B, or to both, x belongs to the set $A \cup B$. The intersection $A \cap B$, is the set

$$A \cap B = \{x \in X : x \in A \text{ and } x \in B\}$$

That is, x belongs to the set $A \cap B$ if it belongs to *both* A and B. This operations can be extended to more than two sets. Suppose we have a collection of sets $\{A_i, i \in I\}$ where I is

[1] $\emptyset$ is a subset of all sets. This is because $\emptyset \subseteq A$ means that $x \in \emptyset \implies x \in A$. However, there is nothing in $\emptyset$, so the antecedent in the previous implication ($x \in \emptyset$) is false. Then, the consequence ($x \in A$) is true. If antecedent is false any consequence is true. See `https://math.stackexchange.com/questions/439987/assumed-true-until-proven-false-the-curious-case-of-the-vacuous-truthifinterested`.

some index set, for example the natural numbers. Then

$$\bigcup_{i\in I} A_i = \{x \in X : \exists i \in I \text{ such that } x \in A_i\}$$
$$\bigcap_{i\in I} A_i = \{x \in X : x \in A_i \ \forall i \in I\}$$

Hence, the union above consists in the set of all $x \in X$ which belong to *at least* one of the A_i's while the intersection consists in the set of all $x \in X$ which belong to *all* A_i's.

Let us examine some properties of unions and intersections

Proposition 2.1. *Let A, B and C be subsets of X. Then the following hold*

(i) Commutative law: $A \cup B = B \cup A$ *and* $A \cap B = B \cap A$.

(ii) Associative law: $(A\cup B)\cup C = A\cup(B\cup C) = A\cup B\cup C$ *and* $(A\cap B)\cap C = A\cap(B\cap C) = A \cap B \cap C$.

(iii) Distributive law: $(A\cup B)\cap C = (A\cap C)\cup(B\cap C)$ *and* $(A\cap B)\cup C = (A\cup C)\cap(B\cup C)$.

The proof of Proposition 2.1 follows form the definitions of union and intersection. I encourage you to do them.

Two sets A and B are disjoint if they have no elements in common, that is, if $A \cap B = \emptyset$. Generally, given a family of sets $\mathcal{A} = \{A_i, i \in I\}$, we say that the elements of $\mathcal{A}$ are pairwise disjoint if

$$A_i \cap A_j = \emptyset \ \forall\, i \neq j$$

A partition of X is a class of pairwise disjoint sets in X such that their union is X. Formally, $\mathcal{A} = \{A_i, i \in I\}$ is a partition of X if for all $i \neq j$

$$A_i \cap A_j = \emptyset \text{ and } \bigcup_{i\in I} A_i = X$$

Given two sets A and B, both subsets of X, $A \setminus B$ denotes the set of elements belonging to A and not to B

$$A \setminus B = \{x \in X : x \in A \text{ and } x \notin B\}$$

The complement of a set $A \subset X$, denoted by A^c is the set containing all elements in X which are not in A

$$A^c = \{x \in X : x \notin A\}$$

Note that $A \setminus B = A \cap B^c$. Another important property is the following

Proposition 2.2 (De Morgan's Laws)**.** *Let $\mathcal{A} = \{A_i, i \in I\}$, then*

(i) $\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c$, *and*

(ii) $\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c$.

Proof. Note that if we want to prove that two sets are equal we need to show that they are subsets of each other. For i) suppose that $x \in (\cup_{i\in I} A_i)^c$, then $x \notin \cup_{i\in I} A_i$, then there exists no $i \in I$ such that $x \in A_i$, this implies that $x \in A_i^c$ for all $i \in I$, hence $x \in \cap_{i\in I} A_i^c$. We conclude that $(\cup_{i\in I} A_i)^c \subseteq \cap_{i\in I} A_i^c$. For the other direction suppose that $x \in \cap_{i\in I} A_i^c$, then $x \in A_i^c$ for all $i \in I$, then $x \notin \cup_{i\in I} A_i$ and $x \in (\cup_{i\in I} A_i)^c$. Hence, $\cap_{i\in I} A_i^c \subseteq (\cup_{i\in I} A_i)^c$. We conclude that $\cap_{i\in I} A_i^c = (\cup_{i\in I} A_i)^c$. ii) is left as an exercise. $\square$

Exercise 2.1. *Let $A_1, A_2, ...$ be subsets of X. Define $B_1 = A_1$, $B_2 = A_2 \setminus A_1$,..., $B_k = A_k \setminus \cup_{i=1}^{k-1} A_i$,... Show that $\{B_i\}_{i=1}^{\infty}$ are disjoint and that*

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i \text{ and } \bigcup_{i=1}^{m} A_i = \bigcup_{i=1}^{m} B_i$$

Mappings between sets can be particularly useful. Because of this we introduce the following concepts

Definition 2.1 (Function)**.** *A function $f : A \to B$ is a rule which associates every element in A with a unique element in B.*

Note that a function needs to map every element in A, that is, we cannot leave out any element in A. However we do not need to map to all elements of B. B is called the co-domain and the set of elements of B to which the function maps to is called the range, denoted R, formally

$$R = \{y \in B : \exists x \in A \text{ such that } f(x) = y\}$$

Definition 2.2 (Injective function)**.** *Every element in R has a unique pre-image in A.*

This means that there are not two distinct elements in the domain which map to the same element in B.

Definition 2.3 (Surjective function)**.** $R = B$.

Definition 2.4 (Bijective function)**.** *A function is called bijective if it is both injective and surjective.*

Definition 2.5. *(i) Sets A and B are said to be equicardinal if there exists a bijection $f : A \to B$, $|A| = |B|$,*

(ii) B has cardinality greater than or equal to A if there exists an injective function $f : A \to B$, $|B| \geq |A|$,

(iii) $|B| > |A|$, if there exists an injective function $f : A \to B$, but A and B are not equicardinal.

Definition 2.6. *(i) A set A is said to be countably infinite if it is equicardinal with $\mathbb{N}$.*

(ii) A set A is countable if it is either finite or countably infinite.

Example 2.1. $\mathbb{Q} \cap [0,1]$ *(Rationals in* $[0,1]$*) is a countable set. This means you can form a bijection with* $\mathbb{N}$*, that is, a list enumerating all elements*

$$\{0, 1, 1/2, 1/3, 2/3, 1/4, 3/4, 1/5, 2/5, 3/5, 4/5, ...\}$$

One can also show that a countable union of sets is countable. Hence,

$$\mathbb{Q} = \bigcup_{i \in \mathbb{Z}} \mathbb{Q} \cap [i, i+1]$$

is countable.

Definition 2.7. *A set* A *is uncountable if it has cardinality strictly larger than that of* $\mathbb{N}$*.*

Examples of uncountable sets are $\mathbb{R}, \mathbb{R} \setminus \mathbb{Q}, 2^{\mathbb{N}}, [0,1]$ or $\{0,1\}^{\infty}$ which is the set of all infinite binary strings. Showing that the latter is uncountable is the first step to show the rest are uncountable. The proof is Cantor's diagonal argument which you can find in many textbooks or just in the Wikipedia. The key from this discussion for us is to have a clear notion of countable (finite and infinite) and uncountable sets.

3 Probability spaces

We start from two *undefined* entities. A *Random Experiment* and an *Outcome*. These are to be understood intuitively from their semantic meaning. However, we will give some examples to clarify these notions. Starting from these two entities we will define all the rest.

Definition 3.1 (Sample space)**.** *the sample space* Ω *is the set of all possible outcomes of a random experiment.*

For example, suppose that your random experiment is to toss a coin once. What is the sample space? It depends on what is of interest to you. You determine what the sample space is. If you are interested in which face shows up, then the set of possible outcomes is $\Omega = \{H, T\}$, where H = Heads and T = Tails. However, you might not be interested in which face shows up but in the number of times the coin flips in the air. In this case, the set of all possible outcomes which are of interest to you is $\Omega = \mathbb{N}$. Another possibility is that what is of interest to you is the velocity at which the coin hits the ground. In this case, the sample sample space is $\mathbb{R}^+$. Note that we have given an example of a finite, countably infinite and uncountable sample space.

