# Introduction to probability theory[1]

A course instructed by Amir Yehudayoff, Department of Mathematics, Technion-IIT
Partially follows notes of Yoav Benyamini

[1] An apology: this text probably contains errors. The intention is the every chapter will roughly take 3 hours.

# Contents

# Chapter 1

# Probability spaces

## 1.1 Introduction

This is a mathematical course, so we shall see definitions, theorems and proofs, but this theory comes with many good examples from real life, physics, economics, etc. Throughout the course, you are encouraged to seek such examples, which will also improve your understanding of the definitions, and your ability to prove theorems.

The basic object we shall study is a *probability space*, which is a model for some system that we are interested in, like a container full of gas in physics, or a market in economics.

One of the objectives of the course, and you will need to challenge yourself to achieve this goal, is to be able to correctly apply the mathematical theory we shall discuss in examples from reality. This ability is crucial in many disciplines, from natural science to economics and social sciences.

Let us start with a simple example. Consider a system consisting of one fair coin. The system has 2 states. Each state has a probability $1/2$ associated with it. We model this system by a probability space.

The definition is slightly more abstract for reasons we shall understand later on.

**Definition 1** (Probability space)**.** *A probability space is a tuple* $(\Omega, \mathcal{F}, \Pr)$ *so that*

- *The sample space* $\Omega$ *is a set.*
- *The set of events* $\mathcal{F} \subseteq 2^{\Omega}$ *is a* $\sigma$*-field (a.k.a.* $\sigma$*-algebra).*
- *The probability function/measure* $\Pr : \mathcal{F} \to [0, 1]$ *is a normalised and* $\sigma$*-additive map.*

*Properties* $2, 3$ *are defined below.*

Intuition:

- The outcome of an experiment is an element $\omega$ in $\Omega$.
- The elements of $\mathcal{F}$ are called events, or measurable sets. They correspond to the type of questions we can ask about the system: for every $F \in \mathcal{F}$, the question "is $\omega \in F$?" may be answered. On one hand, $\mathcal{F}$ should be reach enough so that it contains all questions of interest for us. On the other hand, we shall see that $\mathcal{F}$ can not be too general as then the overall structure collapses.
- The value $\Pr(F)$ corresponds to the probability/chance that the answer to this question is "yes."

## 1.2 Examples

Let us consider several examples first. Later we complete the formal proof.

*A fair die.* In this case, the set of possible outcomes is $\Omega = \{1, 2, 3, 4, 5, 6\} = [6]$. The set of possible questions is all subsets of $\Omega$

$$\mathcal{F} = 2^{\Omega} = \{F \subseteq \Omega\}.$$

The number of events is

$$2^{|\Omega|} = 2^6.$$

The probability of $\{1\}$ is $1/6$. So is $\Pr(\{i\})$ for all $i \in \Omega$. The probability of $\{1, 6\}$ is $1/3$. In general, the probability of $F \in \mathcal{F}$ is

$$\Pr(F) = \sum_{i \in F} \Pr(\{i\}) = \frac{|F|}{|\Omega|} = \frac{|F|}{6}.$$

Specifically, $\Pr(\emptyset) = 0$ and $\Pr(\Omega) = 1$.

*Two dice.* The sample space is

$$\Omega = [6] \times [6] = \{(i, j) : i, j \in [6]\}.$$

The set of events is

$$\mathcal{F} = 2^{\Omega}.$$

The probability of $(1, 1)$ is $1/36$. The probability of

$$E = \{(i, j) : i + j = 8\} = \{(1, 7), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2), (7, 1)\}$$

is

$$\Pr(E) = \frac{7}{36}.$$

In general,

$$\Pr(F) = \frac{|F|}{|\Omega|}.$$

*A random dart.* Here the goal is to model a choice of a unifrom random point in a domain $D$. Consider e.g. the unit circle $D = \{e^{2\pi\theta i} : \theta \in \mathbb{R}\} \subset \mathbb{C}$. What is the sample space?

$$\Omega = D.$$

What is the set of events? We can ask "what is probability of landing in right half of circle?" and similar questions. More general, there is a question for every arc. But there are many more events (two disjoint arcs, a point, ...). We shall not define $\mathcal{F}$ formally for now.

What is the probablity function? We shall not define formally as well (we do not know what is $\mathcal{F}$), but if $E$ is an arc then we set $\Pr(E)$ as the length of the arc divided by length of the circle $2\pi$.

## 1.3 Completing the definition

Having a few examples to keep in mind, we provide a formal definition of $\mathcal{F}, \Pr$.

**Definition 2** (Set of events)**.** *The set $\mathcal{F} \subseteq 2^\Omega$ is a $\sigma$-field, that is, it satisfies the following properties*

- $\emptyset \in \mathcal{F}$.
- *If $F \in \mathcal{F}$ then $F^c = \Omega \setminus F \in \mathcal{F}$.*
- *If $F_1, F_2, \ldots \in F$ then*

$$\bigcup_{i=1}^{\infty} F_i, \bigcap_{i=1}^{\infty} F_i \in F.$$

This property of $\mathcal{F}$ is not so important when $\Omega$ is finite, but it is when $\Omega$ is infinite.

The second property corresponds to saying that every yes/no question can also be a no/yes question.

The third property is related to that questions $Q_1, Q_2$ can give new questions e.g. "is both answers are "yes"?" which corresponds to intersection (infinitely many intersections allows the questions to be more informative).

**Definition 3** (Probability measure/function)**.** *The function $\Pr : \mathcal{F} \to [0, 1]$ is*

- *normalized:* $\Pr(\Omega) = 1$.
- *$\sigma$-additive: For all $F_1, F_2, \ldots \in \mathcal{F}$ that are pairwise disjoint (i.e. $F_i \cap F_j = \emptyset$ for all $i \neq j$),*

$$\Pr\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \Pr(F_i).$$

A function Pr is called additive if it satisfies the second property for finitely many sets. Additivity is a weaker restriction than $\sigma$-additivity.

First simple conclusions and intuition about $\sigma$-additivity:

**Lemma 4.** *If $F_1 \subseteq F_2$ in $\mathcal{F}$ then $\Pr(F_1) \leq \Pr(F_2)$.*

*Proof.*

$$\Pr(F_2) = \Pr(F_1 \cup (F_2 \setminus F_1)) = \Pr(F_1) + \Pr(F_2 \setminus F_1) \geq \Pr(F_1).$$

Explain to yourself the validity of the equalities/inequalities from definitions. □

**Lemma 5.** *If $F_1, F_2, \ldots \in \mathcal{F}$ are pairwise disjoint then $\lim_{i\to\infty} \Pr(F_i) = 0$.*

*Proof.* Assume towards a contradiction that there are infinitely many $i_1, i_2, \ldots$ so that $\Pr(F_{i_j}) \geq \epsilon$ for all $j$ for some $\epsilon > 0$. Then, for all $N \in \mathbb{N}$, using previous lemma,

$$1 = \Pr(\Omega) \geq \Pr\left(\bigcup_{j=1}^{N} F_{i_j}\right) = \sum_{j=1}^{N} \Pr(F_{i_j}) \geq \epsilon N.$$

This is a contradiction for $N > 1/\epsilon$. □

**Lemma 6** (Inclusion-exclusion)**.** *Let[1] $F_1, \ldots, F_n \in \mathcal{F}$. Then,*

$$\Pr\left(\bigcup_{i\in[n]} F_i\right) + \sum_{S\subseteq[n]:S\neq\emptyset} (-1)^{|S|} \Pr\left(\bigcap_{i\in S} F_i\right) = 0.$$

*Proof.* An exercise. Observe similarity to $\sum_{S\subseteq[n]}(-1)^{|S|} = 0$. □

## 1.4 Discrete spaces

Let $\Omega$ be countable (or finite). Let $(p_i : i \in \Omega)$ be non negative real numbers so that $\sum_{i\in\Omega} p_i = 1$. Define

$$\mathcal{F} = 2^{\Omega}.$$

[1]There is no infinite analog: If $F_i = F$ for all integer $i$ then the left hand side does not absolutely converge.

Define for $F \in \mathcal{F}$,

$$\Pr(F) = \sum_{i \in F} p_i.$$

**Claim 7.** *The above is a probability space.*

*Proof.* An exercise. □

This is a generic construction of discrete probability spaces (i.e. when $\Omega$ is countable). The probability functions is first defined over the atoms, elements of $\Omega$. The set of events and probability function is then generically defined. It can be verified that every discrete (namely $\Pr(F) > 0$ for all $F$, and with some regularity assumptions on $\mathcal{F}$) probability space can be defined in this way.

## 1.5 σ-fields?

In the dart example, $\Omega$ is the unit circle. What is the underlying $\mathcal{F}$? It turns out that there are several reasonable definitions for $\mathcal{F}$. One example is call the Borel $\sigma$-field (the minimal one containing all arcs). We shall not discuss it in detail here, it will be deeply addressed in the course on real functional analysis. But roughly $\mathcal{F}$ contains all arcs, and is closed under complements and countable unions and intersections.

Well, why can't we just take $\mathcal{F}$ to be all subsets of $\Omega$? It turns out that such a definition does not make sense (in some sense). Consider the dart example from before.

Let $\Omega$ be the unit circle

$$\Omega = \{e^{i\theta} : \theta \in \mathbb{R}\} \subset \mathbb{C}.$$

Assume we wish to define a uniform distribution on $\Omega$. In that case, if $F \in \mathcal{F}$ and $F' \in \mathcal{F}$ is a rotation of $F$, then $\Pr(F) = \Pr(F')$. We call such a function Pr invariant under rotation.

**Lemma 8.** *There is no $\sigma$-additive normalized function that is invariant under rotation and defined over all subsets of $\Omega$.*

This lemma shows that there is no consistent way to define a probability measure even in this simple case, if we do not restricted Pr to be defined only on some of the subsets.

*Proof.* We shall see that for every Pr that is $\sigma$-additive, normalized and invariant, there is a "non measurable" set[2]. Assume towards a contradiction that such a function exists.

[2]Assuming the axiom of choice.

Define a relation $\sim$ on $\Omega$ by $\theta \sim \phi$ iff

$$e^{2\pi i\theta} = e^{2\pi i(\phi+\sqrt{2}z)}$$

for some $z \in \mathbb{Z}$. Verify that it is an equivalence relation. We thus get a partition of $\Omega$ to equivalence classes. For every such class $C$, let $\theta_C$ be a choice of representative of $C$. Consider the set of representatives

$$F_0 = \{\theta_C\} \subset \Omega.$$

Denote by $F_z$ the rotation of $F_0$ by angle $\sqrt{2} \cdot 2\pi z$.

First, by definition, $\bigcup_z F_z = \Omega$.

Second, the sets $F_0, F_1, F_{-1}, \ldots$ are pairwise disjoint. Indeed, for all $\theta$,

$$e^{2\pi i\theta} \neq e^{2\pi i(\theta+q\sqrt{2})}$$

for every rational $q \in \mathbb{Q}$, since otherwise $q\sqrt{2} = n$ for some $n \in \mathbb{N}$, but $\sqrt{2}$ is not rational.

So, if $\Pr(F_0) = 0$ then

$$0 = \sum_z \Pr(F_z) = \Pr(\Omega) = 1,$$

which is a contradiction. However, if $\Pr(F_0) > 0$ then

$$\infty = \sum_z \Pr(F_z) = \Pr\left(\bigcup_z F_z\right) = \Pr(\Omega) = 1,$$

which is also a contradiction. □

We have seen a need to choose $\mathcal{F}$ carefully enough so that the overall structure makes sense on one hand and useful on other hand.

*Summary:* We have defined a probability space $(\Omega, \mathcal{F}, \Pr)$, seen some examples of simple spaces, and provided some ideas on why this definition is needed.

# Chapter 2

# Independence and conditioning

A probability space is a model for a process or an experiment. We would like to have mathematical tools and notions to formally study and understand such objects. We now introduce two such basic notions (which distinguish it from general measure theory).

## 2.1 Independence

The idea of statistical independence is natural. When we toss a coin twice, experience tells us that the second outcome has nothing to do with the first outcome. The changes of the value of a house in China seems independent of that of a house in Peru. We now provide the formal meaning to this concept.

**Definition 9.** *The events $F_1, F_2, \ldots \in \mathcal{F}$ are independent if for every finite non empty subset $I \subset \mathbb{N}$,*

$$\Pr\left(\bigcap_{i\in I} F_i\right) = \prod_{i\in I} \Pr(F_i).$$

Let us consider some examples:

*Two coins.* Consider two tosses of a fair coin. Choose

$$\Omega = \{0,1\}^2,$$

$\mathcal{F} = 2^\Omega$ and define for all $\omega \in \Omega$,

$$\Pr(\omega) = \Pr(\{\omega\}) = \frac{1}{4}.$$

(Here and in the future we shall abuse notation and replace a singleton set by the element in contains.)

Define $E_1$ as the event that the first coin is 1,

$$E_1 = \{(1,0),(1,1)\}.$$

Define $E_2$ as the event that the second coin is 1. The observation is that $E_1, E_2$ are independent:

$$\Pr(E_1) = \Pr(E_2) = \frac{1}{2},$$

and

$$\Pr(E_1 \cap E_2) = \Pr(\{(1,1)\}) = \frac{1}{4}.$$

Define $E_3$ as the event that the number of 1s is even, $E_3 = \{(1,1),(0,0)\}$. Are $E_1, E_3$ independent? Less clear. Check

$$\Pr(E_1 \cap E_3) = \Pr(\{(1,1)\}) = \frac{1}{4},$$

and indeed they are independent.

*Exercise.* Consider the experiment of tossing two fair dice. What is the probability space? We say that an event $E \subset \mathcal{F}$ depends only on the first die if for any $(i,j) \in \mathcal{F}$ we have $(i,j') \in \mathcal{F}$ for all $j'$. Prove that if $E_1$ depends only on the first die, and $E_2$ depends only the second die, then $E_1, E_2$ are independent.

