Probability and statistics
Prof. Emmanuel Abbé — EPFL

Notes by Joachim Favre

Computer science bachelor — Semester 4
Spring 2023

I made this document for my own use, but I thought that typed notes might be of interest to others. So, I shared it (with you, if you are reading this!); since it did not cost me anything. I just ask you to keep in mind that there are mistakes, it is impossible not to make any. If you find some, please feel free to share them with me (grammatical and vocabulary errors are of course also welcome). You can contact me at the following e-mail address:

joachim.favre@epfl.ch

If you did not get this document through my GitHub repository, then you may be interested by the fact that I have one on which I put my typed notes. Here is the link (go take a look in the "Releases" section to find the compiled documents):

https://github.com/JoachimFavre/EPFLNotesIN

Please note that the content does not belong to me. I have made some structural changes, reworded some parts, and added some personal notes; but the wording and explanations come mainly from the Professor, and from the book on which they based their course.

I think it is worth mentioning that in order to get these notes typed up, I took my notes in LaTeXduring the course, and then made some corrections. I do not think typing handwritten notes is doable in terms of the amount of work. To take notes in LaTeX, I took my inspiration from the following link, written by Gilles Castel. If you want more details, feel free to contact me at my e-mail address, mentioned hereinabove.

https://castel.dev/post/lecture-notes-1/

I would also like to specify that the words "trivial" and "simple" do not have, in this course, the definition you find in a dictionary. We are at EPFL, nothing we do is trivial. Something trivial is something that a random person in the street would be able to do. In our context, understand these words more as "simpler than the rest". Also, it is okay if you take a while to understand something that is said to be trivial (especially as I love using this word everywhere hihi).

Since you are reading this, I will give you a little advice. Sleep is a much more powerful tool than you may imagine, so never neglect a good night of sleep in favour of studying (especially the night before the exam). I will also take the liberty of paraphrasing my high school philosophy teacher, Ms. Marques, I hope you will have fun during your exams!

Version 2023–06–13

To Gilles Castel, whose work has
inspired me this note taking method.

Rest in peace, nobody
deserves to go so young.

Contents

List of lectures

Chapter 1

Summary by lecture

- Presentation of some of the interesting applications of probabilities and statistics.

- Explanation of combinatorics for when repetitions are or are not allowed, and when ordre does or does not matter.
- Definition of binomial coefficient, and explanation of some of its properties.
- Explanation on how to solve problems which require generalised order without repetitions.

- Definition of sample spaces.
- Definition of event spaces, and explanation of their properties.
- Definition of probability distributions, and explanation of their properties.
- Explanation of the inclusion-exclusion formula.
- Definition of probability spaces.
- Definition of equiprobable events, and some examples.

- Definition of conditional probabilities.
- Explanation of the law of total probability and Bayes' theorem.

- Definition of independence.
- Definition of mutually independence, pairwise independence and conditional independence.
- Explanation of Simpson's paradox.

- Definition of discrete random variable.
- Definition of Bernoulli random variable.
- Definition of the probability mass function of some random variable.

- Definition of binomial, geometric, negative binomial, hypergeometric discrete uniform, and Poisson distributions.

Lecture 7 : High expectations — Tuesday 14th March 2023 *p. 34*

- Definition of CDF, and explanation of some of its properties.
- Proof of a formula giving the PMF of a function.
- Definition of expected value.
- Proof of a formula giving the expected value of a function.

Lecture 8 : Le poisson du moment — Thursday 16th March 2023 *p. 36*

- Proof of some properties of the expected value.
- Definition of moments, central moments and factorial moments.
- Definition of variance and standard deviation.
- Explanation of some properties of the variance.
- Proof of a theorem allowing to compute the expected value.

Lecture 9 : Going continuous — Tuesday 21st March 2023 *p. 39*

- Definition of conditional probability mass distributions, and proof that they indeed follow the axioms of PMFs.
- Definition of conditional expected value.
- Definition of convergence in distribution.
- Proof of the law of small numbers.
- Definition of continuous random variables, and their PDF and CDF.

Lecture 10 : Transformations — Thursday 23rd March 2023 *p. 44*

- Definition of uniform, exponential, gamma, Laplace and Pareto random variables.
- Definition of expectation, variance and conditional densities for continuous random variables.
- Definition of quantile.
- Example of transformations.

Lecture 11 : The Professor did not know the error function ☹ — Tuesday 28th March 2023 *p. 47*

- Proof of the theorem of general transformation.
- Definition of the normal distribution, and explanation of its properties.
- Explanation of De Moivre-Laplace's theorem.
- Definition of joint PMF, PDF and CDF for several random variables.

Lecture 12 : Generalisation of our definitions — Thursday 30th March 2023 *p. 53*

- Definition of marginal PMF and PDF, and proof that they work like we want them to.
- Definition of conditional PMF and PDF for an arbitrary set of random variable.
- Definition of the multinomial distribution.

- Explanation and proof of Markov's inequality and some of its corollaries.
- Explanation of Jensen's inequality.
- Explanation of Hoeffding's inequality.
- Definition of almost sure, mean square, probability and distribution convergence, and explanation of which implies which.

- Explanation that a random variable converging in distribution to a constant also converges in distribution.
- Explanation of Slutsky's lemma.
- Explanation and justification of the weak law of large numbers.
- Explanation and justification of the central limit theorem.

- Introduction to statistical inference.
- Definition of the method of moments.
- Definition of the maximum likelihood method.