An (elementary) outcome is denoted by $\omega \in \Omega$. *This* is the source of randomness. You have no control over what ω realizes. You can think about it as $\omega \in \Omega$ being chosen by some Goddess of Chance. Another way of imagining it is as $\omega \in \Omega$ being one of many alternate realities[2]. Suppose that whenever you throw a dice, six alternative realities are created and you do not get to choose in which one you end up. Every time you run the random experiment, an outcome $\omega \in \Omega$ realizes. Another example of a random experiment is to toss

[2]If you like the show Rick and Morty you can think that parallel universe C-137 is the ω which was chosen for the main characters in the show.

a coin n times. If you are interested in the number of faces that show we have $\Omega = \{H, T\}^n$ (finite sample space). Note that we are thinking of this as *one* random experiment and not as n random experiments. We could also think about a random experiment which consists in tossing a coin infinitely many times. Then, $\Omega = \{H, T\}^\infty$ (uncountable sample space). An example of an elementary outcome in this setting is $\omega = \{H, H, T, H, T, T, T, ...\}$. Another random experiment is to throw a dart to the $[0, 1]$ line. Then, $\Omega = [0, 1]$ (uncountable sample space) and an elementary outcome could be $\omega = 0.333$.

Often we are not interested in whether a particular elementary outcome has occurred. We might be interested in whether a *subset* of the sample space has occurred or not. For instance, if your random experiment is tossing a coin once and your sample space is the number of flips in the air ($\Omega = \mathbb{N}$), you might not be interested in the exact number of flips but on whether it flipped more than five times ($\{6, 7, 8, ...\} \subseteq \mathbb{N}$). Or think about the Spanish economy during the last quarter of 2020 as your random experiment and GDP growth as your sample space ($\Omega = \mathbb{R}$). You might not be interested in the exact growth rate but just in whether GDP growth is positive or negative ($\mathbb{R}^+ \subseteq \mathbb{R}$ or $\mathbb{R}^- \subseteq \mathbb{R}$). These kind of subsets of Ω which are of interest to us will be called events (we will give more rigorous definitions in a bit). An event $A \subseteq \Omega$ is said to occur if $\omega \in A$, that is, if the Goddess of Chance has picked an $\omega \in \Omega$ which belongs to A. Importantly, all events are subsets of Ω but not all subsets of Ω are events. For instance, we might not care about GDP growth being a rational number despite the fact that $\mathbb{Q} \subseteq \mathbb{R}$. Ultimately we will want to assign probabilities to events. But first we need to put everything we just said in a more rigorous mathematical scheme. Our goal now is to build a structure of the subsets of Ω to determine what is of interest and what is not of interest, i.e. what is an event and what is not an event. To do this we are going to impose some rules. Intuitively, since Ω always occurs, we should be interested in Ω. Also, if there is some subset A which interests us, we should be also interested in A^c, i.e. in A not occurring. Also, if there are two events A and B which are of interest, we should be interested in at least one of them taking place or in both of them occurring, i.e. $A \cup B$ and $A \cap B$ should be of interest. These rules motivate some mathematical structures of subsets of Ω we will work with.

Definition 3.2 (Algebra*)**.** *Let Ω be the sample space and $\mathcal{F}_0$ be a collection of subsets of Ω. $\mathcal{F}_0$ is called an algebra if*

(i) $\emptyset \in \mathcal{F}_0$,

(ii) if $A \in \mathcal{F}_0$, *then* $A^c \in \mathcal{F}_0$,

(iii) if $A \in \mathcal{F}_0$ *and* $B \in \mathcal{F}_0$, *then* $A \cup B \in \mathcal{F}_0$.

However, one limiting aspect of the algebra is that it is only closed under[3] finite unions (show that it is closed under finite unions). However, we might be interested in a countably infinite union of events. This motivates the definition of a broader structure which will be one of our main tools.

[3]We say that some set is closed under some operation, if you apply this operation to different elements of the set and you get an element of the set. For instance, (ii) can be read as the algebra being closed under complementation.

Definition 3.3 (σ-algebra)**.** *A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra if*

(i) $\emptyset \in \mathcal{F}$,

(ii) if $A \in \mathcal{F}_0$, then $A^c \in \mathcal{F}_0$,

(iii) if $A_i \in \mathcal{F}$, $i = 1, 2, ...$, then $\cup A_i \in \mathcal{F}$.

σ-algebras are closed[4] under complementation and under countable unions. You will encounter different countable union notations, $\cup A_i, \cup_{i=1}^{\infty}, \cup_{i\in\mathbb{N}}$ all are defined as in the mathematical review, that is as the set of elements which belong to at least one of the $A_1, A_2, ...$

Exercise 3.1. *Show that a $\sigma - algebra$ is also closed under countable intersections. Hint: you need (ii) and a property in the mathematical review.*

Now, before we informally called events as those subsets of Ω which are of interest. Formally, an event is an element of the σ-algebra. We will also call the events as $\mathcal{F}$-measurable sets. Note, that $\mathcal{F}$ is a *collection* of sets, hence, its elements are sets. A trivial σ-algebra would be $\mathcal{F} = \{\emptyset, \Omega\}$. Another trivial σ-algebra is $\mathcal{F} = 2^{\Omega}$, that is, all subsets of Ω. If we use the power set as a σ-algebra, we do not lose any information, however, as we will see this will not always be possible[5]. Another example of a σ-algebra is the smallest σ-algebra containing $A \subseteq \Omega$, denoted by $\sigma(A)$, $\sigma(A) = \{\Omega, \emptyset, A, A^c\}$.

Exercise 3.2. *Show that all these examples of σ-algebras are indeed σ-algebras.*

So now we already have some more structure. Specifically, we have a sample space Ω and a σ-algebra defined on it. This tuple $(\Omega, \mathcal{F})$ we call a *measurable space*. It is a space to which a measure can be assigned, if not it would not be measurable, hence the logical step now is to define what a measure is.

Definition 3.4 (Measure)**.** *A measure is a function $\mu : \mathcal{F} \to [0, \infty)$ such that*

(i) $\mu(\emptyset) = 0$,

(ii) if $A_1, A_2, ...$ is a countable collection of $\mathcal{F}$-measurable disjoint sets, then

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

Unsurprisingly, $(\Omega, \mathcal{F}, \mu)$ is called a *measure space.* If $\mu(\Omega) < \infty$, μ is called a finite measure. If there exists a sequence $A_1, A_2, ...$ of subsets of Ω such that $\cup A_i = \Omega$ and $\mu(A_i) < \infty$ for all i, then μ is said to be a σ-finite measure.

Example 3.1 (Counting measure)**.** *Suppose $\Omega = \{a_1, ..., a_n\}$ and that $\mathcal{F}$ is some σ-algebra defined on Ω. Let the counting measure be defined as*

$$\nu(A) = \sum_{i=1}^{n} \delta_{a_i}(A) \quad A \in \mathcal{F},$$

[4] see the footnote above ↑

[5] If Ω is uncountable it will not be feasible to take $\mathcal{F} = 2^{\Omega}$.

where $\delta_{a_i}(A)$ is a set function[6] *which is equal to one if $a_i \in A$ and zero otherwise. Hence it counts how many elementary outcomes are in the $\mathcal{F}$-measurable set A. $(\Omega, \mathcal{F}, \nu)$ is an example of a measure space.*

Example 3.2 (Lebesgue measure)**.** *Suppose $\Omega = \mathbb{R}$ and that we have some σ-algebra $\mathcal{F}$ defined on it which contains closed intervals (i.e. $[a, b] \subseteq \mathbb{R}$)*[7]*. Define the Lebesgue measure as*

$$\lambda([a, b]) = b - a.$$

$(\Omega, \mathcal{F}, \lambda)$ is another example of a measure space.

If $\mu(\Omega) = 1$, μ is called a *probability measure*. Since this is really important let us state a proper definition even though it is almost the same as the definition of a measure.

Definition 3.5 (Probability measure)**.** *A probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that*

(i) $\mathbb{P}(\emptyset) = 0$.

(ii) (Countable additivity) If $A_1, A_2, \ldots$ is a countable collection of $\mathcal{F}$-measurable disjoint sets, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

$(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. In summary, we have introduced the concept of a random experiment, we have defined all its possible outcomes as the sample space Ω, we have defined a collection of subsets of Ω, $\mathcal{F}$, which contains the sets which are of interest, these sets are what we have called events or $\mathcal{F}$-measurable sets, then we have defined $(\Omega, \mathcal{F})$ to be a measure space. Finally, we have assigned a probability measure $\mathbb{P}$ on the measure space which gives the probability of all $\mathcal{F}$-measurable sets and we called $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. Now we introduce some useful properties of probability measures. Note these properties are properties of measures in general for the special case that we deal with a probability measure.