*Pairwise versus general independence.* Let $\Omega$ be all vectors in $\{0,1\}^3$ with even number of ones. $|\Omega| = 4$. Consider the uniform distribution on $\Omega$: for all $x = (x_1, x_2, x_3) \in \Omega$,

$$\Pr(x) = \frac{1}{4}.$$

Denote by $E_i$, $i \in \{1,2,3\}$, the event that the $i$'th bit is one, that is,

$$E_i = \{x \in \Omega : x_i = 1\}.$$

Thus,

$$\Pr(E_i) = \frac{1}{2}.$$

For every $i \neq j$, the two events $E_i, E_j$ are independent. Are $E_1, E_2, E_3$ independent? Well, no.

$$E_1 \cap E_2 \cap E_3 = \emptyset.$$

The fact that there are pairwise independent distributions that are not fully independent is helpful in derandomization of randomized algorithms.

*Geometric meaning.* In the future we shall discuss geometric meaning, but we need more definitions for that.

*Summary.* Saw definition of independence, and some simple examples. Meaning will become clearer in future.

## 2.2 Conditioning

Here we provide a formal meaning to the idea of obtaining more information about the world, or looking at specific part of it.

What is the probability that the height of the first person we meet is less that 1.5 meters? How does our estimate change if we are told that we are standing at the entrance to a kindergarden?

**Definition 10.** *Let $B \in \mathcal{F}$ be so that $\Pr(B) > 0$. The probability of $A \in \mathcal{F}$ conditioned on $B$ is defined as*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

The semantics of $\Pr(A|B)$ is the probability of $A$ when we are guaranteed that $B$ has happened.

Examples

*Independence.* If $A, B$ are independent with $\Pr(B) > 0$, then

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)\Pr(B)}{\Pr(B)} = \Pr(A).$$

In words, knowledge of $B$ does not change the perception of $A$.

*A fair die.* Consider one toss of a fair die. Denote by $A$ the event that the outcome is 4, and by $B$ the event that the outcome is at least 4.

$$\Pr(A|B) = \frac{1/6}{1/2} = \frac{1}{3},$$

and

$$\Pr(B|A) = \frac{1/6}{1/6} = 1.$$

*Medical exams I.* There is a syndrom, let us call it the X syndrom. There are two medical tests $A, B$ that provide statistical information about the possibility of having syndrom X. A person gets the following results: Test $A$ says that his chance of having X is 1 chance in 100. Test $B$ says that his chance of having X is 1 chance in $1,000$. It seems that most likely this person does not have X. From $A, B$ we can get an even more accurate answer by combining the two results together. It turns out that combining the two tests actually yield that the person's chance of having X is 1 in 2!

How can this be? The basic reason is that $X$ may be small in $A$, and $X$ may be small in $B$, but $X$ is large in $A \cap B$. Consider an example...

The theory allows to formally argue. We introduce two general helpful tools.

### 2.2.1 Bayes' rule

A useful tool in this mental exercise is a simple equality called Bayes' rule.

**Theorem 11** (Bayes'). *If $A, B \in \mathcal{F}$ have positive probabilities then*

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A).$$

The importance of this rule, as we shall see in the test example, is that it allows us to move from conditioning on $A$ to conditioning on $B$, and vice versa.

*Proof.* By definition. □

Bayes' rule: Think of $H$ as an hypothesis to be learnt. The quantity $\Pr(H)$ is the prior estimate for the probabilty of $H$. After observation an experiment $E$, we have learnt something, and we get to posterior probability $\Pr(H|E)$. The number $\Pr(E)$ indicates how probable it is to see outcome $E$, and the smaller it is, the larger the impact (the more the posterior information be differ from the prior one).

Let us briefly discuss Bayesian inference. Assume we are observing some system. What determines the behvior of the system is a parameter $\theta_0$. We do not know what the "real" value of $\theta_0$ is, but we have some idea concerning its disribution. The prior disribution is $p(\theta_0)$. Now, we have run some experiment, which does not tell us in general what $\theta_0$ is, but provides some information $x$. Our understanding of the system tells us what $\Pr(x|\theta)$ for all $\theta$ is. (Example: $\theta_0$ is uniform in $[0, 1/n, 2/n, \ldots, 1]$ and the measure $x$ is a coin with bias $\theta$ which we do not know and want to learn.)

We have a prior distribution

$$p_0(\theta) = \Pr(\theta).$$

After a measurement $x_1$, we get a posterior distribution

$$p_1(\theta) = \frac{\Pr(x_1|\theta)\Pr(\theta)}{\Pr(x)}$$

where $\Pr(x) = \sum_{\theta'} P(x|\theta')p_0(\theta')$. This choice is of course inspired by Bayes' theorem. This is a new distribution on the possible values of $\theta$, which is "closer to the truth" than the prior one. We can keep going an eventually we get a distribution that is highly concerntrated at $\theta_0$, which is the underlying parameter we are interested in.

In general it is useful in many statistical applications. Some examples are learning theory, statistical predictions, spam filters, and more.

### 2.2.2 Law of complete probability

Another useful tool in the law of complete probability.

**Theorem 12** (Law of complete probabilty)**.** *Let $B_1, B_2, \ldots \in \mathcal{F}$ be a partition*[1] *of $\Omega$ so that $\Pr(B_i) > 0$ for all $i$. For all $A \in \mathcal{F}$,*

$$\Pr(A) = \sum_i \Pr(A|B_i) \Pr(B_i).$$

The importance of this law is that it allows to break down the perhaps complicated computation of $\Pr(A)$ to simpler parts.

*Proof.* Use the $\sigma$-additivity of Pr and that $A \cap B_1, A \cap B_2, \ldots$ is a partition of $A$. □

Example: *A lottery.* There is a shop that gives 100 USD to a random person that arrives on a certain day. Assume that the probability that $k \geq 0$ people arrive besides me is $2^{-k-1}$. (Note that $\sum_k 2^{-k-1} = 1$.) What is the probability that I win?

(What is the probablity space?)

Denote by $A$ the event that I win. Denote by $B_k$ the event that $k$ people arrived besides me. Thus,

$$\Pr(A|B_k) = 1/(k+1).$$

The events $B_k$ partition $\Omega$. So,

$$\Pr(A) = \sum_k 2^{-k-1}/(k+1) = \ln(2) \approx 0.69.$$

(To compute this sum, use that the derivative of $\sum_{k\geq 1} x^k/k$ is a geometric sum, which we know as a function of $x$.)

Does it make sense? I win with probability roughly 70 percent? What about the second person to come? How can he win with probability 70 percent as well?

### 2.2.3 One more example

*Medical exam II.* There is some illness. People can be either healthy or sick. There is a medical test for illness. The test can either say + or −, where + should be interpreted as positive or sick. Thus,

$$\Omega = \{h, s\} \times \{+, -\}.$$

[1] That is, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $\Omega = \bigcup_i B_i$.

Think of meaning of each of the 4 options.

There are the healthy people

$$H = \{(h, +), (h, -)\},$$

and the sick people $S = \Omega \setminus H$. There are people who got positive answer

$$P = \{(h, +), (s, +)\},$$

and the people who got negative answer $N = \Omega \setminus P$.

Most people are healthy

$$\Pr(H) = 0.99.$$

The test is pretty accurate

$$\Pr(P|H) = 0.01, \ \Pr(P|S) = 0.99.$$

What is the probability that one is sick, given that her test was positive? What is

$$\Pr(S|P)?$$

It seems as if the exam is pretty accurate, so a positive answer should mean that given a positive answer the person is most likely sick.

We know $\Pr(P|S)$ and we want to compute $\Pr(S|P)$. Bayes' rule implies

$$\Pr(S|P) = \frac{\Pr(P|S)\Pr(S)}{\Pr(P)}$$

and the law of complete probability

$$\Pr(P) = \Pr(P|S)\Pr(S) + \Pr(P|H)\Pr(H) = 0.99 \cdot 0.01 + 0.01 \cdot 0.99.$$

The answer is thus

$$\Pr(S|P) = \frac{1}{2}.$$

So although the test greatly improves the evidence that the person is sick, it is still far from fully confirming it. A high level reason for this phenomenon is that the accuracy of the test is roughly 0.01 while it is supposed to distinguish a rare event that happens 0.01 of the time.

*Summary.* Defined independence and conditioning. Gave examples for applications of these notions.

# Chapter 3

# Bernoulli trials

There are many specific examples of probability spaces that are applicable to daily scenarios. We shall consider the example of a glass cups factory. Assume this factory manufactures $n = 1,000$ cups a day, but the probability that a cup is broken in the process is $p = 0.01$.

How many cups does the factory actually produce? There are $n$ cups roughly $p$ of them are broken so overall roughly $pn = 100$ cups are broken and $(1-p)n = 9,900$ cups are good. This is still just a rough estimate. What is probability that exactly $k = 100$ cups are broken? Well, it is

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

But what is this number? Later on we shall see how to analyse it pretty accurately, but for now we just give an estimate.

**Definition 13.** *A $(n, p)$ Bernoulli or Binomial space is defined by $\Omega = \{0, 1, \ldots, n\}$, by $\mathcal{F} = 2^\Omega$, and by*

$$\Pr(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

*for every $k \in \Omega$.*

**Theorem 14.** *Let $\delta > 0$. The probability of the event*

$$E = \{k \in \Omega : |k - pn| \geq \delta n\}$$

*in a $(n, p)$ Bernoulli space is at most*

$$\Pr(E) \leq \frac{2}{\delta^2 n}.$$

*Specifically, when $n \to \infty$, this probability tends to 0.*

This theorem is a special case of the weak law of large numbers, which says that the average of many independent samples tends to be close to what one expects. We shall address it later in the course. We also note that the upper bound on $\Pr(E)$ is far from sharp, but is a first evidence of this phenomenon.

We shall not prove the theorem here, because we shall see a more general proof later on. But we provide some directions for a possible proof. We have

$$\Pr(E) = \sum_{0 \le k < (p-\delta)n} b_k + \sum_{(p+\delta)n < k \le n} b_k$$

where

$$b_k = \binom{n}{k} p^k (1-p)^{n-k}.$$

There are several ways to continue. One is to show that say for $k < (p-\delta)n$ the ratio $p_k/p_{k+1}$ is small (and the opposite for large $k$). Another is to estimate $p_k$ using Stirling's approximation.

Before seeing some applications of this theorem, let us describe a different space in which Bernoulli trials can reside. Let $\Omega = \{0,1\}^n$, let $\mathcal{F} = 2^\Omega$, and let

$$\Pr(x) = p^{|x|}(1-p)^{n-|x|}$$

for $x \in \Omega$ where $|x|$ is the number of ones in $x$. This space is a refinement of the previous space. It provides more information about the process, i.e., it says exactly which trials failed instead of just counting the number of failing trials. This examples shows that there is no uniqueness in defining a space for a given scenario.

## 3.1 Approximation theory

Here we provide Bernstein's proof of a classical theorem of Weirstrass, about approximating continuous functions by polynomials.

**Theorem 15.** *For every continuous $f : [0,1] \to \mathbb{R}$, there is a sequence of polynomials $f_1, f_2, \ldots$ that uniformly converge to $f$. That is, for all $\epsilon > 0$, there is $N$ so that for all $n \ge N$ we have $|f(x) - f_n(x)| < \epsilon$ for all $x \in [0,1]$.*

Bernstein in fact explicitly described how to approximate $f$.

*Proof.* Define $f_n(x)$ as the "average[1] of $f$ with respect to a Bernolli $(n, x)$ experiment"

$$f_n(x) = \sum_{k=0}^{n} f(k/n)\binom{n}{k}x^k(1-x)^{n-k}.$$

For example $f_0(x) = f(0)$, $f_1(x) = (1-x)f(0)+xf(1)$ and $f_2(x) = (1-x)^2f(0)+2x(1-x)f(1/2)+x^2f(1)$.

Let $\epsilon > 0$. Choose $\delta > 0$ so that $|f(x) - f(y)| < \epsilon/2$ for all $|x - y| < \delta$. Let $M = \max\{|f(x)| : x \in [0, 1]\}$.

Estimate, since $\sum_k \binom{n}{k}x^k(1-x)^{n-k} = 1$,

$$\begin{aligned}
|f(x) - f_n(x)| &\le \left|\sum_k (f(x) - f(k/n))\binom{n}{k}x^k(1-x)^{n-k}\right| \\
&\le \left|\sum_{k:|x-k/n|<\delta} (f(x) - f(k/n))\binom{n}{k}x^k(1-x)^{n-k}\right| \\
&+ \sum_{k:|x-k/n|\ge\delta} (|f(x)| + |f(k/n)|)\binom{n}{k}x^k(1-x)^{n-k} \\
&:= I + II.
\end{aligned}$$

To bound $I$, use choice of $\delta$,

$$|I| \le \sum_k (\epsilon/2)x^k(1-x)^{n-k} = \epsilon/2.$$

To bound $II$, use the weak law of large numbers for a $(n, x)$ Bernoulli,

$$|II| \le 2M\frac{2}{\delta^2 n} \le \epsilon/2,$$

as long as $N$ is large enough (it is important here that the bound from weak law for Bernouli trials does not depend on $x$). □

## 3.2 Normal numbers

Here we consider an infinite variant of Bernoulli trials, and its connection to "numerically nice" numbers. This short discussion is less formal than usual.

[1] We shall later give a formal meaning to it.

*Normal numbers in base 2.* Every number $x \in [0,1]$ can be represented in binary as $x = 0.x_1x_2\ldots$ where $x_i \in \{0,1\}$ for all $i$. Define the fraction of ones in $x$ up to $n$ as

$$\rho_n(x) = \frac{|\{i \le n : x_i = 1\}|}{n}.$$

A number is called normal if

$$\lim_{n\to\infty} \rho_n(x) = \frac{1}{2}$$

(specifically the limit exists). That is, roughly half of its coordinates are 1 and half are 0. The number $0.01010101\ldots$ is normal.