Proposition 3.1 (Properties of Probability Measures)**.**

(i) (Finite additivity) If $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint then

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

Proof. Follows from countable additivity, take a countable sequence of sets $B_1, B_2, \ldots$ such that $B_i = A_i$ if $i \leq 0$ and $B_i = \emptyset$ if $i > 0$. Then $\cup_{i=1}^{\infty} B_i = \cup_{i=1}^{n} A_i$. Hence

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \mathbb{P}(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{i+1}^{\infty} \mathbb{P}(B_i) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

□

[6] Takes a set as an input and gives a number as an output.

[7] We will talk in detail about how such a σ-algebra can be created, for now take it as given.

(ii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(iii) (Monotonicity) If $A \subseteq B$, $A, B \in \mathcal{F}$, *then*

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof. Note that $B = A \cup B \setminus A$. A and $B \setminus A$ are disjoint, so

$$\mathbb{P}(B) = \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0} \implies \mathbb{P}(B) \geq \mathbb{P}(A).$$

□

(iv) If $A_1, ..., A_n \in \mathcal{F}$, *then*

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - ... + (-1)^{n-1}\,\mathbb{P}(\cap_{i=1}^n A_i).$$

Proof. I do it for two sets A and B, the general proof uses induction. We can divide $A \cup B$ into the union of three disjoint sets

$$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B).$$

We can write

$$A = (A \cap B^c) \cup (A \cap B) \implies \mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B)$$
$$B = (B \cap A^c) \cup (A \cap B) \implies \mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B).$$

Hence,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

□

(v) (Continuity) If $A_1, A_2, ... \in \mathcal{F}$, *then*

$$\mathbb{P}(\cup_{i=1}^\infty A_i) = \lim_{m \to \infty} \mathbb{P}(\cup_{i=1}^m A_i)$$

Proof. What this property means is far from obvious. To remember it you can think of it as "taking the limit inside" as you would do with a continuous function. However, that is not really what is going on. It is a really useful property for proofs. Here I give a sketch of the proof. We define $B_1 = A_1$, $B_2 = A_2 \setminus A_1$,..., $B_n = A_n \setminus \cup_{i=1}^{n-1} A_i$,... By Exercise 2.1 we know that the B_i's are disjoint and that $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty B_i$. Hence

$$\mathbb{P}(\cup_{i=1}^\infty A_i) = \mathbb{P}(\cup_{i=1}^\infty B_i) = \sum_{i=1}^\infty \mathbb{P}(B_i) = \lim_{m \to \infty} \sum_{i=1}^m \mathbb{P}(B_i) = \lim_{m \to \infty} \mathbb{P}(\cup_{i=1}^m B_i) = \mathbb{P}(\cup_{i=1}^m A_i).$$

We have used Exercise 2.1, countable additivity and Exercise 2.1, definition of infinite sum, finite additivity, Exercise 2.1 respectively in each equality above. $\square$

(vi) *(Corollaries of continuity, some textbooks define this as continuity) If $A_1 \subset A_2 \subset A_3 ...$ (nested increasing) (or $A_1 \supset A_2 \supset A_3 ...$ (nested decreasing, like Russian dolls)) and $A_1, A_2, ... \in \mathcal{F}$, then*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \lim_{m\to\infty} \mathbb{P}(A_m) \quad \Big(\mathit{or}\ \mathbb{P}(\cap_{i=1}^{\infty} A_i) = \lim_{m\to\infty} \mathbb{P}(A_m) \Big),$$

which can also be written as $\mathbb{P}(\lim_{m\to\infty} A_m) = \lim_{m\to\infty} \mathbb{P}(A_m)$.

(vii) *(Subadditivity) If $A_1, A_2, ... \in \mathcal{F}$, then*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. Again define $B_i = A_i \setminus \cup_{j=1}^{i-1} A_j$. We know that $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$ and that the B_i's are disjoint by Exercise 2.1. Then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \mathbb{P}(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i),$$

since $\mathrm{B}_i \subseteq A_i$, $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ for all i, so

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{P}(B_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) \text{ for all } n \geq 1 \implies & \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \\ \implies & \mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \end{aligned}$$

$\square$

4 Discrete Probability Spaces

If Ω is countable (finite or countably infinite) we can always take the σ-algebra to be $\mathcal{F} = 2^{\Omega}$, the collection of all subsets of Ω. Now, we have a definition of a probability measure and we have a σ-algebra, our task is to assign probabilities. This means that we have to come up with a probability measure which satisfies the definition and assigns probabilities to all events ($\mathcal{F}$-measurable sets) contained in 2^{Ω}.

The probability of each $A \in \mathcal{F}$ is going to be defined in terms of the probabilities of the singleton subsets[8], $\mathbb{P}(\{\omega\})$. For any $A \in \mathcal{F}$ we are gonna assign a probability measure such that

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\{\omega\}) \text{ and } \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = 1.$$

Example 4.1.

[8] A singleton subset is $\{\omega\}$, which is different from an elementary outcome ω. One is a set (even if it contains only one element) and the other is an element.

(i) $\Omega = \{H, T\}$ *and* $\mathcal{F} = 2^{\Omega}$. *We need to assign a probability to each singleton. We can let* $\mathbb{P}(\{H\}) = p$ *and* $\mathbb{P}(\{T\}) = 1 - p$ *for* $0 \leq p \leq 1$.

(ii) $\Omega = \mathbb{N}$ *and* $\mathcal{F} = 2^{\mathbb{N}}$. *We need to assign* $\mathbb{P}(\{k\})$ *for each* $k = 1, 2, \ldots$ *such that* $\sum_{k=1}^{\infty} \mathbb{P}(\{k\}) = 1$. *There are many ways of doing this. One way would be*

$$\mathbb{P}(\{k\}) = \frac{1}{2^k} \text{ for } k \in \mathbb{N},$$

another way

$$\mathbb{P}(\{k\}) = (1-p)^{k-1} p \text{ for } k \in \mathbb{N} \text{ and } 0 \leq p \leq 1.$$

(iii) $\Omega = \{0\} \cup \mathbb{N}$ *and* $\mathcal{F} = 2^{\Omega}$, *one valid probability assignment would be*

$$\mathbb{P}(\{k\}) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots \text{ and } \lambda > 0.$$

These are examples of different ways of assigning probabilities in discrete probability spaces. Intuitively, you can think about this procedure as having sum mass which adds up to one which you have to distribute among a countable number of points.

5 Uncountable Ω

Here we are going to follow a specific motivating example. Suppose that your random experiment consists in throwing a dart to the $(0, 1]$ line. Then your sample space is $\Omega = (0, 1]$. Suppose you want to assign probabilities in such a way that makes it equally likely for you to hit any part of the line.

Now, assigning probabilities to singletons is not going to work. If you assign a probability to some $\{\omega\}$ it cannot be strictly positive. This is because we want to distribute probability uniformly, implying that we would want all singletons to have the same probability. But then if we assign a strictly positive probability everything blows up. This implies that we can only put zero probability to all singletons.

One of the defining properties of a probability measure is that it is countably additive. If we have assigned $\mathbb{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$ and we want to measure an interval, for instance $[1/2, 1]$, we cannot. $[1/2, 1]$ is an *uncountable* union of singletons. The solution to this problem is to forget about singletons altogether and focus on subsets of $\Omega = (0, 1]$ which are of interest. Remember this is a specific motivating example in which we are trying to assign a *uniform* probability measure to $(0, 1]$, that is, a probability measure which tells us that the dart is equally likely to land anywhere. If we want this probability measure to be uniform, there are two conditions it must satisfy (asides from the definition of a probability measure)

(i) (We do not care about singletons condition) For $a, b \in (0, 1]$, $a \leq b$

$$\mathbb{P}((a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b)) = \mathbb{P}([a, b]).$$

(ii) (Translation invariance): we want intervals of the same length to have the same probability (i.e. probability does not change if you move the set around).

Now we state a general (for any measure not only probability measures) impossibility theorem which we will not prove

Theorem 5.1 (Impossibility Theorem)**.** *There exists no measure $\mu(A)$ defined on 2^Ω (i.e. all subsets of $[0,1]$), satisfying (i) and (ii).*

The takeaway is that when Ω is uncountable we cannot pick $\mathcal{F} = 2^\Omega$. We cannot keep all subsets, we need to pick less subsets which means we need to specify a smaller σ-algebra. Hence, we cannot dismiss the question of what is interesting to us just by picking all subsets as we do with discrete probability spaces. What subsets should we pick? Borel and Lebesgue when faced with this problem while developing measure theory found that focusing on intervals is a good solution. Of course the collection of all intervals is not a σ-algebra since the complement of an interval is not necessarily an interval. Hence, we have to somehow generate a σ-algebra which contains all intervals. In the following suppose that our collection of subsets of interest is $\mathcal{C}$ (e.g. collection of intervals). We are going to see how to generate a σ-algebra from $\mathcal{C}$.

5.1 Generated σ-algebras

Let $\mathcal{C}$ be an arbitrary collection of subsets of Ω.

Theorem 5.2. *There exists a unique σ-algebra, say $\sigma(\mathcal{C})$, which is the smallest σ-algebra containing $\mathcal{C}$. That is, if $\mathcal{H}$ is any σ-algebra that contains $\mathcal{C}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{H}$. $\sigma(\mathcal{C})$ is called the σ-algebra generated by $\mathcal{C}$.*

Proof. Let $\{\mathcal{F}_i, i \in I\}$ be the collection of *all* σ-algebras which contain $\mathcal{C}$ (a collection of collections of sets!). Note that this collection will not be empty since 2^Ω will always be there. We can prove that

$$\sigma(\mathcal{C}) = \cap_{i\in I}\mathcal{F}_i.$$

To prove it we need to show three things

(i) $\sigma(\mathcal{C})$ is a σ-algebra (Exercise)

(ii) $\mathcal{C} \subseteq \sigma(\mathcal{C})$. This is true since all σ-algebras in the intersection contain $\mathcal{C}$.

(iii) It is the smallest σ-algebra. To see this, let $\mathcal{H}$ a σ-algebra such that $\mathcal{C} \subseteq \mathcal{H}$, then $\mathcal{H} \in \{\mathcal{F}_i, i \in I\}$ since $\{\mathcal{F}_i, i \in I\}$ is the collection of all σ-algebras which contain $\mathcal{C}$. Hence, there exists an $i \in I$ such that $\mathcal{H} = \mathcal{F}_i$ which implies that $\sigma(\mathcal{C}) \subseteq \mathcal{H}$.

$\square$

Exercise 5.1. *Show that the intersection of σ-algebras are σ-algebras (note this is not true for unions).*

5.2 Borel σ-algebra

Let $\Omega = (0,1]$ and $\mathcal{C}_0$ be the collection of all open intervals of Ω.

Definition 5.1. *$\sigma(\mathcal{C}_0)$ is called the Borel σ-algebra on $(0,1]$, denoted by $\mathcal{B}((0,1])$.*

Definition 5.2. *Elements of $\mathcal{B}((0,1])$ are called Borel-measurable sets, or simply Borel sets.*

Note that $\mathcal{B}((0,1])$ is well-defined by Theorem 5.2. However, nothing we have done tells us that it is not 2^Ω. However, it turns out (it can be shown) that it is much smaller than 2^Ω. It actually has the same cardinality as $\mathbb{R}$. Also, it is quite hard to find sets in 2^Ω which are not in $\mathcal{B}((0,1])$. In sum, the Borel σ-algebra buys us a lot with very little sacrifice. Now we show some useful propositions.

Proposition 5.1. *Let $b \in (0,1]$. Then the singleton $\{b\}$ is a Borel set.*

Proof. It is true because it can be written as a countable intersection of Borel sets

$$\{b\} = \bigcap_{n=1}^{\infty} \left[\left(b - \frac{1}{n}, b + \frac{1}{n} \right) \cap \Omega \right].$$

Since all sets in the intersection contain b, $\{b\}$ is a subset of the intersection. Now, to show that the intersection is a subset of $\{b\}$, take any $c \in (0,1]$ which is different than b. I can find an n_0 such that for all $n \geq n_0$,

$$c \notin \left(b - \frac{1}{n}, b + \frac{1}{n} \right).$$

□

Proposition 5.2. *$(a,b], [a,b], [a,b)$ are Borel sets.*

Proof. $(a,b] = (a,b) \cup \{b\}$ which is a countable union of Borel sets. $[a,b] = \{a\} \cup (a,b) \cup \{b\}$ which is also a countable union of Borel sets. $[a,b) = \{a\} \cup (a,b)$ which is also a countable union of Borel sets. □

Curiosities: there are really weird sets which are Borel sets. For instance, Cantor sets, which have an interesting fractal like behaviour, are Borel sets. An example of a non Borel-measurable set is the Vitali set. These two sets have extensive Wikipedia articles you can check if interested. An interesting paradox which comes out when dealing with non-measurable sets is the Banach-Tarski paradox[9] which states that you can decompose a ball into a finite number of disjoint subsets and then put these subsets back together in a way in which you get two balls identical to the original one. As you might suspect, these disjoint subsets are not Borel sets.

So remember our motivating example was to put a uniform measure on $(0,1]$. Now we are going to work with the measurable space $((0,1], \mathcal{B}((0,1]))$. You can consider the next part to be optional (if you are studying this for the first part in an Econometrics course for

[9] A really cool video with amazing visualizations of the paradox: https://www.youtube.com/watch?v=s86-Z-CbaHA.

instance). What we are going to do is the same as we did with discrete probability spaces. Which is to assign a measure to all sets belonging to the σ-algebra. In our case we are gonna assign a measure to all Borel sets. However, to do this with Borel sets we need a more advanced mathematical machinery. If you skip the next section, all you need to know is that we are going to assign a measure to half-closed intervals $(a, b]$ proportional to their length (i.e. $\mu((a, b]) = b - a$. Then there will be a theorem (which we will not prove) which says that this measure we have put on half-closed intervals extends uniquely to all Borel sets.

5.3 *Caratheodory's Extension

Define $\mathcal{F}_0$ as the collection of subsets of Ω which are finite unions of disjoint intervals of the form $(a, b]$ plus the empty set. A typical element of $\mathcal{F}_0$ is $(a_1, b_1] \cup (a_2, b_2] \cup ... \cup (a_n, b_n]$, where $a_1 < b_1 \leq a_2 < b_2 ... \leq a_n < b_n$. We want a measure of each of these intervals which is proportional to their length (i.e. $b_1 - a_1$). We want to extend this measure to all (including really weird sets) Borel sets. For this we will use Caratheodory's[10] Extension theorem. But first we need a lemma about the collection $\mathcal{F}_0$.

Lemma 5.1.

(i) *$\mathcal{F}_0$ is an algebra.*

Proof. $\Omega \in \mathcal{F}_0$ clearly and finite unions are also elements of $\mathcal{F}_0$, hence it is an algebra. □

(ii) *$\mathcal{F}_0$ is* not *a σ-algebra.*

Proof. Consider

$$A_n = \left(0, \frac{n}{n+1}\right] \quad n = 1, 2, ...$$

Note that $A_n \in \mathcal{F}_0$ for all n. However,

$$\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} \left(0, \frac{n}{n+1}\right] = (0, 1) \notin \mathcal{F}_0$$

□

(iii) *$\sigma(\mathcal{F}_0) = \mathcal{B}((0, 1])$.*

Proof. To show that $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}((0, 1])$, it is enough to show that $\mathcal{F}_0 \subseteq \mathcal{B}((0, 1])$, since this implies that $\sigma(\mathcal{F}_0)$ is a sub-σ-algebra of $\mathcal{B}((0, 1])$ because it is the smallest σ-algebra containing $\mathcal{F}_0$. We have shown that $(a, b] \in \mathcal{B}((0, 1])$ (in Proposition 5.2), so $\mathcal{F}_0 \subseteq \mathcal{B}((0, 1])$ which implies that $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}((0, 1]) = \sigma(\mathcal{C})$.

To show that $\mathcal{B}((0, 1]) \subseteq \sigma(\mathcal{F}_0)$, we proceed as follows

$$(a, b) = \cup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right] \implies \mathcal{C} \subseteq \sigma(\mathcal{F}_0) \implies \sigma(\mathcal{C}) \subseteq \sigma(\mathcal{F}_0).$$

[10] Greek mathematician who lived from 1873 to 1950

□

So we want to define a measure. We will call it $\mathbb{P}$ since it will be a probability measure given that we are working with the zero-one interval. This measure has to be proportional to length of the intervals. Hence, one property that we desire is that for $F = (a_1, b_1] \cup (a_2, b_2] \cup ... \cup (a_n, b_n] \in \mathcal{F}_0$ we have that

$$\mathbb{P}(F) = \sum_{i=1}^{n} (b_i - a_i).$$

However, measures are defined on σ-algebras not just algebras. So we cannot define a measure $\mathbb{P}$ on $\mathcal{F}_0$ and call it a measure. The step from algebra to σ-algebra is given by Caratheodory's Extension theorem which we state but do not prove.