*The probability space.* Let $\Omega = \{0,1\}^{\mathbb{N}}$. That is, there is an infinite sequence of experiments. All events are generated from subsets of the form

$$F = \{x \in \Omega : x_1 = a_1, x_2 = a_2, \ldots, x_m = a_m\}$$

for some $a_1, \ldots, a_m \in \{0,1\}$. (This is the product topology, and such events are cylinders.) The probability of such an $F$ is

$$\Pr(F) = (1/2)^{|a|}(1/2)^{m-|a|}.$$

It can be shown that Pr can be extended to all of $\mathcal{F}$.

Almost every $\omega \in \Omega$ corresponds to a unique $x \in [0,1]$ as above. The only exceptions are rational numbers, but since there are only countably many of them, their total probability mass is 0. Indeed, for every $\omega \in \Omega$ it holds that $\Pr(\omega) = 0$, and $\Pr(\mathbb{Q}) = 0$ by $\sigma$ additivity.

Let

$$A_i = \{x \in \Omega : x_i = 1\}.$$

The set $A_1$ corresponds to the left half of $[0,1]$, the set $A_2$ corresponds to two quarters, and so forth. Each $A_i$ is a union of $2^{i-1}$ dyadic intervals. It can be verified that $A_1, A_2, \ldots$ are independent (as events).

The theorem above says that for all $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left(\{x \in [0,1] : |\rho_n(x) - 1/2| < \delta\}\right) = 1.$$

This does not formally mean that the measure of normal numbers is 1, or that the measure of numbers that are not normal is 0. This stronger statement turns out to be true:

$$\Pr(\{x \in [0,1] : x \text{ is normal in base } 2\}) = 1,$$

and follows from the strong law of large number which we discuss later on.

*Normal numbers.* A normal number is a number that is normal in every base $b$. Since there are countably many bases, $\sigma$ additivity actually implies that

$$\Pr(\{x \in [0,1] : x \text{ is normal}\}) = 1.$$

There are very few explicit examples of numbers that were proven to be normal, and it is believed that $\sqrt{2}$ or $\pi$ are normal, but no proof is known.

# Chapter 4

# Random variables

We now discuss one of the basic and useful notions in this theory. Random variables represent measurement of the system, like its temprature, pressure, price, etc. We have already seen one such measurement: How many coin tosses were tails?

**Definition 16.** *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is a random variable if for every interval $I \in \mathbb{R}$ the set $X^{-1}(I)$ is an event in $\mathcal{F}$.*

In other words, a random variable is a measurable real valued function on $\Omega$. The reason for the measurability is that we want to be able to answer questions of the form "what is the chance that $X$ is positive?"

A random variable is a generalisation of the notion of event. Why? Given an event $A \in \mathcal{F}$, we can define the random variable that is 1 on $A$ and 0 elsewhere. It is called the characteristic variable of $A$, and denoted $\mathbf{1}_A$.

**Definition 17.** *A random variables $X$ is discrete if for every $x \in X(\Omega)$ we have $\Pr(X = x) > 0$.*

Such random variables usually count something, like number of successes, people, storms.

Every random variable over a discrete space is discrete. But there are non discrete spaces that have interesting discrete random variables. For example, a store when many people arrive at random times, but an interesting quantity is how many people arrived in a single day.

**Claim 18.** *If $X$ is discrete, then $X(\Omega)$ is countable.*

*Proof.* The size of $\{x : \Pr(X = x) \geq 1/n\}$ is at most $n$. □

We shall often treat a discrete random variable over the probability space it defines $(\Omega_X, \mathcal{F}, \Pr)$ where $\Omega_X$ is the countable set of values $X$ attains of $\mathbb{R}$ and $\mathcal{F}$ with $2^{\Omega_X}$.

**Binomial.** An $(n,p)$-Binomial random variable, with $p \in [0,1]$ and $n \in \mathbb{N}$, takes values in $\{0,1,\ldots,n\}$ and

$$\Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for all $k \in \{0,1,\ldots,n\}$. We denote this by $X \sim Bin(n,p)$.

Most random variables come with a story. The story behind this one is: It counts the number of successes in $n$ independent experiments, where the probability of success in each is $p$.

## 4.1 Operations

**Claim 19.** *The following operations yield random variables: sums, products, limits, supremum, infimum.*

*Proof.* Let us consider one case for example. If $X_1, X_2, \ldots$ is a sequence of random variables and $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ exists for all $\omega$ then

$$\{\omega : X(\omega) < a\} = \bigcup_{N\in\mathbb{N}} \bigcup_{q\in\mathbb{Q}:q<a} \bigcap_{n>N} \{\omega : X_n(\omega) < q\}.$$

This $\supseteq$ is easier: If there is $N$ and $q < a$ so that for all $n > N$ we have $X_n(\omega) < q$ then $X(\omega) \le q < a$.

In the other direction, if $X(\omega) < a$ then $X(\omega) = a - r$ for some $r > 0$. Let $q$ be so that $a - r < q < a - r/2$. And so $X_n(\omega) < q$ for all $n$ large enough. □

## 4.2 More examples

**Geometric.** We write $X \sim Geom(p)$ if $X$ takes values in $\{1,2,3,\ldots\}$ and for every $k$ in this set

$$\Pr(X=k) = (1-p)^{k-1}p.$$

It marks the first head in an infinite sequence of random independent coin tosses, each is head with probability (w.p.) $p$. Check normalization. Graph.

Let us see what is the probability of $\{X > k\}$. (This is how we denote events from now on.)

$$\Pr(X>k) = \sum_{\ell=k+1}^{\infty} (1-p)^{\ell-1}p = (1-p)^k \sum_{\ell=1}^{\infty} p(1-p)^{\ell-1} = (1-p)^k.$$

**Claim 20** (Memoryless)**.** *If $X \sim Geom(p)$ then it is memoryless, that is, for all $k, \ell$ positive natural numbers,*

$$\Pr(X > \ell + k | X > \ell) = \Pr(X > k).$$

*This property charaterizes all geometric distributions (if $X$ takes values in $\{1, 2, 3, \ldots\}$).*

In words, if in the first $\ell$ experiments we did not get head, then the next experiments are independent.

*Proof.* First,

$$\begin{aligned}\Pr(X > \ell + k | X > \ell) &= \frac{\Pr(X > \ell + k, X > \ell)}{\Pr(X > \ell)} \\ &= \frac{\Pr(X > \ell + k)}{\Pr(X > \ell)} \\ &= \frac{p^{\ell+k}}{p^\ell} = p^k.\end{aligned}$$

The other direction follows by induction since $\Pr(X > k+1) = \Pr(X > k)\Pr(X > 1)$ for all $k \geq 0$, which implies what $\Pr(X = k)$ is (as a function of $\Pr(X > 1) = (1 - p)$). □

**Poisson.** We write $X \sim Pois(\lambda)$ for $\lambda > 0$ if for all $k \in \{0, 1, 2, \ldots\}$,

$$\Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

It counts the number of phone calls received in an office, or the number of electronic a radioactive particle emits in a minute. The number $\lambda$ indicates how many calls arrive on average. Normalization. Graph.

It may be obtained as a limit. Partition the time interval $[0, 1]$ to $n$ interval of equal length. Let $X_i$ for $i \in [n]$ be the number of calls that arrived in the $i$'th interval. Assume that $n$ is large enough so that w.h.p. no $X_i$ is more than one (a simplifying assumption). What should be $\Pr(X_i = 1)$ so that $\sum_i X_i$ will have average $\lambda$? It should be $\lambda/n$, as for binomials. In fact, $X$ is close to $Bin(n, \lambda/n)$. (Shall not formally define.) Now, for

fixed $k$,

$$\begin{aligned}\Pr(X_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{1}{k!}\frac{n(n-1)\dots(n-k+1)}{n^k}\lambda^k\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k} \\ &\to_{n\to\infty} \frac{1}{k!}\lambda^k e^{-\lambda}.\end{aligned}$$

It is part of a more general family of Poisson processes, which have nice properties. For every every time interval $I = [t_1, t_2]$, we can assign a variable $X_I$ with distribution $Pois(\lambda|I|)$ that measures how many things occurred in $I$. With the property that if $I \cap I' = \emptyset$ then $X_I, X_{I'}$ are independent (have not formally defined yet).

We have seen 3 examples. There are many more (e.g. negative binomial) but we shall not discuss them for now.

## 4.3 Cumulative distribution function

Given a random variable $X$, we can plot a function that completely describes it.

**Definition 21.** *The cumulative distribution function of $X$ denoted by $F_X : \mathbb{R} \to \mathbb{R}$ is $F_X(t) = \Pr(X \le t)$.*

Graph.

Examples:

**Geometric.** If $X \sim Geo(p)$ then $\Pr(X \le k) = 1 - \Pr(X > k) = (1-p)^k$ for every integer $k$. For $t$ not integer, fill accordingly.

**Discrete.** If $X$ is discrete then

$$F_X(t) = \sum_{k \le t} \Pr(X = k).$$

These sums are not always easy to calculate or estimate.

Exercise: Order $\mathbb{Q}$ as $a_1, a_2, \dots$. For $q = a_n \in \mathbb{Q}$, define

$$\Pr(X = q) = 2^{-n}.$$

This is a probability distribution. What is $F_X$? It is not constant in every interval. The only points of continuity are the irrationals.

**Theorem 22.** *The function $F = F_X$ satisfies the following properties.*

- $F(t) \to 0$ *at* $-\infty$.
- $F(t) \to 1$ *at* $\infty$.
- $F$ *is monotone non decreasing.*
- $F$ *is right continuous.*
- $\Pr(X = t) = F(t) - \lim_{s \to t^-} F(s)$.

We shall not write a formal proof, but discuss reasons.

**Claim 23.** *If $F$ satisfies the above 5 properties then $F$ is the cumulative function of some random variable.*

*Proof.* Choose $\Omega = \mathbb{R}$, $\mathcal{F}$ Borel, and $\Pr((-\infty, t]) = F(t)$. Choose $X(\omega) = \omega$. It remains to verify... □

## 4.4 Continuous random variables

We have seen discrete random variables. Their cumulative distribution function (CDF) $F$ is a step function.

What if $F$ is continuous? Then $X$ is called continuous.

What if $F$ is differentiable?

**Definition 24.** *A random variable $X$ is called (absolutely) continuous if there is a function $f_X : \mathbb{R} \to \mathbb{R}$ so that*[1]

$$F_X(t) = \int_{-\infty}^{t} f_X(s)ds.$$

*The function $f_X$ is called the density of $X$.*

$F$ is differentiable where $f$ is continuous, and then $F' = f$.

For every event $A \subseteq \mathbb{R}$,

$$\Pr(X \in A) = \int_{s \in A} f_X(s)ds.$$

Specifically,

$$\int_{-\infty}^{\infty} f_X(s)ds = 1$$

[1] This is Lebesgue integral which we do not discuss in detail in this course.

and w.l.o.g. for all $s$,

$$f(s) \geq 0.$$

**Examples.**

*Uniform.* A random variable $X$ is uniform in the interval $[a, b]$ if its density is $f_X(s) =?$. It should model a random point in the line, or the position of a random dart. Need to be positive, constant, and with integral 1 on $[a, b]$. So $f_X(s) = 1/(b - a)$ if $s \in [a, b]$ and is zero elsewhere. What is CDF? For $t \leq a$, it is zero. For $t \geq b$, it is one. For $t \in [a, b]$, it is

$$F_X(t) = (t - a)/(b - a).$$

We denote this by $X \sim U[a, b]$.

*Normal/gaussian.* $X$ is gaussian with parameters $\mu, \sigma$ is defined by

$$f_X(s) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}}.$$

The graph of $f_X$ is a bell shaped curved with maximum at $\mu$ and width $\sigma$. What is formal meaning of width? It is the distance between transition points between concave and convex.

This CDF has no closed form. But what is $F_X(\mu)$? Since $f_X$ is symmetric around $\mu$, $F_X(\mu) = 1/2$. In general, to know $F_X(s)$ one needs to look at a table (there are many online).

Normalization? It is difficult to compute the integral of $f_X$ directly, but the square of the integral is easy and gives 1, using change of variables $(x, y) \to (r, \theta)$.

*Exponential.* $X \sim Exp(\mu)$ is defined by

$$f_X(s) = \mu e^{-\mu s}$$

for $s \geq 0$ and 0 elsewhere. So,

$$F_X(t) = \int_0^t \mu e^{-\mu s} ds = 1 - e^{-\mu t}.$$

It is a model for a random alarm clock that on average starts ringing at time $\mu$.

**Claim 25.** *An exponential random variable is memoryless, that is, for all $s, t \geq 0$,*

$$\Pr(X > s + t | x > t) = \Pr(X > s).$$

This is similar to geometric but for a r.v. that takes values in all of $\mathbb{R}$.

There is a tight connection between exponential r.v.'s and Poisson. Recall Poisson is number of calls in a minute (say). Exponential represents the time the call actually

arrived. So, if $X_1, X_2, \ldots$ are independent (haven't formally defined yet) $Exp(\mu)$ r.v.'s then what is $Pois(\lambda)$? The number of phone calls, that is,

$$\min\{n \in \mathbb{N} : X_1 + X_2 + \ldots + X_n > 1\} - 1$$

is Possion. What is $\lambda$? The first clock rings on average at time $\mu$. So, $\lambda$ should be $1/\mu$. This turns out to be true, but we shall not prove formally for now.

## 4.5 Mixed random variables

What if $F$ is not continuous nor a step function?

Example: Imagine a car that reaches a traffic light at a random time in $[0, 1]$. In $[0, 1/2)$ the light is red, and in $[1/2, 1]$ it is green. Denote by $T$ this the car crosses the light. So, $T \in [1/2, 1]$. What is $\Pr(T = 1/2)$? Well $\{T = 1/2\}$ if the car reaches in $[0, 1/2]$, so $\Pr(T = 1/2) = 1/2$. And what is $\Pr(T = 3/4)$? Well it is zero. So $F_T$ is a step at $1/2$ and then it is continuous (draw).

We can generically decompose a r.v. to a discrete part and a continuous part.

**Theorem 26.** *Let $F_X$ be a CDF of a random variable $X$. Then there are $Z, Y$ with $Z$ discrete and $Y$ continuous so that*

$$F_X = \alpha F_Z + (1 - \alpha) F_Y$$

*with $\alpha \in [0, 1]$.*

In words, a general random variable is a convex combination of a discrete and continuous random variables (not necessarily absolutely continuous!).

One can think of choosing $X$ as follows. Let $B \sim Ber(\alpha)$. Let $Y, Z$ be two independent r.v.'s that are distributed correctly. If $B = 1$ then $X = Z$ and if $B = 0$ then $X = Y$.

*Proof.* Let $a_1, a_2, \ldots$ be the points of discontinuity of $F_X$. (Why are there countably many?) Let $Z$ take values in $a_1, a_2, \ldots$, where

$$\Pr(Z = a_i) = c(\Pr(X \leq a_i) - \Pr(X < a_i))$$

where $c \geq 0$ is normalisation constant. What is $c$? It is one over

$$\alpha = \sum_i \Pr(Z = a_i).$$

It is the total sum of jumps. (If $\alpha = 0$ then $X$ is continuous.)

What is $Y$? Want $Y$ to be continuous.

$$(1-\alpha)F_Y = F_X - \alpha F_Z$$

so $F_Y$ is continuous because by construction all of the jumps of $F_X$ were cancelled. And $F_Y$ tends to 1 at infinity since $F_X$ to 1 and $\alpha F_Z$ to $\alpha$. □

## 4.6 Functions of random variables

Example: Let $X \sim U[0,1]$. Denote by $Y$ the area of a square with side length $X$. It is also a random variable $Y = X^2$. How is $Y$ distributed? It also takes values in $[0,1]$. But for $t \in [0,1]$,

$$\Pr(Y \leq t) = \Pr(X^2 \leq t) = \Pr(X \leq \sqrt{t}) = \sqrt{t}.$$

So $Y$ is not uniform in $[0,1]$. It is more concentrated around 0 than around 1. The map $x \to x^2$ shrinks around 0.

This construction works in general. If $X$ is a r.v. and $h$ is a measurable function then $h(X)$ is a also a r.v. Here measurable means that $h^{-1}((-\infty, t])$ is in the Borel $\sigma$-field for all $t$.

**Theorem 27.** *If $h$ is measurable and strictly increasing then*

$$F_{h(X)}(t) = \Pr(h(X) \leq t) = \Pr(X \leq h^{-1}(t)) = F_X(h^{-1}(t)).$$

*So if $X$ has density $f_X$ and $h$ is differentiable then*

$$f_{h(X)}(t) = f_X(h^{-1}(t))(h^{-1})'(t) = f_X(h^{-1}(t))(h^{-1})'(t).$$

*If $h$ is decreasing, there is a minus sign.*

What if $h$ is not monotone or invertible? Just partition to parts in which is invertible. Or even better, analyze $F_Y$ first, and then compute $f_Y$ if needed.

For example, let $X \sim [-1,2]$ and $Y = X^2$. Then $Y$ is supported on $[0,4]$. For every $t \in [0,4]$,

$$\Pr(Y \leq t) = \Pr(-\sqrt{t} \leq X \leq \sqrt{t}).$$

For $t \in [0,1]$ this is

$$\frac{2\sqrt{t}}{3},$$

and for $t \in [1,3]$ this is

$$\frac{1+\sqrt{t}}{3}.$$

To know what is $f_Y$ take derivative of $F_Y$. The value $f_Y(1)$ is of no importance.

## 4.7 Random vectors

What if we have more that one r.v., and we want to understand the joint distribution of several of them?

**Definition 28.** *A random vector is a map $X : \Omega \to \mathbb{R}^n$ so that each of its coordinates is a random variable (i.e. measurable).*

As for random variables, we will mostly ignore the underlying probability space.
Example: Let $X_1 \sim N(0,1)$. Let $X_2 = 2X_1$. Then $X = (X_1, X_2)$ is a random vector.
We can define a CDF by

$$F_X(t) = \Pr(X_1 \le t_1, X_2 \le t_2, \ldots, X_n \le t_n).$$

It has similar properties to that of random variables. In two dimension, it is probability of $X$ being in an "infinite box with vertex at $t$."

A discrete random vector takes countably many values, and then

$$F_X(t) = \sum_{i_1 \le t_1, \ldots, i_n \le t_n} \Pr(X_1 = i_1, \ldots, X_n = i_n).$$

How is $X_1$ distributed? What is $F_{X_1}$? Well,

$$F_{X_1}(t_1) = \Pr(X_1 \le t_1) = \lim_{t_2, \ldots, t_n \to \infty} F_X(t).$$

A similar statement holds for all $i \in [n]$, and in fact for all subset $I$ of $[n]$.

What about probability function of $X_1$?

$$\Pr(X_1 = t_1) = \sum_{t_2, \ldots, t_n} \Pr(X = t).$$

That is, the marginal probability functions are sums or integrals over the global functions.

If $X = (X_1, X_2)$ has density $f_X$ then the density of $X_1$ is

$$f_{X_1}(t_1) = \int_{t_2} f_X(t) dt_2.$$

Example, a uniform point in unit ball in plane.

If $f(t_1, t_2)$ is continuous and

$$F(t_1, t_2) = \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} f(s_1, s_2) ds_1 ds_2$$

is its integral then

$$f = \frac{\partial^2 F}{\partial t_1 \partial t_2}.$$

Comment: the data of $X$ determines that of $X_1, X_2$, but the other direction does not hold. E.g., $X_1 \sim Ber(p)$ and $X_2 = X_1$, is very different than when $X_1, X_2$ are independent (which we now formally define).

## 4.8 Independence

We now define the formal meaning of two random variables being independent. The intuition is that the value of one does not tell us anything about the value of the other. The formal definition is more elaborate.

Given a random variable $X$ over $(\Omega, \mathcal{F}, \Pr)$, the $\sigma$-filed $\mathcal{F}_X$ it generates is defined as the set of all $X^{-1}(I)$ where $I$ is a Borel subset of $\mathbb{R}$.

Given a random vector $(X, Y)$ over $(\Omega, \mathcal{F}, \Pr)$, the $\sigma$-field $\mathcal{F}_{X,Y}$ it generates is the set of all $(X, Y)^{-1}(I)$ where $I$ is a Borel subset of $\mathbb{R}^2$.

For example, consider $\Omega = \mathbb{R}^2$ with $\mathcal{F}$ Borel, and $X_1(\omega_1, \omega_2) = \omega_1$. Then $\mathcal{F}$ is collection of all measurable subsets of plane, and $\mathcal{F}_{X_1}$ is cylinders. Similarly define $X_2(\omega) = \omega_2$. The intersection of a set in $\mathcal{F}_{X_1}$ and a set in $\mathcal{F}_{X_2}$ is a "rectangle."

**Definition 29.** *Let $\mathcal{F}_1, \mathcal{F}_2$ be two sub $\sigma$-fields of $\mathcal{F}$. They are called independent over $(\Omega, \mathcal{F}, \Pr)$ if every $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$ are independent. The two random variables $X_1, X_2$ are called independent if $\mathcal{F}_{X_1}, \mathcal{F}_{X_2}$ are independent. Similarly define for several $\sigma$-fields and r.v.'s.*

**Claim 30.** *If $X = (X_1, X_2)$ is discrete then $X_1, X_2$ are independent iff $\Pr(X = t) = \Pr(X_1 = t_1) \Pr(X_2 = t_2)$ for all $t = (t_1, t_2)$.*

Example: If $X$ is a uniform random point in a finite domain $D \subset \mathbb{Z}^2$ then $X_1, X_2$ are independent iff $D$ is an axis parallel rectangle. Indeed, if $(x_1, x_2)$ and $(x_1', x_2')$ are in $D$ then it follows that $(x_1, x_2')$ is in $D$. Draw in plane the 3 points. This means that $D$ is a rectangle (a set of form $A \times B$).

**Claim 31.** *$X_1, X_2$ are independent iff $F_X = F_{X_1} F_{X_2}$ with $X = (X_1, X_2)$.*

*Explanation.* If $X_1, X_2$ are independent then for all $t = (t_1, t_2)$,

$$\Pr(X_1 \le t_1, X_2 \le t_2) = \Pr(X_1 \le t_1) \Pr(X_2 \le t_2).$$

In the other direction, let us give an example. Consider the event $\{a_1 < X_1 \le b_1, a_2 < X_2 < b_2\}$:

$$\begin{aligned}
&\Pr(a_1 < X_1 \le b_1, a_2 < X_2 < b_2) \\
&= \Pr(X_1 \le b_1, X_2 \le b_2) - \Pr(X_1 \le a_1, X_2 \le b_2) \\
&- \Pr(X_1 \le b_1, X_2 \le a_2) + \Pr(X_1 \le a_1, X_2 \le a_2) \\
&= \Pr(X_1 \le b_1) \Pr(X_2 \le b_2) - \Pr(X_1 \le a_1) \Pr(X_2 \le b_2) \\
&- \Pr(X_1 \le b_1) \Pr(X_2 \le a_2) + \Pr(X_1 \le a_1) \Pr(X_2 \le a_2) \\
&= (\Pr(X_1 \le b_1) - \Pr(X_1 \le a_1))(\Pr(X_2 \le b_2) - \Pr(X_2 \le a_2)) \\
&= \Pr(a_1 < X_1 \le b_1) \Pr(a_2 < X_2 \le b_2).
\end{aligned}$$

□

**Claim 32.** *If $X = (X_1, X_2)$ has density $f_X$ then $X_1, X_2$ are independent iff $f_X = f_{X_1} f_{X_2}$.*

## 4.9 Functions of random vectors

If $h : \mathbb{R}^2 \to \mathbb{R}$ is measurable and $X$ is a 2 dimensional random vector, then $h(X)$ is a random variable.

There are many examples, but an important one is addition. Let $h(x_1, x_2) = x_1 + x_2$. Assume $X_1, X_2$ are independent and are absolutely continuous. Then $X_1 + X_2$ has the following CDF

$$F_{X_1+X_2}(t) = \int_{s_1} \int_{s_2 \le t - s_2} f_X(s) ds = \int_{s_1} f_{X_1}(s_1) \left( \int_{s_2 \le t - s_2} f_{X_2}(s_2) ds_2 \right) ds_1.$$

Taking derivative gives

$$f_{X_1+X_2}(t) = \int_{s_1} f_{X_1}(s_1) f_{X_2}(t - s_1) ds_1.$$

This is called the convolution of $f_{X_1}$ and $f_{X_2}$, and is denote

$$f_{X_1+X_2} = f_{X_1} * f_{X_2}.$$

# Chapter 5

# Moments

So far we have defined relatively abstract notions (and also gave examples of concrete special cases). Here we start discussing more concrete properties.

Imagine an economic system, consisting of individuals that get salaries. We have an abstract model of the society, and this model allows to abstractly answer any question we are interested in. But what are the simplest questions one can ask? The most basic properties of the society seems to be the average salary in it, and the average differences between the salaries (the amount of social "inequality").

We now discuss the mathematical notions that capture these properties, and more.

## 5.1 Expectation

The expectation of a random variable is what is commonly thought of as "average value."

**Definition 33.** *Let $X$ be a discrete random variable. It expectation is*

$$\mathbb{E}X = \sum_k \Pr(X = k) \cdot k,$$

*if this sum absolutely converges.*

Another way to think of expectation is as a center of mass: Imagine mass $\Pr(X = k)$ distributed at a point $k$. Then $\mathbb{E}X$ is the center of mass of the total mass of 1.

Another useful way to compute expectation: If $\Omega$ is also discrete this is also equal to

$$\sum_{\omega \in \Omega} \Pr(\{\omega\}) X(\omega).$$

Indeed, by Fubini:

$$\begin{aligned}\mathbb{E}X &= \sum_k \Pr(X=k)k \\ &= \sum_k \sum_{\omega: X(\omega)=k} \Pr(\{\omega\})k \\ &= \sum_\omega \Pr(\{\omega\})X(\omega).\end{aligned}$$

**Definition 34.** *Let $X$ be an absolute continuous random variable. It expectation is*

$$\mathbb{E}X = \int_{-\infty}^{\infty} f_X(s)sds,$$

*if this integral absolutely converges.*

There is a general definition, but we shall not provide it in this course. We shall state general theorem about $\mathbb{E}X$, but shall prove only for r.v.'s from the two types above.

### 5.1.1 Examples

*Binomial.* $X \sim Bin(n,p)$. What should $\mathbb{E}X$ be? $np$. Let us see:

$$\begin{aligned}\mathbb{E}X &= \sum_{k=0}^{n} \Pr(X=k)k \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} k \\ &= \sum_{k=1}^{n} n\binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} = pn.\end{aligned}$$

*Uniform.* $X \sim U[a,b]$. Guess? $(a+b)/2$. Indeed,

$$\mathbb{E}X = \int_a^b \frac{1}{b-a} sds = \frac{1}{b-a}\frac{1}{2}(b^2-a^2) = (a+b)/2.$$

*Geometric.* If $X \sim Geo(p)$ then $\mathbb{E}X = 1/p$.

*Normal.* If $X \sim N(0,1)$ then

$$\mathbb{E}X = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} s ds.$$

Since it is an odd function, its integral is 0. Similarly, if $X \sim N(\mu, \sigma)$ then $\mathbb{E}X = \mu$.