Theorem 5.3 (Caratheodory's Extension Theorem)**.** *Let $\mathcal{F}_0$ be an algebra on Ω and let $\mathcal{F} = \sigma(\mathcal{F}_0)$. Suppose $\mathbb{P}_0 : \mathcal{F}_0 \to [0, 1]$ such that $\mathbb{P}_0(\Omega) = 1$ and $\mathbb{P}_0$ is countably additive in $\mathcal{F}_0$. Then $\mathbb{P}_0$ can be* uniquely *extended to a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. That is, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that*

$$\mathbb{P}(A) = \mathbb{P}_0(A) \text{ for all } A \in \mathcal{F}_0.$$

Note that $\mathbb{P}_0$ is *not* a probability measure. Also, we have stated the theorem for the special case of probability theory, note that the general theorem holds for general measures. When we say that $\mathbb{P}_0$ is countably additive in $\mathcal{F}_0$, we mean that it is countably additive for those countable union in $\mathcal{F}_0$ (since $\mathcal{F}_0$ is an algebra and not a σ-algebra, a countable unions of elements in $\mathcal{F}_0$ need not belong in $\mathcal{F}_0$).

So, back to our motivating example. We have that $\mathcal{F}_0$ defined as collection of finite unions of half-closed intervals plus the empty set is an algebra. We showed that $\sigma(\mathcal{F}_0) = \mathcal{B}((0, 1])$. Let us define $\mathbb{P}_0 : \mathcal{F}_0 \to [0, 1]$ such that $\mathbb{P}_0(\emptyset) = 0$ and $\mathbb{P}_0(F) = \sum_{i=1}^{n}(b_i - a_i)$. Next, we need to verify countable additivity of $\mathbb{P}_0$ in $\mathcal{F}_0$.

Exercise 5.2. *For any $F_1, F_2, ... \in \mathcal{F}_0$ such that $\cup_{i=1}^{\infty} F_i \in \mathcal{F}_0$ we have that $\mathbb{P}_0(\cup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} \mathbb{P}(F_i)$.*

Now, all conditions of Caratheodory's Theorem hold. Then, by the theorem we have that there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{B}((0, 1]))$ which agrees with $\mathbb{P}_0$ on $\mathcal{F}_0$. Which this basically means is that for all Borel sets in $(0, 1]$ I can define a unique probability measure which corresponds to the notion of length. This unique measure corresponds to the notion of length because $\mathbb{P}_0$ *is* length and $\mathbb{P}$ is the unique measure defined on all Borel sets which agrees with $\mathbb{P}_0$. Again, since it is a measure, it is defined for all Borel sets. So even if you give it a weird Borel set (such as a Cantor) it will assign to it a measure which corresponds with length. Basically, if you recall the impossibility theorem, $\mathbb{P}$ is a measure defined on $\mathcal{B}((0, 1])$ (not on 2^{Ω}) which satisfies the two conditions we desired[11]. This measure $\mathbb{P}$ is called the *Lebesgue measure* on $(0, 1]$.

[11] The impossibility theorem stated that there existed not measure in 2^{Ω} which satisfied those two properties. We have basically settled down for a much smaller σ-algebra, the Borel σ-algebra and found a unique measure on this σ-algebra which does satisfy those two conditions.

So we know this Lebesgue measure $\mathbb{P}$ exists and that it is unique, however we only know its explicit form for elements of $\mathcal{F}_0$, not for all Borel sets. Now we are going to see how to construct the Lebesgue measure for some commonly encountered Borel sets which are not in $\mathcal{F}_0$.

5.4 *Lebesgue measure on $(0, 1]$

Example 5.1. *Singleton $\{b\}$:*

$$\mathbb{P}(\{b\}) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} \underbrace{\Big(b - \frac{1}{n}, b + \frac{1}{n}\Big] \cap \Omega}_{\equiv B_n}\Big)$$

Note that $B_1, B_2, \ldots$ are nested decreasing sets! Hence, we can use continuity of the probability measure

$$\mathbb{P}(\{b\}) = \lim_{n\to\infty} \mathbb{P}(B_n) \leq \lim_{n\to\infty} \Big(b + \frac{1}{n} - b + \frac{1}{n}\Big) = \lim_{n\to\infty} \frac{2}{n} = 0.$$

The $\leq$ follows from the fact that B_n is intersected with Ω. One can show that $\mathbb{P}((a, b)) = \mathbb{P}([a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b]) = b - a$.

Example 5.2. $\mathbb{P}($*all rational numbers*$) = \mathbb{P}(\mathbb{Q} \cap \Omega) = 0$. *Note that rationals are Borel sets since $\mathbb{Q}$ is countable and singletons are Borel sets with Lebesgue measure zero. In fact, any countable subset of $(0, 1]$ will have zero probability.*

Typical confusion: the probability of an event being zero *does not* mean it cannot occur. It is an elementary outcome in your sample space so it can occur. Since the sample space is the set of all possible outcomes. As long as the event is not empty it can occur.

Example 5.3. $\mathbb{P}($*irrationals*$) = 1 - \mathbb{P}(\mathbb{Q} \cap \Omega) = 1$. *However this does not mean that irrationals happen for sure! It is not the case that the set of possible outcomes is equal to the set of all irrationals, $\Omega \neq \{$irrationals$\}$. What we say is that an irrational happens almost surely (a.s) or with probability one.*

6 *Lebesgue measure on $\mathbb{R}$

The Lebesgue measure can be defined also for $\mathbb{R}$ and not only for $(0, 1]$. However note that the Lebesgue measure in $\mathbb{R}$ is not a probability measure anymore.

Definition 6.1. *Let $\mathcal{C}_0$ be the collection of all open intervals in $\mathbb{R}$, then $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{C}_0)$.*

Definition 6.2. *Let $\mathcal{D}$ be the collection of semi-infinite intervals*

$$\mathcal{D} = \{(-\infty, x] : x \in \mathbb{R}\}$$

Then, $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{D})$.

We have defined the Borel σ-algebra in $\mathbb{R}$ in two different ways. We should prove that both definitions are equivalent, i.e. $\sigma(\mathcal{C}_0) = \sigma(\mathcal{D})$. The first definition is the one that is usually given. The second one is a more operational definition. But they are indeed equivalent. To construct the Lebesgue measure on $\mathbb{R}$ we have to repeat the exact same story as we did for $(0,1]$. Define an algebra $\mathcal{F}_0$, take a pseudo-measure λ_0 which can be interpreted as length. Check that Caratheodory's theorem holds. Then by the theorem there exists a unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which is the Lebesgue measure on $\mathbb{R}$.

7 Conditional Probability

Let us work now with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let B be an event such that $\mathbb{P}(B) > 0$.

Definition 7.1. *The conditional probability of $A \in \mathcal{F}$ given B is defined as*

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

We cannot condition of zero probability events. For instance, if we throw a dart to the $[0,1]$ line, the question of what is the probability of hitting within the interval $[0,1/2]$ given that the dart has landed on a rational number is not well-defined.

Theorem 7.1. *Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then $\mathbb{P}(. \mid B) : \mathcal{F} \to [0,1]$ is a probability measure on $(\Omega, \mathcal{F})$.*

Proof. We need to show that $\mathbb{P}(. \mid B) \in [0,1]$ and that $\mathbb{P}(\cup_{i=1}\infty A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for disjoint $A_1, A_2, ...$ Note the following

$$\begin{aligned}
\mathbb{P}(\Omega \mid B) &= \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1, \\
\mathbb{P}(\emptyset \mid B) &= \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0, \\
0 \leq \mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leq \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = 1 \text{ for all } A \in \mathcal{F}.
\end{aligned}$$

So $\mathbb{P}(. \mid B) \in [0,1]$. Now take $A_1, A_2, ...$ disjoint, then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i \mid B) = \frac{\mathbb{P}((\cup_{i=1}^{\infty} A_i) \cap B}{\mathbb{P}(B)} = \frac{\mathbb{P}((\cup_{i=1}^{\infty} (A_i \cap B)))}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B).$$

Where we have used the definition of conditional probability, then the distributive law for unions and intersections of sets, the fact that $A_i \cap B$ are disjoint sets and countable additivity, and lastly the definition of conditional probability again. $\square$

Note that we are working with the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ all the time. The random experiment does not change. For instance if the random experiment is to throw a dice, the conditioning event could be $B =$ even number and $A =$ getting a two. Everything is done on the basis of the same random experiment. Conditional probabilities are probabilities so they satisfy the properties of probability measures. Now we define further properties which are specific to conditional probabilities.