*There are* many more examples.

*Non examples.* Let $\Pr(X = k) = c/|k|^2$ for some $c > 0$ and $k$ non zero integer. The expectation does not converge.

## 5.2 Tail formula

There is another useful way to compute expectation.

**Theorem 35.** *If $X$ is a non negative random variables then*

$$\mathbb{E}X = \int_0^{\infty} \Pr(X \geq t) dt.$$

*In general,*

$$\mathbb{E}X = \int_{-\infty}^{0} F_X(t) dt + \int_0^{\infty} (1 - F_X(t)) dt.$$

*Proof.* Let us prove only first part when $X$ is discrete. Use Fubini:

$$\begin{aligned}
\mathbb{E}X &= \sum_{k \geq 0} \Pr(X = k) k \\
&= \sum_{k \geq 0} \Pr(X = k) \int_0^k 1 ds \\
&= \int_0^{\infty} \sum_{k \geq s} \Pr(X = k) ds \\
&= \int_0^{\infty} \Pr(X \geq s) ds.
\end{aligned}$$

□

Examples:

- $X_1, X_2 \sim Exp(\lambda)$ independently. Let $Y = \min X_1, X_2$. What is $\mathbb{E}(Y)$? By definition? Using tail: $\Pr(Y > t) = \Pr(X_1 > t) \Pr(X_2 > t)$.

- Convex combination. If $F_X = \alpha F_Y + (1 - \alpha) F_Z$ then

$$\mathbb{E}X = \alpha \mathbb{E}Y + (1 - \alpha) \mathbb{E}Z.$$

## 5.3 Properties

**Linearity.** A very important property is linearity.

**Theorem 36** (Linearity)**.** *If $X, Y$ are two r.v. with finite expectation then $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for all $a, b \in \mathbb{R}$.*

Try to prove from definition for discrete, say.

*Proof.* Let us consider a discrete $\Omega$. The basic reason is the formula

$$\mathbb{E}X = \sum_{\omega \in \Omega} \Pr(\{\omega\}) X(\omega).$$

□

This property is called linearity of expectation. It is an extremely important and useful property. Part of the importance is that it holds for any two random variables, not necessarily independent. Examples:

*Binomial.* Let $X_1, \ldots, X_n$ be independent $Ber(p)$. How is their sum $X$ distributed? It is $Bin(n, p)$. Thus,

$$\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}X_i = np.$$

*Random permutation.* Let $f : [n] \to [n]$ be a uniformly chosen random permutation. Denote by $X$ the number of fixed points in $X$. What is $\mathbb{E}X$? Let $X_i$ be the indicator of the event $\{f(i) = i\}$. Thus $\mathbb{E}X_i = \Pr(f(i) = i) = 1/n$. So, $\mathbb{E}X = n/n = 1$. The random variables e.g. $X_1, X_2$ are not independent.

*Birthday paradox.* What is minimum number $k$ so that if there are $k$ people in a room the expected number of pairs who share a birthday is at least 1? Denote by $X_{i,j}$ the indicator for the event that person $i$ and person $j$ share a birthday, for $1 \le i < j \le k$. Thus $\mathbb{E}X_{i,j} = 1/365$ and

$$\mathbb{E} \sum_{i,j} X_{i,j} = \binom{k}{2} / 365.$$

This means that the minimum $k$ is roughly $\sqrt{730} \approx 27$.

*Maximum.* We have seen how to use tail formula to compute expectation of minimum. What about maximum? We can use linearity:

$$X + Y = \min\{X, Y\} + \max\{X, Y\}.$$

So if we know $\mathbb{E}X, \mathbb{E}Y$ and $\mathbb{E}\min\{X, Y\}$ we also know $\mathbb{E}\max\{X, Y\}$.

**Respect order.** A basic property of expectation is that it respects order.

**Claim 37** (Respect order)**.** *If $X, Y$ are two r.v. so that $X(\omega) \geq 0$ for all $\omega \in \Omega$, then $\mathbb{E}X \geq 0$.*

This is obvious in the cases we considered.

**Expectation of a function.** Example: Let $X \sim U[0,1]$ and $Y = X^2$. By definition,

$$\mathbb{E}Y = \int f_Y(t)tdt.$$

To compute this, we need to compute $f_Y$ which requires computing $F_Y$. There is a quicker way.

**Theorem 38** (Expectation of a function)**.** *Let $X$ be a random variable and $h$ be a measurable function. If $X$ has density $f_X$ then*

$$\mathbb{E}h(X) = \int_{-\infty}^{\infty} f_X(s)h(s)ds$$

*and if $X$ is discrete*

$$\mathbb{E}h(X) = \sum_k \Pr(X = k)h(k).$$

This is not obvious from definition.

*Proof.* Let us prove for $\Omega$ discrete:

$$\begin{aligned}\sum_k \Pr(X = k)h(k) &= \sum_k \sum_{\omega: X(\omega)=k} \Pr(\{\omega\})h(X(\omega)) \\ &= \sum_\omega \Pr(\{\omega\})(h \circ X)(\omega).\end{aligned}$$

□

Example: $X \sim N(0,1)$ and $Y = e^X$. Thus,

$$\mathbb{E}Y = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^t dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t^2-2t+1)/2} e^{1/2} dt = e^{1/2}.$$

## 5.4 Inequalities

Inequalities are extremely useful in mathematics, and usually harder to prove than equalities. Here we consider two well known ones.

**Theorem 39** (Cauchy-Schwartz)**.** *If $X, Y$ are two random variables over the same space so that $\mathbb{E}X^2, \mathbb{E}Y^2$ are finite then*

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2\mathbb{E}Y^2}.$$

*Proof.* There are several options to prove. One is the use the known inequality for finitely supported random variables and then use approximations. Another is to consider

$$p(t) = \mathbb{E}(|X| + t|Y|)^2$$

which is always non negative. Observe that it is finite for all $t \in \mathbb{R}$ since $2ab \leq a^2 + b^2$ for all $a, b > 0$. Write

$$p(t) = \mathbb{E}Y^2 \cdot t^2 + 2\mathbb{E}|X||Y|t + \mathbb{E}X^2.$$

So,

$$4(\mathbb{E}|X||Y|)^2 - 4\mathbb{E}Y^2\mathbb{E}X^2 \leq 0,$$

since otherwise $p$ gets a negative value. □

In other words, we can think of r.v.'s as a vector space and of $\mathbb{E}XY$ as an inner product (if it is defined).

What is a convex function? There are several equivalent definitions.

1. $h''$ is non negative (if it is defined).
2. The tangent to $h$ at every point is below the graph.
3. The area above $h$ is convex.
4. For every $x_1, x_2$ and $p \in [0, 1]$,

$$h(px_1 + (1-p)x_2) \leq ph(x_1) + (1-p)h(x_2).$$

This exactly corresponds to a r.v. that takes 2 values. To prove for general r.v. use induction to get all finitely supported r.v.'s and then use approximations to get all.

**Theorem 40** (Jensen)**.** *If $h$ is a convex function on the image of $X$ then $h(\mathbb{E}(X)) \leq \mathbb{E}h(X)$.*

Examples: $|x|^p$ for $p \geq 1$. $a^x$ for $a \geq 1$. Specifically, $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

## 5.5 Variance

We have discussed the "average salary" in the society. What about the "amount of inequality"? How to measure? It should be the average difference from the average.

**Definition 41.** *The expectation of a random variables $X$ is defined as*

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2,$$

*if it is defined.*

Use $(X - \mathbb{E}(X))^2$ to measure distance instead of $|X - \mathbb{E}(X)|$ since is easier to work with, and is similar to $L_2$ norm.

The variance is always non negative and is zero iff $X = \mathbb{E}(X)$ almost surely.

**A different formula.**

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

**Standard deviation.** Variance is not measured in the same units as expectation. The square root of it does. The standard deviation of $X$ is

$$\sigma_X = \sqrt{Var(X)}.$$

The name explains it self.

**Examples:**

*Binomial.* Let $X \sim Bin(n, p)$. Let $X_1, \ldots, X_n$ be independent $Ber(p)$ so that $X = \sum_i X_i$.

$$\mathbb{E}X^2 = \sum_{i,j \in [n]} \Pr(X_i = 1, X_j = 1) = np + n(n-1)p^2.$$

So,

$$Var(X) = np + n(n-1)p^2 - (pn)^2 = np - np^2 = np(1-p).$$

Again, we see how useful is linearity of expectation.

**Variational definition of expectation:** What is $a$ that minimizes $\mathbb{E}(X - a)^2$? Write

$$\mathbb{E}(X-a)^2 = \mathbb{E}X^2 - 2a\mathbb{E}X + a^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 + (\mathbb{E}X - a)^2 \geq Var(X).$$

**Translations and dilation.** For every $a \in \mathbb{R}$,

$$Var(a + X) = Var(X).$$

And

$$Var(aX) = a^2 Var(X).$$

**Sums.** When does $Var(X+Y) = Var(X) + Var(Y)$ hold? Let us see:

$$Var(X+Y) = \mathbb{E}(X - \mathbb{E}X)^2 + \mathbb{E}(Y - \mathbb{E}(X))^2 + 2\mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)).$$

This holds iff

$$\mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) = 0.$$

This quantity is called covariance:

$$cov(X,Y) := \mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

If $X, Y$ are independent then

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

which means

$$cov(X,Y) = 0.$$

For example, when $(X,Y)$ are discrete,

$$\mathbb{E}(XY) = \sum_k \Pr(XY = k)k = \sum_{k,\ell} \Pr(X = \ell)\Pr(Y = k/\ell)\ell \cdot (k/\ell) = \mathbb{E}(Y) \cdot \mathbb{E}(X).$$

In this case, $X, Y$ are called uncorrelated.

The other direction does not always holds. For example, if $X \sim N(0,1)$ and $Y = X^2$ then

$$\mathbb{E}X = 0$$

and

$$cov(X,Y) = \mathbb{E}X^3 = 0$$

since it is the integral of an odd function.

## 5.6 Higher moments

We have seen the importance of $\mathbb{E}X, \mathbb{E}X^2$. In general, these are the 1st and 2nd moments of $X$. The nth moment is defined as $\mathbb{E}X^n$ if it is defined, and the nth absolute moment is defined as $\mathbb{E}|X|^n$. Observe if the 2nd moment is defined then so is the 1st: $\mathbb{E}X \le \sqrt{\mathbb{E}X^2}$. Similarly, for all $m < k$,

$$\mathbb{E}|X|^m = \mathbb{E}(|X|^k)^{m/k} \le (\mathbb{E}|X|^k)^{m/k},$$

due to convexity of $x \mapsto x^{m/k}$.

# Chapter 6

# Laws of large numbers

## 6.1 Probability estimates

One of the most useful things to do is to obtain good probability estimates.

We start with the simplest example. If the average height of a person is 1.7 meters, what is probability that a person have height larger than 3.4 meters? One half.

**Theorem 42** (Markov's inequality)**.** *Let $X$ be a non negative random variables. For every $a > 0$,*

$$\Pr[X > a\mathbb{E}(X)] < 1/a.$$

*Proof.* Consider discrete $X$ for example.

$$\mathbb{E}X = \sum_{k > a\mathbb{E}(X)} \Pr(k)k + \sum_{0 \leq \ell \leq a\mathbb{E}(X)} \Pr(\ell)\ell \geq a\mathbb{E}(X) \Pr(X > a\mathbb{E}(X)).$$

□

This is an extremely simple inequality but many others are built on top of it.

One example that you have seen in the exercise. Let $G \sim G(n, 1/2)$ be a random graph. The expected number of cliques of size $k = 3 \log n$ is $G$ is at most $n^k 2^{-k(k-1)/2} < 2/n$. So, if $n$ is large, the probability that $G$ has a clique of size larger than $k$ is at most $2/n$.

Markov's inequality is tight. For every $a$ there is a r.v. $X$ so that

$$\Pr(x \geq a\mathbb{E}(X)) = 1/a.$$

It is the random variables that takes the value $a\mathbb{E}(X)$ w.p. $1/a$ and the value 0 otherwise.

Non negativity is important.

We can apply Markov's inequality to more complicated random variables and get better estimates.

**Theorem 43** (Chebyshev's inequality)**.** *Let $X$ be a random variables with finite variance. For all $a > 0$,*

$$\Pr[|X - \mathbb{E}(X)| > a\sigma] < 1/a^2,$$

*where $\sigma = \sqrt{Var(X)}$ is the standard deviation of $X$.*

*Proof.* Denote $Y = |X - \mathbb{E}(X)|$. It is a non negative random variables. Notice $\mathbb{E}Y^2 = Var(X)$. Apply Markov's inequality on $Y$.

$$\Pr[Y > a\sigma] = \Pr[Y^2 > a^2 Var(X)] < 1/a^2.$$

□

This inequality roughly means that if $\sigma_X \ll \mathbb{E}(X)$ then the probability that $X$ takes values that are far from it expectation is small.

## 6.2 Weak law of large numbers

We can go back to a property we saw, and described applications of, but did not prove because we did not have the right tools (we could have proven it with a some amount of calculations). We now prove it.

The special case we saw is:

**Theorem 44.** *Let $X \sim Bin(n, p)$ with $p \in (0, 1)$. For all $\delta > 0$,*

$$\Pr(|X - pn| \geq \delta n) \leq \frac{p(1-p)}{\delta^2 n} \leq \frac{1}{4\delta^2 n}.$$

The general theorem with finite variance is:

**Theorem 45** (Weak law)**.** *Let $X_1, \ldots, X_n$ be independent random variables with expectation $\mu$ and variance $\sigma_i^2 < \sigma$. Let $X = (1/n)\sum_i X_i$. For all $\delta > 0$,*

$$\Pr(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2 n}.$$

*Specifically, the probability tends to 0 as $n \to \infty$.*

*Proof.* Due to independence,

$$\mathsf{Var}(X) = \sum_i Var(X_i/n) \leq n(\sigma/n)^2 = \sigma^2/n.$$

Using Chevyshev,

$$\Pr(|X-\mu| \geq \delta) = \Pr\left(|X-\mu| \geq \delta \frac{\sqrt{n}}{\sigma}\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\sigma^2}{\delta^2 n}.$$

□

The condition on the variance being finite in not necessary.

**Theorem 46** (Khintchine)**.** *Assume $X_1, X_2, \ldots$ are i.i.d. random variables, each distributed as $X$ so that $\mathbb{E}X = \mu$. Let $S_n = (1/n)\sum_{i=1}^n X_i$. Then,*

$$\lim_{n\to\infty} \Pr(|S_n - \mu| > \delta) = 0.$$

*Sketch.* The proof is by an appropriate approximation. Consider discrete $X$ for example. Without loss of generally assume, $\mathbb{E}X = \mu = 0$. Let $\epsilon > 0$. Let $n$ be large enough.

**Truncation and its properties.** Truncate $X$ as

$$X = X_{\leq n} + X_{>n},$$

where

$$X_{\leq n} = X \cdot 1_{|X|\leq n}.$$

The rational is the we can't bound the moments of $X$, however we can bound the moments of $X_{\leq n}$, and we think of $X_{>n}$ as the error term that we just need to control.

**Partition $S_j$ to two parts.** Define

$$(S_j)_{\leq n} = (1/j)\sum_{i\leq j}(X_i)_{\leq n}$$

and

$$(S_j)_{>n} = (1/j)\sum_{i\leq j}(X_i)_{>n}.$$

This is not the truncation of $S_j$ but the average of truncations of $X$.

**Bounding error.** The sequence

$$\mathbb{E}(|X_{>n}|) = \sum_k \Pr(X=k)|k|1_{|k|>n}$$

of $n$ is decreasing and bounded from below, so its limit is its infimum which is 0 since $\mathbb{E}|X| < \infty$ (it is the tail of a converging sum). Thus, $\mathbb{E}|X_{>n}| \to 0$ when $n \to \infty$. So, by convexity,

$$\mathbb{E}|(S_j)_{>n}| \leq \epsilon\delta.$$

Markov's inequality implies

$$\Pr(|(S_j)_{>n}| \geq \delta) \leq \epsilon.$$

**Bounding bulk.** First, expectation: as above, if $n$ is large then

$$|\mathbb{E}(S_j) - \mathbb{E}(S_j)_{\leq n}| = |\mathbb{E}(S_j)_{>n}| \leq \epsilon\delta.$$

Second, variance:

$$Var((S_j)_{\leq n}) = \sum_{i \leq j} Var(X_{\leq n}/j) \leq (1/j)^2 jn^2 \leq n^2/j,$$

using independence.

So,

$$\begin{aligned}
\Pr\left(|(S_j)_{\leq n} - \mu| \geq \delta\right) &\leq \Pr\left(|(S_j)_{\leq n} - \mathbb{E}(S_j)_{\leq n}| \geq \delta - \delta\epsilon\right) \\
&\leq \Pr\left(|(S_j)_{\leq n} - \mathbb{E}(S_j)_{\leq n}| \geq \delta/2\right) \\
&\leq \frac{4Var((S_j)_{\leq n})}{\delta^2} \leq \frac{4n^2}{\delta^2 j} < \epsilon,
\end{aligned}$$

for $j$ large.

**Together.** By the union bound,

$$\Pr(|S_j - \mu| \geq 2\delta) \leq \Pr(|(S_j)_{\leq n} - \mu| \geq \delta) + \Pr(|(S_j)_{>n}| \geq \delta) \leq 2\epsilon.$$

**Choice of j.** We first chose $n = n(X, \epsilon, \delta)$ large, and then chose $j = j(n, \epsilon, \delta) = j(X, \epsilon, \delta)$ large. □

We can see how much more difficult it is to prove without assumption of finite variance. This is a phenomenon that often occurs.

There are also stronger variants of this law, but we shall not discuss now (we have briefly discussed before).

## 6.3 Types of convergence

The weak law says that a certain sequence of probabilities tends to 1. This is one type of convergence, called weak or in probability.

**Definition 47** (Convergence in probability)**.** *The sequence of random variables $(X_n)$ converges to $X$ in probability or weakly if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} \Pr(|X_n - X| > \epsilon) = 0.$$

In this language, the weak law says that the average of i.i.d. variables converges weakly to its expectation (which is a constant r.v.).

There is also a stronger type of convergence.

**Definition 48** (Convergence almost surely)**.** *The sequence* $(X_n)$ *converges to* $X$ *almost surely (a.s.) or strongly if there is a set* $A \in \mathcal{F}$ *so that* $\Pr(A) = 1$ *and* $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ *for every* $\omega \in A$*.*

This is indeed a stronger type of convergence.

**Theorem 49.** *If* $(X_n)$ *converges to* $X$ *a.s. then it converges to* $X$ *in probability.*

*Proof.* Let $\epsilon > 0$ and let

$$B_k = \{\omega : |X_k(\omega) - X(\omega)| > \epsilon\}.$$

Our goal is to show that if $k$ is large then $\Pr(B_k) < \epsilon$.

Let $E$ be the set of $\omega \in \Omega$ that belong to infinity many $B_n$'s, i.e.

$$E = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} B_k.$$

By strong convergence, every $\omega \in A$ belongs to at most finitely many $B_k$'s. Namely, $E$ is disjoint from $A$, and so $\Pr(E) = 0$

Consider the sequence of sets

$$E_m = \bigcap_{n=1}^{m} \bigcup_{k=n}^{\infty} B_k \supseteq B_m.$$

Hence

$$E_1 \supseteq E_2 \supseteq \cdots \supseteq E.$$

$\sigma$-additivity implies

$$\Pr(E_1 - E) = \sum_{m=1}^{\infty} \Pr(E_m - E_{m+1}) < 1.$$

So, if $n$ is large then

$$\Pr(E_n - E) < \epsilon$$

since (by $\sigma$-additivity) it is the tail of a converging sum, which implies

$$\Pr(B_n) \leq \Pr(E_n) < \epsilon + \Pr(E) = \epsilon.$$

□

The other direction does not always hold. Let $Y \sim U[0,1]$. Define $(X_{n,k} : n \in \mathbb{N}, 1 \leq k \leq n)$ as follows. The variables $X_{n,k}$ is the indicator of the event

$$\{Y \in [k/n, (k+1)/n]\}.$$

Specifically $\Pr(X_{n,k} \neq 0) = 1/n \to 0$ as $(n,k) \to \infty$. So this sequence weakly converges to 0. But for every $\omega \in [0,1]$, the sequence $X_{n,k}(\omega)$ does not even converge.

## 6.4 Strong law of large numbers

The weak law roughly says that the average of i.i.d. variables converges weakly to the expectation. The strong law says that the same holds with strong convergence. As in the proof of weak law, with a bound on the higher moments the proof is much simpler.

**Theorem 50** (Strong law)**.** *Let $X_1, X_2, \ldots$ be i.i.d. so that $\mathbb{E}X_1^4 < \infty$ and $\mathbb{E}X = \mu$. Let $S_n = \sum_{i \leq n} X_i/n$. Then,*

$$\Pr(\lim_{n\to\infty} S_n \neq \mu) = 0.$$

We will not prove the most general statement (due to Khintchine): The strong law holds even under the condition that the expectation is finite.

The weak law does not tell us anything about normal numbers. The strong law tells us that almost every number is normal.

To prove the strong law we need a criterion (as simple as possible) that guarantees strong converges.

**Lemma 51** (Borel-Cantelli)**.** *Let $A_n$ be a sequence of events so that $\sum_n \Pr(A_n) < \infty$. Then,*

$$\Pr\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\right) = 0.$$

*In words, the probability measure of the set of $\omega$'s that appear in infinitely many $A_n$'s is zero.*

*Proof.* For every $k$, by $\sigma$-additivity,

$$\Pr\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\right) \leq \Pr\left(\bigcap_{n=k}^{\infty} \bigcup_{m=n}^{\infty} A_m\right) \leq \sum_{n=k}^{\infty} \Pr(A_n) \to 0,$$

as $k \to \infty$ since it is the tail of a converging sum. □

**Corollary 52.** *If* $\sum_n \Pr(|X - X_n| > \epsilon) < \infty$ *for all* $\epsilon > 0$ *then* $X_n$ *strongly converges to* $X$*.*

The proof is left as an exercise. This is not an "iff" condition (exercise). The assumption immediately implies weak convergence since the tail of a converging sum converges to 0. But it actually says that this convergence to 0 is fast, which suffices to prove strong convergence.

*Proof of Theorem 50.* Without loss of generality, $\mu = 0$. A higher moment version of Chebyshev says

$$\Pr(|S_n| > \epsilon) \leq \frac{\mathbb{E}(S_n^4)}{\epsilon^4}.$$

Let us estimate $\mathbb{E}S_n^4$. Using independence, and Cauchy-Schwartz,

$$\begin{aligned} n^4 \mathbb{E}S_n^4 &= \sum_{i,j,k,\ell=1}^{n} \mathbb{E}(X_i X_j X_k X_\ell) = \sum_{i,j,k,\ell=1}^{n} \mathbb{E}(X_i)\mathbb{E}(X_j)\mathbb{E}(X_k)\mathbb{E}(X_\ell) \\ &\leq 10 \left( \sum_{i,j=1}^{n} \mathbb{E}(X_i^2)\mathbb{E}(X_j^2) + \sum_{i=1}^{n} \mathbb{E}(X_i^4) \right) \leq Cn^2. \end{aligned}$$

So,

$$\Pr(|S_n| > \epsilon) \leq \frac{C}{\epsilon^4 n^2}.$$

Borel-Cantelli finishes the proof. □

*Summary.* We have defined the moments of random variables, used them to get probability estimates, and to prove convergence of two types.

# Chapter 7

# More on random vectors

We now add some more properties of random vectors. In other words, we just discuss interaction between several random variables in more detail.

## 7.1 Expectation

If $X$ is an $n$ dimensional random vector, and $h : \mathbb{R}^n \to \mathbb{R}$ is a measurable function then $h(X)$ is a random variable.

**Theorem 53.** *If $X$ is absolutely continuous then*

$$\mathbb{E}h(X) = \int f_X(x)h(x)dx.$$

*If $X$ is discrete then*

$$\mathbb{E}h(X) = \sum_k \Pr(X = k)h(k).$$

*These are integrals and sums in $n$ dimensional space.*

## 7.2 Conditioning

We consider the 2 dimensional case. Let $(X, Y)$ be a 2 dimensional random vector. We would like to understand the distribution of $Y$ conditioned on the value of $X$.

**Discrete.** When $X$ is discrete this is relatively straightforward. For every $k$ so that $\Pr(X = k) > 0$, we can consider the distribution of $Y$ conditioned on the event $\{X = k\}$. Thus, $\Pr(Y = \ell | X = k)$ is the probability function of $Y$ conditioned on $X = k$. We also have

$$F_{Y|X}(\ell|k) = \Pr(Y \leq \ell | X = k).$$

A more elaborate definition is of conditional expectation.

$$\mathbb{E}(Y|X=k) = \sum_{\ell} \Pr(Y=\ell|X=k)\ell.$$

This is a map $k \mapsto \mathbb{E}(Y|X=k)$. In other words, we may think of $\mathbb{E}(Y|X)$ as a random variable.

**Absolutely continuous.** When $(X,Y)$ has a density, then $\Pr(X=x)=0$ so we can not condition on the event $\{X=x\}$. The way to handle this is to define a conditional density. Instead of considering $\{X=x\}$, consider the event $E_\delta = \{X \in [x, x+\delta]\}$. Now, if $f_{X,Y}$ is continuous then

$$\mathcal{F}_{Y|E}(y|E_\delta) = \frac{F_{X,Y}(x+\delta,y) - F_{X,Y}(x,y)}{F_X(x+\delta) - F_X(x)} \to_{\delta\to 0} \frac{\frac{\partial}{\partial x}F_{X,Y}(x,y)}{f_X(x)}.$$

This suggests the following definition. Assume $(X,Y)$ are absolutely continuous. Then,

$$F_{Y|X}(y|x) = \frac{\frac{\partial}{\partial x}F_{X,Y}(x,y)}{f_X(x)}$$

and

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y}F_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

This is similar to but different than definition of conditioning for events. The conditional expectation is thus defined as

$$\mathbb{E}(Y|X=x) = \int f_{Y|X}(y|x) y dy.$$

There are analogs of properties we already saw:

$$f_Y(y) = \int f_{Y|X}(y|x) f(x) dx$$

and

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)}$$

if it is defined.

Example: Let $X$ chosen with density $f_X(x)$ supported say on $[0,1]$. Given $X=x$, choose $Y$ to be $U[0,x]$. That is, $f_{Y|X}(y|x)$ is the density function of $U[0,x]$. We can now compute $f_Y, \mathbb{E}Y$ using the formulas above.

*Best estimator:* We mentioned that $\mathbb{E}(Y)$ can be thought of as the best a priori estimate of $Y$. It is the number that minimized $\mathbb{E}(|Y-a|^2)$.

Similarly, $\mathbb{E}(Y|X=x)$ can be thought of as the best predictor of $Y$ given that $X=x$. This corresponds to cases when we are interested in the value of some quantity $Y$, that we can not measure. The only thing we can measure is $X$. Given that we measured $X=x$, our best estimate of the value of $Y$ is $\mathbb{E}(Y|X=x)$.

### 7.2.1 Law of total expectation

Above we have defined $\mathbb{E}(Y|X)$. We mentioned that it is a random variable (it is a measurable function of $X$). It turns out to be useful in many cases.