Proposition 7.1 (Properties of Conditional Probabilities). *(i) (Law of Total Probability) Let $B_i \in \mathcal{F}$, $i = 1, 2, ...$ be a partition of Ω (i.e. $\cup_{i=1}^{\infty} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$). Then*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i).$$

Proof.

$$\sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \cup B_i) = \mathbb{P}(\cup_{i=1}^{\infty}(A \cap B_i)) = \mathbb{P}(A \cap (\cup_{i=1}^{\infty} B_i)) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A).$$

Where we used the definition of conditional probabilities, the fact that $A \cap B_1$, $A \cap B_2$,... are disjoint, countable additivity and the distributive law. One useful example of this property is: $\mathbb{P}(A) = \mathbb{P}(A \mid B)\, \mathbb{P}(B) + \mathbb{P}(A \mid B^c)\, \mathbb{P}(B^c)$. One common example is the following. Suppose you have 1,2,... urns with red and blue balls. The probability of picking a red ball is the sum of the probabilities of picking a red ball in urn 1, times probability of urn 1, plus the probability of picking a red ball in urn 2 times the probability of urn 2 and so on. ☐

(ii) (Bayes Rule) Let $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ and B_i, $i = 1, 2, ...$ as in (i). Then

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j)}.$$

Proof.

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j)} = \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} = \mathbb{P}(B_i \mid A),$$

where we used property (i). Example: given that you took a red ball what is the probability it came from urn i (posterior). ☐

(iii) Let $A_i \in \mathcal{F}$, $i = 1, 2, ...$ Then

$$\mathbb{P}(\cap_{i=1}^{\infty} A_i) = \mathbb{P}(A_i) \prod_{i=2}^{\infty} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j),$$

with $\mathbb{P}(\cap_{j=1}^{i-1} A_j) > 0$ for all i.

Proof.

$$\begin{aligned}\mathbb{P}(\cap_{i=1}^{\infty} A_i) &= \lim_{n\to\infty} \mathbb{P}(\cap_{i=1}^{n} A_i) = \lim_{n\to\infty} \mathbb{P}(A_1) \prod_{i=2}^{n} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j) \\ &= \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j).\end{aligned}$$

☐

8 Independence

8.1 Independence of events

So we are working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we now give a definition of independence of events ($\mathcal{F}$-measurable sets).

Definition 8.1. *Events A and B are said to be independent under $\mathbb{P}$ (or simply independent when measure $\mathbb{P}$ is unambiguous) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.*

We say independent *under* $\mathbb{P}$, since events might be independent under some probability measure but dependent under another probability measure. This is the definition. It is advisable to forget or not pay much attention to previous intuitions you might have. Things like "they are independent if they have nothing to do with each other" can be misleading. Independence of events is what is stated in the definition, not more, not less. You might have encountered another definition, mainly that A and B are independent if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. This is a consequence of the definition not a definition itself. It adds the requirement that $\mathbb{P}(B) > 0$, which is not required for independence.

Example 8.1. *Statement "A and B are independent if knowing B tells me nothing about A" can be misleading:*

Take $([0,1], \mathcal{B}([0,1]), \mathbb{P})$. Then, $\mathbb{P}(\textit{rational} \cap \textit{irrational}) = \mathbb{P}(\emptyset) = 0$. And also, $\mathbb{P}(\textit{rational}) = 0$ and $\mathbb{P}(\textit{irrational}) = 0$. Hence $\mathbb{P}(\textit{rational} \cap \textit{irrational}) = \mathbb{P}(\textit{rational})\,\mathbb{P}(\textit{irrational})$. So the event that the elementary outcome is a rational number and the event that it is an irrational number are independent. However, if I know the elementary outcome is an irrational number, I know for sure it is not a rational number. The occurrence of one event rules out the other one, still, they are independent. This happens because they have zero probability, if both events had strictly positive probability the intuition would be fine.

Example 8.2. *Can an event be independent from itself? Yes!*

$$\mathbb{P}(A \cap A) = \mathbb{P}(A)\,\mathbb{P}(A) \iff \mathbb{P}(A) = \mathbb{P}(A)^2 \iff \mathbb{P}(A) = 0 \text{ or } \mathbb{P}(A) = 1.$$

Definition 8.2. *$A_1, A_2, \dots, A_n \in \mathcal{F}$ are independent if for all non-empty $I_0 \subseteq \{1, 2, \dots, n\}$, we have*

$$\mathbb{P}(\cap_{i \in I_0} A_i) = \prod_{i \in I_0} \mathbb{P}(A_i)$$

The definition above basically tells you that for any finite collection of events, to say they are independent you need to check all combinations. The following extends this to arbitrary collections of events.

Definition 8.3. *Let $\{A_i, i \in I\}$ be an arbitrary collection of events (index does not need to be countable). These events are said to be independent if for all non-empty and finite $I_0 \subseteq I$, we have*

$$\mathbb{P}(\cap_{i \in I_0} A_i) = \prod_{i \in I_0} \mathbb{P}(A_i)$$

Note that in the definition above there might be infinite non-empty finite subsets of I.

Example 8.3. *Suppose we toss a coin infinite times. Then $\Omega = \{0,1\}^\infty$. Let $\mathcal{F}_k$ be the collection of subsets of Ω whose occurrence can be decided by looking at the first k tosses. For instance, the event that there are at least three heads in the first 15 tosses belongs to $\mathcal{F}_{15}$. You can show $\mathcal{F}_n \subseteq \mathcal{F}_m$ for all $n \leq m$ and that $\mathcal{F}_k$ is a σ-algebra (left as an exercise). Then we can show that A_i and A_j, $i \neq j$ are independent events. Without loss of generality suppose that $j > i$, then $A_i, A_j \in \mathcal{F}_j$. Define the following probability measure on $\mathcal{F}_j$, $\mathbb{P}(A) = |A|/2^j$ (check it is a probability measure). Then*

$$\mathbb{P}(A_i \cap A_j) = \frac{2^{j-2}}{2^j} = \frac{1}{4}, \quad \mathbb{P}(A_i) = \frac{2^{i-1}}{2^i}$$

FINISH

Exercise 8.1. *Consider a sequence of events $A_1, A_2, \ldots$. Show that if A_i's are independent, then A_i^c's are independent as well.*

8.2 Independence of σ-algebras

Remember we are working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. since $\mathbb{P}$ is defined on $\mathcal{F}$, we are going to look at independence between sub-σ-algebras of $\mathcal{F}$. This is because we need $\mathbb{P}$ to be defined to have a notion of independence. So, for instance, if $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ and $\mathcal{F}_1 = \{\emptyset, \Omega\}$, then $\mathcal{F}_1 \subseteq \mathcal{F}$. Or if $\Omega = [0,1]$, $\mathcal{F} = 2^\Omega$ and $\mathcal{F}_1 = \mathcal{B}([0,1])$, then $\mathcal{F}_1 \subseteq \mathcal{F}$.

Definition 8.4. *Two sub-σ-algebras, $\mathcal{F}_1$ and $\mathcal{F}_2$ of $\mathcal{F}$, are said to be independent if for any $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$, A_1 and A_2 are independent.*

Definition 8.5. *Let $\{\mathcal{F}_i, i \in I\}$ be an arbitrary collection of sub-σ-algebras of $\mathcal{F}$ (I might be uncountable). These $\mathcal{F}_i$'s are said to be independent if for any choice of $A_i \in \mathcal{F}_i$, $i \in I$, we have that $\{A_i, i \in I\}$ are independent events*[12].

Comparing all possible events can be extremely difficult. Hence, later we will see that there exists a simpler way of checking independence of sub-σ-algebras. Namely, that if you prove independence in collections of events which are closed under finite intersections (these collections are called π-systems) you are actually done.

9 Borel-Cantelli Lemmas

Now we introduce two important results which will be useful later on.

Lemma 9.1 (First Borel-Cantelli Lemma)**.** *If $A_1, A_2, \ldots$ is a sequence of events such that $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$, then a.s only finitely many A_n's will occur.*

Lemma 9.2 (Second Borel-Cantelli Lemma)**.** *If $A_1, A_2, \ldots$ are independent events such that $\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty$, then a.s infinitely many A_n's will occur.*

[12] As a curiosity: this definition depends on the axiom of choice, which states that for an indexed collection of non-empty sets you can always pick one element of each member of the collection. Even if the index is uncountable. Although this axiom was controversial in its origin, now it is commonly used by most mathematicians and is included in the standard form of axiomatic set theory. Check the wikipedia for more.