**Theorem 54** (Law of total expectation)**.** *Let $(X,Y)$ be a random vector so that $\mathbb{E}Y < \infty$. Then,*

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

Let us consider example from above, with $Y \sim U[0,X]$. Since $\mathbb{E}(Y|X) = X/2$ we have

$$\mathbb{E}Y = \mathbb{E}X/2 = (1/2)\mathbb{E}X.$$

*Proof.* Prove for discrete.

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y|X)) &= \sum_x \Pr(X=x)\sum_y \Pr(Y=y|X=x)y = \sum_x\sum_y \Pr(X=y, X=x)y \\ &= \sum_y \Pr(Y=y)y = \mathbb{E}Y.\end{aligned}$$

□

It will sometimes be useful to use the following. If $(X,Y)$ is discrete and $h$ measurable then

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

A similar statement holds in general (under some conditions on $h$ that we do not specify now).

*Polya's urn:* Imagine an urn with 1 white ball and 1 black ball. We take a ball out of the urn, and put 2 of the same color. So there are now 3 ball, but their color is random. We keep on going for $n$ more steps. Denote by $X_k$, $0 \le k \le n$, the fraction of white balls at time $k$ ($X_0 = 1/2$). At time $k$ there are $k+2$ balls. What is $\mathbb{E}X_n$?

$$\begin{aligned}\mathbb{E}X_n &= \mathbb{E}(\mathbb{E}(X_n|X_{n-1})) \\ &= \mathbb{E}\left(\frac{X_{n-1}(X_{n-1}(n+1)+1) + (1-X_{n-1})X_{n-1}(n+1)}{n+2}\right) = \mathbb{E}X_{n-1}.\end{aligned}$$

And by induction $\mathbb{E}X_n = X_0 = 1/2$ for all $n$.

## 7.3 Covariance

We have seen definition of covariance. It is a measure for the correlation between two random variables. We have seen two way to compute. One is

$$cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

**Theorem 55** (Properties of covariance)**.**

1. $cov(X,X) = Var(X)$.
2. $cov(X,Y) = cov(Y,X)$.
3. *Linearity:* $cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$ *and* $cov(aX,Y) = acov(X,Y)$ *for* $a \in \mathbb{R}$.
4. *If* $X, Y$ *are independent then* $cov(X,Y) = 0$. *Shows that it is not inner product.*
5. *Cauchy-Schwartz:* $|cov(X,Y)|^2 \leq Var(X)Var(Y)$.

Covariance is not a normalized measure. E.g. $cov(X,Y) = 1$ is not universally meaningful. To normalize is the *(Pearson) correlation coefficient* (mekadem mitam):

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

if the variances are non zero. It has similar properties to variance, except that it is normalized to be between $-1, 1$. It measures the linear correlation between $X, Y$. Roughly, the larger $\rho_{X,Y}$ is the larger the linear correlation is. E.g. if $\rho_{X,Y} = 1$ then $Y = aX + b$ a.s. with $a, b \in \mathbb{R}$.

Example: Let $X$ be some random variable taking positive value, and let $Y \sim Exp(X)$. Is the covariance between $X, Y$ positive or negative? Guess: The larger $X$, the smaller $Y$ typically is, so it should be non positive. Indeed,

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY|X)) = \mathbb{E}(X\mathbb{E}(Y|X)) = \mathbb{E}(X(1/X)) = 1,$$

and similarly

$$\mathbb{E}(Y) = \mathbb{E}(1/X)$$

so

$$cov(X,Y) = 1 - \mathbb{E}(X)\mathbb{E}(1/X) \leq 0$$

since $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$ by convexity.

## 7.4 Covariance matrix

We represent the covariances between the entries of $X = (X_1, \ldots, X_n)$ in an $n \times n$ matrix:

$$\sigma_X(i,j) = cov(X_i, X_j),$$

for $i, j \in [n]$. This matrix captures the correlation between the entries of $X$. It has the following properties:

1. $\sigma_X$ is symmetric. This implies that it has $n$ real eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ with $n$ orthogonal eigenvectors. That is, there is a real matrix $U$ so that $UU^T = I$ and $\sigma_X = U\Lambda U^T$ with $\Lambda$ diagonal so that $\Lambda(i,i) = \lambda_i$ for all $i$.

2. $\sigma_X(i,i) = Var(X_i)$ for all $i$.

3. $\sigma_X$ is positive semi-definite. That is, for every vector $u$ we have $\langle \sigma_X u, u \rangle \geq 0$ with the standard inner product over $\mathbb{R}$ or $\mathbb{C}$. In other words, $\lambda_1 \geq 0$. Indeed, for every $u = (u_1, \ldots, u_n)$, consider $Y = \sum_i u_i X_i$,

$$0 \leq Var(Y) = cov(Y,Y) = \sum_{i,j} u_i u_j \sigma_X(i,j) = \langle u\sigma_X, u \rangle.$$

4. If $M \in \mathbb{R}^{m \times n}$ and $Y = MX$ then

$$\sigma_Y = M\sigma_X M^T.$$

   Indeed,

$$\sigma_Y(i,j) = cov((MX)_i, (MX)_j) = \sum_{k,\ell=1}^{n} M_{k,i} M_{\ell,j} \sigma_X(k,\ell).$$

## 7.5 Gaussians

We now define one of the most important distributions on real vectors. Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite, and let $\mu \in \mathbb{R}^n$. Define a density on $\mathbb{R}^n$ by

$$f(x) = f_{A,\mu}(x) = ce^{-\frac{\langle (x-\mu)A, x-\mu \rangle}{2}}$$

with

$$c = \sqrt{\frac{det(A)}{(2\pi)^n}} > 0$$

a normalization constant (perhaps discuss its value later). The integral is finite, since $A$ is positive definite (if there is a kernel or a negative eigenvalue the integral is infinite). A random variable with this density is an $n$-dimensional Gaussian.

When $A = I$ and $\mu = 0$, it is a normal Gaussian with density

$$f(x) = ce^{-\frac{\sum_{i=1}^n x_i^2}{2}}.$$

It is radial symmetric. Its expectation is therefore 0. The density is a product density, and hence the coordinates are independent (here we know $c = (1/2\pi)^{n/2}$).

General properties: If $X \sim Gau(A, \mu)$ then

1. $\mathbb{E}X_i = \mu_i$ for all $i$. In vector form $\mathbb{E}X = \mu$.

2. The maximum of the density is at $\mu$. The contour shapes of the density are ellipses centered at $\mu$. In dimension 2, this is a bell shaped curve.

3. For every $S \subset [n]$, denote by $X_S = (x_i : i \in S)$. The vector $X_S$ is also Gaussian in $|S|$ dimensions.

   To prove, integrate in $\mathbb{R}^{[n]-S}$ and get density of similar form. In the integration, there are cross terms involving $x_i$ and $x_j$ for $i \in S$ and $j \notin S$. The integration is done by completing to squares, which moves terms of the form $e^{cx_i^2}$ for $i \in S$ out side the integral. But overall we still have a density of the correct form. Below we shall see what are $A', \mu'$ defining $X_S$.

   Example, for $n = 2$ and $S = \{1\}$:

$$\begin{aligned} f_{X_1}(x_1) &= \int ce^{-(a_{1,1}x_1^2 + 2a_{1,2}x_1x_2 + a_{2,2}x_2^2)/2} dx_2 \\ &= e^{-a_{1,1}x_1^2} e^{a_{1,2}^2 x_{1,1}^2/a_{2,2}} \int ce^{-(\sqrt{a_{2,2}}x_2 - a_{1,2}x_1/\sqrt{a_{2,2}})^2} dx_2, \end{aligned}$$

   the right integral is 1, and the left function of $x_1$ is of the correct form.

4. If $M$ is $n \times n$ and invertible then $Y = MX$ is also Gaussian.

   By the chain rule for high dimensional transformations, since we are doing a linear transformation:

$$f_Y(y) = c_M f_X(M^{-1}x) = c'_M e^{-\frac{\langle A(M^{-1}x-\mu), M^{-1}x-\mu\rangle}{2}} = c'_M e^{-\frac{\langle (M^{-1^T}AM^{-1})(x-M\mu), x-M\mu\rangle}{2}},$$

   where $c_M, c'_M$ are constants that depends on $M$.

5. If $X$ is normal ($A = I$, $\mu = 0$) and $U$ is orthogonal then $UX$ is also normal. In words, a normal/standard Gaussian is invariant under rotations.

6. The covariance matrix of $X$ is
$$\sigma_X = A^{-1}.$$
This gives a simple formula also for the behavior under affine transformations.

Proof: W.l.o.g. $\mu = 0$. The matrix $A$ is positive definite so it has a square root $A^{1/2}$, which is symmetric as well. The matrix of the vector $A^{1/2}X$ is therefore $I$ which means that it is normal. The covariance matrix of $A^{1/2}X$ is therefore $I$ as well (easy calculation). So,
$$\sigma_X = \sigma_{A^{-1/2}A^{1/2}X} = A^{-1/2}IA^{-1/2}.$$

Examples:
- If $(X_1, X_2)$ are defined with $I, \mu = 0$ and
$$(Y_1, Y_2) = \frac{1}{\sqrt{2}}(X_1 + X_2, X_1 - X_2)$$
then $Y$ is also defined with $I, \mu$. These are different vectors with the same distribution. When drawing the rows of $I$ in the plane we get to standard basis vector, that define a circle. The rotation given by $X \mapsto Y$ leaves the circle as is.

- However, if $X$ in the plane is defined by $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ then the rows of $A$ define an ellipse, and the same rotation also rotate the ellipse so that the entries of $Y$ are no longer independent.

You shall discuss conditioning in the exercise.

*Summary.* In this chapter we have discussed connections and measures that related several random variables, how to condition one random variable on another, and an important distribution on a collection of random variables (Gaussians).

# Chapter 8

# Central limit theorem

The central limit theorem is similar to the laws of large numbers in that it provides a universal behavior for system with many independent parts. However, the central limit theorem is more accurate than the laws of large numbers. In the laws of large number, the only promise we get is that a certain event happens with high/small probability. The central limit theorem provide a quantitive estimate.

**Theorem 56.** *Let $(X_n)$ be a sequence of i.i.d. random variables with expectation $\mu$ and variance $\sigma^2$. Denote*

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}}.$$

*Let $Z \sim N(0,1)$. Then, for every $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \Pr(Z_n \leq t) = \Pr(Z \leq t).$$

*The normalization of $Z_n$ is so that $\mathbb{E}Z_n = 0$ and $Var(Z_n) = 1$ for all $n$.*

This theorem is one of the most important theorems in probability theory. It shows the universality of the normal distribution (which also explains it name), and the reason that the "bell shaped curve" appears in so many places.

Assume $\mu = 0$ and $\sigma = 1$. The laws of large numbers say that $S_n = \frac{1}{n}\sum_{i=1}^n X_i \to 0$. The central limit theorem is much more accurate, it says what is roughly the rate of convergence. Not only that $S_n \to 0$, but $Z_n = \sqrt{n}S_n$ which is much larger than $S_n$ tends to a limit $Z$.

The high level idea of the proof we shall see is very simple. We want to show that $F_{Z_n} \to F_Z$ pointwise when $n \to \infty$. We shall consider the characteristic function $\varphi_{Z_n}$, which can be thought of as the Fourier transform of $F_{Z_n}$, and show that it converges to $\varphi_Z$. This will be quite easy. Then we shall explain why if it works for $\varphi$ it also work for $F$. This will be non trivial.

The reason that Fourier transform is useful here is: $Z_n$ is a sum of $n$ i.i.d. random variables. In other words, it corresponds to the convolution of $n$ functions. The Fourier transform of convolution is just a product, which is much easier to understand. This idea appear in many proof throughout mathematics.

Before actually proving the theorem, we need to build some theory.

## 8.1 Weak convergence of monotone functions

We have defined when a sequence of random variables converges to a limit. We now consider the cumulative disturbution functions.

**Definition 57.** *Let $F, F_1, F_2, \ldots$ be uniformly bounded monotone functions. The sequence $(F_n)$ converges weakly to $F$ if $F_n(x) \to F(x)$ for every $x$ in which $F$ is continuous, when $n \to \infty$. We denote this by $F_n \to_w F$ when $n \to \infty$.*

We are mostly interested in function that are right continuous (CDFs). The limit of such functions is not necessarily right continuous; for example, $F_n = \mathbf{1}_{[1/n,\infty)}$ pointwise converges to $\mathbf{1}_{(0,\infty)}$ but it weakly converges to $\mathbf{1}_{[0,\infty)}$. In general, there are only countably many points in which such a function is not continuous.

Note that when $F_{X_n} \to_w F_X$, the function $F_X$ is right continuous, so the example above shows that if $X_n \to_w X$ then it does not hold that $F_{X_n} \to F_X$ pointwise, but it does hold that $F_{X_n} \to_w F_X$.

A basic property of weak convergence is that the space of bounded monotone functions is compact.

**Theorem 58** (Helly's selection)**.** *Let $(F_n)$ be a sequence of monotone functions so that $|F_n(x)| \leq 1$ for all $x, n$. Then, there is a subsequence that weakly converges to a right continuous non decreasing function $F$.*

Even when $F_n$ are CDF the limit is not necessarily such, e.g., if $F_n = \mathbf{1}_{[n,\infty)}$ then $F_n \to 0$. The condition that guarantees that $F$ corresponds to a random variable is roughly $F(-\infty) = 0$ and $F(\infty) = 1$.

## 8.2 Three types of convergence

We can now add a 3rd type of convergence to our list. Before we formally define it, let us recall the previous 2 definitions:

**Strong.** The sequence $X_1, X_2, \ldots$ over the same probability space converges strongly to $X$ if

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

This is the strongest version of convergence.