Before we prove these Borel-Cantelli (BC) lemmas let us make some comments. First let us formally consider the event that infinitely many A_n's occur. This is

$$\{A_n \ i.o\} \equiv \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i$$

To understand this object, let $B_n = \cup_{i=n}^{\infty} A_i$, then $\{A_n \ i.o\} = \cap_{n=1}^{\infty} B_n$. B_n is the event that at least one of $A_n, A_{n+1}, ...$ occurs, we can call it the n-th tail event. Then, $\{A_n \ i.o\}$ is the intersection of these n-th tail events. The event that all B_n's occur, that is, for all n, B_n occurs. Insisting, this means that for every n, at least one of the A_n's occurs. No matter how big your n is, no matter how far you go, you will have at least one of the A_n's after that n occurring. The first BC lemma says that $\mathbb{P}(\{A_n \ i.o\}) = 0$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. The second BC lemma syas that $\mathbb{P}(\{A_n \ i.o\}) = 1$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and A_n's are independent. Let us do the proofs now.

Proof of first BC lemma. So we need to show that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}(\{A_n \ i.o\}) = 0$. We can show that $\mathbb{P}(\{A_n \ i.o\}^c) = 1$

$$\{A_n \ i.o\}^c = \left(\cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i\right)^c = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i^c,$$

this event is the event that there exists n_0 such that for all $n \geq n_0$, each of the A_n failed to occur. Note that B_n are nested decreasing ($B_1 \supseteq B_2 \supseteq ...$), hence

$$\mathbb{P}(\cap_{n=1} B_n) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}(\cup_{i=n}^{\infty} A_i) \leq \lim_{n\to\infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i) = 0,$$

where we have used continuity of probability measures, subadditivity of probability measures and the fact that if $\sum_{k=1} b_k < \infty$, then the sequence of tail sums $\sum_{n=k}^{\infty} b_n$ converges to zero. It is in this last step where we used that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. $\square$

Proof of second BC lemma. First we prove another lemma which we will need

Lemma. Suppose $0 \leq p_i \leq 1$ is such that $\sum_{i=1}^{\infty} p_i = \infty$. Then $\prod_{i=1}^{\infty}(1 - p_i) = 0$.

Proof. We know that $\ln(1 - x) \leq -x$ for all $x \in [0, 1)$, then

$$\ln \prod_{i=1}^{\infty}(1 - p_i) = \ln\left(\lim_{n\to\infty} \prod_{i=1}^{n}(1 - p_i)\right) = \left(\lim_{n\to\infty} \ln \prod_{i=1}^{n}(1 - p_i)\right) \leq \ln \prod_{i=1}^{k}(1 - p_i).$$

The second equality follows since ln is a continuous function so it can be interchanged with the limit, and the inequality follows since $0 \leq p_i \leq 1$, so if we stop at some $k < \infty$, we will get a larger number. Now

$$\ln \prod_{i=1}^{k}(1 - p_i) = \sum_{i=1}^{k} \ln(1 - p_i) \leq \sum_{i=1}^{k}(-p_i) \text{ for all } k \geq 1.$$

Hence, taking $k \to \infty$, $\ln \prod_{i=1}^{\infty}(1 - p_i) \to -\infty$ which implies that $\prod_{i=1}^{\infty}(1 - p_i) \to 0$. So the lemma is proved.

Now, to show that $\mathbb{P}(\{A_i \ i.o\}) = 1$, note that

$$1 - \mathbb{P}(\{A_i \ i.o\}) = \mathbb{P}(\cup_{n=1}^{\infty} B_n^c) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n^c).$$

We need to show that the above is zero. This amounts to showing that $\mathbb{P}(B_n^c) = 0$ for all $n \geq 1$. Fix n, and $m \geq n$. Then (if A_i's are independent, A_i^c's are independent too, see exercise 8.1)

$$\mathbb{P}(\cap_{i=n}^{m} A_i^c) = \prod_{i=n}^{m} (1 - \mathbb{P}(A_i)),$$

where we have used independence of events. Hence,

$$\mathbb{P}(B_n^c) = \lim_{m \to \infty} \mathbb{P}(\cap_{i=n} m A_i^c) = \prod_{i=n}^{\infty} (1 - \mathbb{P}(A_i)) = 0 \text{ for all } n \geq 1,$$

where we use continuity of probabilities and the lemma we just proved. Therefore, $\mathbb{P}(\{A_i \ i.o\}) = 1$. □

Example 9.1. *Consider $\Omega = \{0,1\}^{\infty}$, for instance the random experiment of tossing a coin infinite times. Suppose we have on it a σ-algebra which (among others) contains events of the form A_i, denoting the event that the i-th toss is heads. Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$ such that $\mathbb{P}(A_n) = 1/n^2$ for all $n \geq 1$ (there might be many probability measures satisfying this). Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, the first BC lemma implies that a.s. only finitely many heads will occur. The idea is that if heads are becoming more and more unlikely* fast, *there exists n_0 after which you do not get anymore heads with probability 1.*

Now suppose $\mathbb{P}(A_n) = 1/n$ for all $n \geq 1$ and that the A_n's are independent. Heads are also becoming more and more unlikely but slower. Since now $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, the second BC lemma implies that a.s. infinitely many heads will occur. That is, there exists no n_0 such that after it you do not get heads anymore with probability one. No matter how big n is, with probability 1 there will be a head in the following tosses. Even though the probability of the head is decreasing at rate n^{-1}! Even if n is one billion, you know that a head will still occur for sure.

Independence in the second BC lemma is sufficient but not necessary. For this reason, there are many more BC lemma's which relax independence in different ways. An example in which there is not "enough independence" for the second BC lemma to hold is the following.

Example 9.2. *Suppose $A_n = E$ for all $n \geq 1$ and that $\mathbb{P}(E) \in (0,1)$. Then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ however, it will not be the case that A_n will occur infinitely often with probability one since $\mathbb{P}(\{A_n \ i.o.\}) = \mathbb{P}(E)$. This happens because of the strong dependence of the A_n's.*

10 Measurable functions and Integration

In this section we will talk about functions between measure spaces and their integration. For now we are going to forget about probabilities and keep a more general measure theoretic approach. However, I will hint at what is coming, which is to define random variables and their expectations. This will just be a special case of what we cover here. Suppose you have two measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$. We are going to work with functions $f : \Omega \to \Lambda$. One key aspect we require from these functions is that they are measurable.

Definition 10.1 (Measurable function)**.** *A function $f : \Omega \to \Lambda$ is said to be $\mathcal{F}$-measurable if for every $G \in \mathcal{G}$, the pre-image is $\mathcal{F}$-measurable, $f^{-1}(G) \in \mathcal{F}$, where $f^{-1}(G) = \{\omega \in \Omega : f(\omega) \in G\}$.*

So take any $\mathcal{G}$-measurable set, $G \in \mathcal{G}$. The set of all $\omega \in \Omega$ which map to this set G under the function $f(.)$, is the pre-image of G under f, $f^{-1}(G)$. This subset of Ω need not be an element of $\mathcal{F}$. If it is, and this happens for all possible $G \in \mathcal{G}$, then $f(.)$ is a measurable function. As we will see and state properly in the next section, a random variable X is just going to be a measurable function from Ω to $\mathbb{R}$.

11 Random variables

11.1 Definition and c.d.f.

11.2 Discrete Random Variables

11.3 Continuous Random Variables

Here goes the Radon-Nikodym derivative

11.4 σ-algebras generated by random variables

11.5 Several Random variables

11.6 Independent Random variables

12 Transformation of Random Variables

13 Conditional Expectation

14 Moment Generating function and Characteristic function

15 Concentration Inequalities

Concentration inequalities give you probability bounds on r.v.s taking values in some range. This why they are called concentration inequalities, since it is about probability concentrating in some range.