**Weak (in probability).** The sequence $X_1, X_2, \ldots$ over the same probability space converges weakly to $X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1.$$

**In distribution.** This is the weakest type of convergence we shall discuss.

**Definition 59.** *A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to $X$ if*

$$F_{X_n} \to_w F_X$$

*when $n \to \infty$.*

It is similar in spirit to weak convergence, but does not require the variables to leave over the same space. This marks a key difference between convergence in distribution and weak convergence, and makes the two notions incomparable somehow.

## 8.3 Characteristic function

Here we define the characteristic function $\varphi_X$ of a random variable $X$. It is the Fourier transform of it probability/density function.

**Definition 60.** *The characteristic function $\varphi_X$ of a random variable $X$ is*

$$\varphi_X(t) = \mathbb{E}e^{itX}.$$

*Here $i = \sqrt{-1}$.*

This is a complex integration, which can be thought of as 2 real integrals.

It is important to note that $\varphi_X$ is always defined since $|e^{itX}| = 1$ so the integral is well defined. For every $t$, the value $\varphi_X(t)$ is a point in the complex unit ball (it is the average of points in on the unit circle).

Examples:

- If $X \sim U[\{-1, 1\}]$ then

$$\varphi_X(t) = \frac{1}{2}\left(e^{it} + e^{-it}\right) = \cos(t).$$

Draw in $\mathbb{C}$.

- If $X \sim N(0, 1)$ then

$$\varphi_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{itx} dx = e^{-t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-it)^2/2} dx = e^{-t^2/2}.$$

This is a line integral in the complex plane. The last equality uses Cauchy's theorem, which say that the integral of the holomorphic function $e^{-z^2/2}$ in the closed path

$$(R,0) \to (R,-it) \to (-R,-it) \to (-R,0) \to (R,0)$$

is zero. The part from $(-R,0)$ to $(R,0)$ is roughly 1, the part opposite is what we want, and the two other parts are small (the modulus is $e^{-R^2-s^2}$ for $0 \le s \le t$ which is very small). The details are left as exercise.

We now list important properties of $\varphi_X$:

1. $\varphi_X(0) = 1$ and $|\varphi_X(t)| \le 1$ for all $t$. The second property follows from convexity.

2. If $X, Y$ are independent then

$$\varphi_{X+Y}(t) = \mathbb{E}e^{it(X+Y)} = \mathbb{E}e^{itX}\mathbb{E}e^{itY} = \varphi_X(t)\varphi_Y(t).$$

   This corresponds to that Fourier transform of convolution is product of transforms.

   The other direction does not hold in general. Let $X$ be a random variable so that $\varphi_X(t) = e^{-|t|}$. A random variable with this characteristic function is Cauchy; its density is

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

   The function $\varphi_X(t)$ is the integral along the real line of a complex function $g(t)$. This follows from the residue theorem (the singularities of $g$ are $\pm i$). We shall not prove. Thus,

$$\varphi_{X+X}(t) = \varphi_{2X}(t) = \varphi_X(2t) = \varphi_X(t)\varphi_X(t).$$

3. $\varphi_X$ is uniformly continuous on $\mathbb{R}$. That is, for every $\epsilon > 0$ there is $\delta > 0$ so that for all $s, t$ if $|s-t| < \delta$ then $|\varphi_X(t) - \varphi_X(s)| < \epsilon$. Indeed, let $M = M(\epsilon)$ be so that

$$\Pr(|X| > M) < \epsilon$$

then

$$\begin{aligned}
|\varphi_X(t) - \varphi_X(s)| &= \left|\mathbb{E}e^{itX} - e^{isX}\right| \\
&= \left|\mathbb{E}e^{isX}(e^{i(t-s)X} - 1)\right| \\
&\leq \mathbb{E}|e^{i(t-s)X} - 1| \\
&= \Pr(|X| > M)\mathbb{E}(|e^{i(t-s)X} - 1|\big||X| > M) \\
&\quad + \Pr(|X| \leq M)\mathbb{E}(|e^{i(t-s)X} - 1|\big||X| \leq M) \\
&\leq 2\Pr(|X| > M) + \mathbb{E}(|e^{i(t-s)X} - 1|\big||X| \leq M) \\
&\leq 2\epsilon + \epsilon,
\end{aligned}$$

as long as say $\delta M < \epsilon/10$.

4. Its derivatives gives the moments: If $\mathbb{E}|X|^k < \infty$ then $\varphi_X$ is differentiable $k$ times at 0 and
$$\varphi_X^{(k)}(0) = i^k \mathbb{E}X^k.$$
The proof shows that we can switch integration and derivative in this case. It hints at that $\varphi_X$ contains all information about $X$.

5. The map $F_X \mapsto \varphi_X$ is one-to-one.

6. The most important property for us:

   **Theorem 61** (Levy continuity)**.** *Let $X, X_1, X_2, \ldots$ be a sequence of random variables. The following are equivalent:*

   *(a) $\varphi_{X_n} \to \varphi_X$ pointwise.*

   *(b) $X_n \to X$ in distribution.*

   *(c) $(X_n)$ is tight, that is, for every $\epsilon > 0$ there is $M$ so that for all $n$,*
   $$\Pr[|X_n| > M] < \epsilon.$$

## 8.4 Proof of central limit theorem

To prove the theorem, we therefore just need to prove:

**Theorem 62.** *If $X_1, X_2, \ldots$ are i.i.d. with expectation 0 and variance 1 and*

$$Z_n = \sum_{j=1}^{n} X_j/\sqrt{n}$$

*then*

$$\varphi_{Z_n} \to \varphi_Z$$

*pointwise, where* $Z \sim N(0,1)$.

Comment: If $X_1, X_2, \dots$ are $N(0,1)$ then $Z_n$ is normal $N(0,1)$ as well, and so this holds for every $n$. This can be thought of the property of normal variables that underlines the central limit theorem.

*Proof.* Let us start with a proof sketch. I.i.d. implies that

$$\varphi_{Z_n}(t) = \prod_{j=1}^{n} \varphi_{X_j/\sqrt{n}}(t) = (\varphi_{X_1/\sqrt{n}}(t))^n.$$

So we need to understand

$$\varphi_{X_1/\sqrt{n}}(t) = \mathbb{E}e^{iX_1 t/\sqrt{n}} \approx \mathbb{E}1 + iX_1 t\sqrt{n} + (iX_1 t/\sqrt{n})^2/2 = 1 + 0 - t^2/n.$$

Thus,

$$\varphi_{Z_n}(t) \approx (1 - t^2/(2n))^n \to e^{-t^2/2},$$

as needed. We just need to make the $\approx$ accurate. For this we use the following (recall that we do not control the 3rd moment): for every $x \in \mathbb{R}$,

$$|e^{ix} - 1 - ix| \le |x|^2/2$$

and

$$|e^{ix} - 1 - ix + x^2/2| \le |x|^3/6.$$

We need to break $\varphi_{X/\sqrt{n}}(t)$ to 2 parts, where $X$ is say $X_1$. Use the following two estimates:

$$\begin{aligned}\left|e^{itX/\sqrt{n}} - 1 - iXt/\sqrt{n} + X^2t^2/(2n)\right| &\le \left|e^{itX/\sqrt{n}} - 1 - iXt/\sqrt{n}\right| + |X|^2|t|^2/(2n) \\ &\le X^2t^2/n\end{aligned}$$

and

$$\left|e^{itX/\sqrt{n}} - 1 - iXt/\sqrt{n} + X^2t^2/(2n)\right| \le |X|^3|t|^3/(6n^{3/2}).$$

Consider the event $A = \{|X| > \delta\sqrt{n}\}$ for $\delta > 0$ to be determined. Hence, a.s.

$$\left|e^{itX/\sqrt{n}} - 1 - iXt/\sqrt{n} + X^2t^2/(2n)\right| \le \mathbf{1}_A X^2t^2/n + \mathbf{1}_{A^c}|X|^3|t|^3/(6n^{3/2}).$$

Take expectation and get

$$\mathbb{E}\left|e^{itX/\sqrt{n}} - 1 - iXt/\sqrt{n} + X^2t^2/(2n)\right| \leq (t^2/n)\mathbb{E}(\mathbf{1}_A X^2) + |t|^3/(6n^{3/2})\mathbb{E}(\mathbf{1}_{A^c}|X|^3).$$

We would like to prove that for every $\epsilon > 0$, if $n$ is large then the expression above is at most $\epsilon/n$. (Plugging this to the proof sketch we started with completes the proof.)

Estimate the second term by:

$$\leq |t|^3/(6n^{3/2})\mathbb{E}(\mathbf{1}_{A^c}|X|^2\delta\sqrt{n}) \leq \delta|t|^3/(6n) \leq \epsilon/n. \quad (\mathbb{E}X^2 = 1, \text{ for small } \delta = \delta(\epsilon, t))$$

Estimate the first term as follows: When $n \to \infty$, the term

$$\mathbb{E}(\mathbf{1}_A X^2)$$

is a tail of a converging sum, so for large enough $n$ it is at most $\epsilon$. For large $n$, the first term is at most

$$\leq (t^2/n)\mathbb{E}(\mathbf{1}_A X^2) \leq \epsilon/n.$$

□

## 8.5 Discussion

The central limit theorem is one of the most useful mathematical theorems, and has applications is all areas of science where statistics is applied.

The theorem it self does not guarantee the rate of convergence. This makes the version we proved not so useful in formal applications, but luckily many scientific areas do not care much about it (e.g. 30 experiments can be considered enough).

However, if we allow control of the 3rd moment, then we do get an estimate on the rate of convergence.

**Theorem 63** (Berry-Esseen)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with expectation 0, variance 1 and $\mathbb{E}|X|^3 < \infty$. Let $Z_n = (X_1 + \ldots + X_n)/\sqrt{n}$. Let $Z \sim N(0,1)$. Then, for all $x, n$,*

$$|F_{Z_n}(x) - F_Z(x)| < \frac{C\mathbb{E}|X|^3}{\sqrt{n}},$$

*where $C$ is a constant.*

We shall not prove, but one can use Fourier analysis to prove as well.

There are also different methods of proof of the central limit theorem. We mention one (Lindberg's method). We want to show that $(X_1+X_2+\ldots+X_n)/\sqrt{n}$ is close to $Z$. As

we mentioned, $Z$ is distributed as $(Y_1+Y_2+\ldots+Y_n)/\sqrt{n}$ where $Y_1, Y_2, \ldots$ are i.i.d. $N(0,1)$. So we want to show that $(X_1 + X_2 + \ldots + X_n)/\sqrt{n}$ is close to $(Y_1 + Y_2 + \ldots + Y_n)/\sqrt{n}$. This can roughly be achieved as follows: We use a hybrid argument showing that

$$(X_1 + X_2 + \ldots + X_i + Y_{i+1} + \ldots + Y_n)/\sqrt{n}$$

is close to

$$(X_1 + X_2 + \ldots + X_i + X_{i+1} + Y_{i+2} + \ldots + Y_n)/\sqrt{n}$$

for all $i$, and then sum the distances. There are also more methods (Stein's, moments...).

One can use the CLT to estimate the size of certain sets. As you saw, the size of $A = \{S \subseteq [n] : n/2 \leq |S| \leq n/2 + 3\sqrt{n}\}$ is for large $n$ close to

$$2^n \int_0^6 \frac{1}{\sqrt{2\pi}} e^{-x^2} dx.$$

Indeed, $|A|/2^n$ is the probability that

$$\sum_{i=1}^{n} X_i \in [n/2, n/3 + 3\sqrt{n}]$$

or

$$Z_n = \sum_{i=1}^{n} \frac{X_i - 1/2}{\sqrt{n}/2} \in [0, 6]$$

for $X_1, X_2, \ldots, X_n$ i.i.d. $Ber(1/2)$.

Another simple example is the behavior of a one dimensional random walk. Let $W_0, W_1, \ldots$ be defined as

$$W_0 = 0$$

and

$$W_t = \sum_{i=1}^{t} X_i,$$

where $X_1, X_2, \ldots$ are i.i.d., each uniform in $\{-1, 1\}$. $W_t$ is the position of a walker that steps right or left independently at time $t$, when started at 0. In general, $W_t \in [-t, t]$. However, the central limit theorem tells us that in most cases $|W_t|$ is roughly $\sqrt{t} \ll t$.

Specifically, we know that if $t$ is large then

$$\Pr(|W_t| < \sqrt{t}) > 0.9 \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 0.95.$$

We can deduce that

$$\binom{t}{t/2} \geq \frac{2^t}{10\sqrt{t}}$$

for large $t$, since the middle binomial coefficient is maximum, and there are roughly $2\sqrt{t}$ coefficients that sum to approximately $0.95 \cdot 2^t$.

Lastly, let us talk about a property of the normal distribution that distinguishes it. The principle of maximum entropy roughly says that the distribution that best describes a situation is the one that maximizes the entropy under what is currently known. We shall not define entropy, but without any prior knowledge the distribution that maximizes the entropy among all distributions on $[n]$ is the uniform distribution. The notion of entropy for continuous random variables is captured by: If $g$ is a density then its entropy is

$$\int g(x) \ln(1/g(x))dx,$$

if it is defined. The observation is that among all densities with expectation 0 and variance 1, the one that maximizes the entropy is that of $N(0,1)$, which we denote by $f$. First,

$$\int f(x) \ln(1/f(x))dx = 1/2.$$

Second, indeed,

$$\begin{aligned}
\int g(x) \ln(1/g(x))dx - 1/2 &= \int g(x) \ln(1/g(x))dx - \int g(x) \ln(1/f(x))dx \\
&= \int g(x) \ln(f(x)/g(x))dx \\
&\leq \ln\left(\int g(x)f(x)/g(x)dx\right) = 0. \qquad \text{(ln is concave)}
\end{aligned}$$

*Summary.* We have proved the central limit theorem. In the process, we discussed various types of convergence, in various spaces. We used Fourier analysis, but there are also other methods. We saw a few applications.