Proposition 15.1 (Markov's Inequality)**.** *If X is a non-negative r.v. with $\mathbb{E}[X] < \infty$, then for any $\alpha > 0$,*

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof.

$$\mathbb{E}[X] = \underbrace{\mathbb{E}[X\mathbb{1}(X \le \alpha)]}_{\ge 0} + \mathbb{E}[X\mathbb{1}(X > \alpha)] \ge \mathbb{E}[X\mathbb{1}(X > \alpha] \ge \alpha\,\mathbb{E}[\mathbb{1}(X > \alpha)] = \alpha\,\mathbb{P}(X > \alpha).$$

□

Note that this is only useful for $\alpha > \mathbb{E}[X]$, otherwise the RHS is above 1. This is a very loose inequality. It just says that the probability of being above α decays at rate $1/\alpha$. However, this probability often decays much faster. If we add further assumptions we can get faster rates. For instance, assuming that finite variance (or finite second moment) we get the following

Proposition 15.2 (Chebyshev's Inequality). *If X is a r.v. with mean μ and variance $\sigma^2 < \infty$, then for any $\alpha > 0$*

$$\mathbb{P}(|X - \mu| > \alpha) \le \frac{\sigma^2}{\alpha}$$

Alternatively, we can write the result as $\mathbb{P}(|X - \mu| > \alpha\sigma) \le 1/\alpha^2$, which tells us that the probability that $|X - \mu|$ is α times the standard deviation decays at rate $1/\alpha^2$. The proof follows from applying the Markov Inequality to the r.v. $|X - \mu|^2$. Still, Chebyshev's Inequality is not very sharp. Assuming the existence of the expectation of increasing transformations of X increases our decay rates.

Proposition 15.3 (Extended Markov Inequality). *If $\varphi(\cdot)$ is a monotone increasing non-negative function on the positive reals, X is a r.v. with $\mathbb{E}[\varphi(X)] < \infty$, then*

$$\mathbb{P}(X \ge a) \le \frac{\mathbb{E}[\varphi(X)]}{\varphi(a)}.$$

The proof follows from the fact that $\mathbb{P}(X \ge a) = \mathbb{P}(\varphi(X) \ge \varphi(a))$ and from applying Markov's Inequality to r.v. $\varphi(X)$. This result allows us to find much sharper bounds. For instance, under assumptions of the moment generating function $\varphi_X(t) = \mathbb{E}[e^{tX}]$ we can get exponential decays.

Proposition 15.4 (Chernoff bound). *FINISH*

16 Convergence of Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_1, X_2, \ldots$ be a sequence of r.v.s.

Definition 16.1 (Pointwise or sure convergence). *The sequence $\{X_n, n = 1, 2, \ldots\}$ is said to converge* pointwise *or* surely *to X, $X_n \to_{pw} X$, if*

$$X_n(\omega) \to X(\omega) \text{ for all } \omega \in \Omega \text{ as } n \to \infty.$$

Note that for a fixed $\omega \in \Omega$, $\{Xn(\omega), n = 1, 2, \ldots\}$ is just a sequence of real numbers. Remember that that a sequence of real numbers $\{a_n, n = 1, 2, \ldots\}$ is said to converge to a, written as $a_n \to a$ or $\lim_{n\to\infty} a_n = a$, iff for any $\varepsilon > 0$, there exists $N \ge 1$ such that for all

$n > N$, $|a_n - a| < \varepsilon$. In fact, $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ is *exactly* pointwise convergence of functions[13]. Most of the times this is a very strong notion of convergence. This is because it also requires convergence in sets with probability 0. Hence, it is rarely used. Note that a technicality which we do not show here is that for the definition to make sense one would have to show that the limit of a sequence of measurable functions is measurable, i.e. if $X_1, X_2, ...$ are r.v.s so is X. Now we define a convergence notion which allows convergence not to happen in probability zero events.

Definition 16.2 (Almost sure or wp1 convergence)**.** *X_n converges to X almost surely or wp1, $X_n \to_{a.s.} X$, if $X_n(\omega) \to X(\omega)$ on a set of probability 1, that is*

$$\mathbb{P}\Big(\{\omega : X_n(\omega) \to X(\omega)\}\Big) = 1.$$

Sometimes this is also called strong convergence. Here there is also a technicality, for the above to make sense we need $\{\omega : X_n(\omega) \to X(\omega)\}$ to be an event (i.e. to be measurable, to be an element of $\mathcal{F}$, etc...). To do this one can show that the event is a countable union/intersection of events, we do not do it here but the reader is encouraged to do it. We now define a weaker concept of convergence which is widely used.

Definition 16.3 (Convergence in probability)**.** *X_n converges in probability to X, $X_n \to_p X$, if for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \varepsilon\Big) = 0.$$

Note that a.s. convergence and convergence in probability are very different. In a.s. convergence, the limit is inside the probability while in convergence in probability the limit is outside of the probability. In fact, it could be said that convergence of X_n in probability is not the most suitable name for this concept since it is not really X_n that is converging but a sequences of probabilities. In fact, there might be sets with probability strictly greater than 0 in which $X_n(\omega)$ and $X(\omega)$ are not "close". What we really have is that the sequence $\mathbb{P}_n(\varepsilon) \to 0$ where $\mathbb{P}_n(\varepsilon) \equiv \mathbb{P}(|X_n - X| > \varepsilon)$. In contrast, a.s. convergence does refer to convergence of the sequence $X_1, X_2, ...$ We will later show that a.s. convergence implies convergence in probability but not the other way around, meaning a.s. convergence is a stronger concept of convergence.

Definition 16.4 (Convergence in r-th mean)**.** *X_n converges to X in r-th mean, $X_n \to_{r-th} X$, if*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

For $r = 2$, X_n is said to converge in mean-squared sense.

Definition 16.5 (Convergence in distribution)**.** *X_n converges to X in distribution, $X_n \to_d X$ or $X_n \rightsquigarrow X$, if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \textit{ for all } x \textit{ where } F_X(x) \textit{ is continuous.}$$

[13] A sequence of functions $\{f_n, n = 1, 2, ...\}$ with domain D and codomain C is said to converge pointwise to some function $f : D \mapsto C$ iff $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in D$.

Convergence in distribution is also called weak convergence since, as we will show it is the weakest form of convergence. Again, this is really convergence of a sequence of c.d.f.s not really convergence of a sequence of r.v.s, that is X_n and X might be far but have close c.d.fs. Now we state the hierarchy between these convergence notions and in the proofs it will be clear what we mean when we say that for some of them we do not require X_n and X to be "close".

Proposition 16.1 (Hierarchy of Convergence modes)**.** *Do GRAPH.*

1. *$X_n \to_{pw} X$ implies $X_n \to_{a.s.} X$,*
2. *$X_n \to_{a.s.} X$ implies $X_n \to_p X$,*
3. *$X_n \to_p X$ implies $X_n \to_d X$*
4. *$X_n \to_{r-th} X$ for $r \geq 1$ implies $X_n \to_p X$*

Any other relationship does not hold.

To prove the above proposition we need to prove three implications asides from sure convergence implying a.s. convergence which is direct. We also have to provide five counterexamples. So let's start doing this in turn.

Proof: $X_n \to_{r-th} X$ implies $X_n \to_p X$ for $r \geq 1$. We use Markov's Inequality:

$$\mathbb{P}\Big(|X_n - X| > \varepsilon\Big) = \mathbb{P}\Big(|X_n - X|^r > \varepsilon^r\Big) \leq \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r},$$

so

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \varepsilon\Big) \leq \lim_{n\to\infty} \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r} = 0$$

□

Proof: $X_n \to_p X$ implies $X_n \to_d X$. Note that

$$\begin{aligned} F_{X_n}(x) = \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \\ &\leq F_X(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

Where the inequality comes from the fact that $\{X_n \leq x\} \cap \{X \leq x+\varepsilon\} \subseteq \{X \leq x+\varepsilon\}$ and that $\{X_n \leq x\} \cap \{X > x+\varepsilon\} \subseteq \{|X_n - X| > \varepsilon\}$. Similarly, we can show that $F_X(x-\varepsilon) \leq F_{X_n}(x) + \mathbb{P}(|X - X_n| > \varepsilon)$. Putting the two inequalities together we get

$$F_X(x-\varepsilon) - \mathbb{P}(|X - X_n| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

As $n \to \infty$ the above becomes

$$F_X(x-\varepsilon) \leq F_{X_n}(x) \leq F_X(x+\varepsilon).$$

So, by sending $\varepsilon \to 0$ we get that $F_{X_n}(x) \to F_X(x)$ if F_X is continuous at x. □

17 Law of Large Numbers

18 Central Limit Theorem

References

Georg Cantor. On a property of the class of all real algebraic numbers. *Crelle's Journal for Mathematics*, 77:258–262, 1874.

Angel De la Fuente. *Mathematical methods and models for economists.* Cambridge University Press, 2000.

Andrei Nikolaevich Kolmogorov and Sergeui Vasilevich Fomin. *Introductory real analysis.* Courier Corporation, 1975.