- Definition of the M-estimation method.
- Definition of the bias of an estimator.
- Definition of the MSE of an estimator, and proof of its link with variance and biases. Confidence intervals the efficiency of estimators.
- Definition of pivot.
- Explanation of how to construct a confidence interval.

- Definition of the standard error, and proof of the empirical CLT.
- Definition of null and alternative hypothesis.
- Definition of false positive and false negative.
- Definition of the Chi-square distribution.
- Definition of Pearson statistic, and proof of its link with the Chi-square distribution.

- Definition of P-value.
- Definition of test significance.

- Explanation and proof of the Neymnan-Pearson lemma.
- Explanation of optimal hypothesis test under average error probability, and computation of a nice way to compute the average error probability.

Tuesday 21st February 2023 — **Lecture 1 : An interesting start**

Chapter 2

Combinatorics

Proposition: Order with repetitions

The number of ways to choose k elements amongst n distinct elements, when order matters and we allow repetitions, is given by:

$$n^k$$

Justification We have n possibilities for the first element, n for the second one, and so on until the k^{th} element. This gives us:

$$n \cdot n \cdots n = n^k$$

Proposition: Order without repetitions

The number of ways to choose k elements amongst n distinct elements, when order matters but we don't want any repetition, is given by:

$$\frac{n!}{(n-k)!}$$

Justification The idea is that there are n possibilities for the first element. Then, there are $n-1$ possibilities for the second one (since we want any element, except for the first one), and so on until $n-(k-1)$ (since we want k elements in total). This gives us:

$$n(n-1)\cdots(n-(k-1)) = \frac{n!}{(n-k)!}$$

Proposition: No order without repetitions

The number of ways to choose k elements amongst n distinct elements, when order does not matter and we don't want any repetition, is given by:

$$\frac{n!}{(n-k)!k!} = \binom{n}{k}$$

called the **binomial coefficient**.

Justification The idea is that we consider the case where we allow order, and we just divide it by $k!$ (this is the number of different permutations since every element is unique). This gives us:

$$\frac{n!}{(n-k)!k!}$$

Properties of binomial coefficients

Let $n, m, r \in \mathbb{N}$ where $r \leq n$. Then, we have the following properties:

1. $\binom{n}{r} = \binom{n}{n-r}$.

2. $\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$.
3. $\binom{m+n}{r} = \sum_{j=0}^{r} \binom{m}{j}\binom{n}{r-j}$.
4. $(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$, which is also known as Newton's binomial theorem.
5. $(1-x)^{-n} = \sum_{j=0}^{\infty} \binom{n+j-1}{j} x^j$ for some $|x| < 1$, which is the generalisation of the last line to negative powers.
6. $\lim_{n\to\infty} n^{-r}\binom{n}{r} = \frac{1}{r!}$ for some finite r.

Indeed:

1. The number of ways of choosing r objects from n is the same as the number of ways of choosing the $n-r$ elements we would leave behind.
2. To choose r objects from $n+1$, we can consider one element to be special: we consider the case where we take this element and thus we have to take $r-1$ elements out of the $n-1$ left, and the case where we don't take the element, and thus that we need to take r elements from the n left. This property yields that $\binom{n}{r}$ is the coefficient in row n and column r of Pascal's triangle.
3. Let us consider that we have n blue hats and m red hats. The number of ways to choose r hats from our total is the sum on all possible j of taking j blue hats and $r-j$ red hats.

Remark The first property must be known, the others might be useful at some point.

Proposition: No order with repetitions The number of ways to choose k elements amongst n distinct elements, when order does not matter but we allow repetitions, is given by:

$$\binom{n-1+k}{k}$$

Justification We need to switch our viewpoint in order to understand this well. The goal is basically to take a total of k elements from n buckets, which we can represent as a n-tuple: the i^{th} element represents the number of objects we took from the i^{th} box. Since we want to pick a total of k objects, the sum of the elements of the tuple is equal to k.
We can encode this as a string: we use $n-1$ characters as splitters (for instance "|"), and k characters as objects we picked from a box (for instance "•"). For example, $(1,2,0,1)$ could be represented by •| • •||• (there are one object in the first delimited box, 2 in the second one, then 0 and finally 1, indeed representing our vector). This is a bijective representation: any tuple can be represented as such a string, and any string represents exactly one tuple. This means that we only need to count the number of possible strings.
Our string has length $n+k-1$, with k and $n-1$ repeated elements. This gives us:

$$\frac{(n+k-1)!}{n!(k-1)!} = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

using a property of the binomial coefficient.

Summary The number of ways to choose k elements amongst n distinct elements is given by:

	Order does not matter	Order matters
Repetitions not allowed	$\frac{n!}{(n-k)!k!} = \binom{n}{k}$	$\frac{n!}{(n-k)!}$
Repetitions allowed	$\binom{n-1+k}{k}$	n^k

Remark The purpose of this table is not to be learnt by heart and to just try and find the correct to use, but to be able to know their intuition, and to be able to find such formula back during an exercise.

Proposition: Generalised order without repetitions

Let's say we have $n = n_1 + \ldots + n_r$ elements of r different types, where n_i is the number of objects of type i that are indistinguishable from one another. Then, the number of permutations without repetitions is given by:

$$\frac{n!}{n_1! \cdots n_r!} = \binom{n}{n_1, \ldots, n_r}$$

called the **multinomial coefficient**.

Justification Let's think of it though an example: let's say we have a category a with a_1 and a_2 (which are really undistinguishable, but we distinguish them here for understanding purposes), and another category b with element b_1. If we just compute the basic order without repetitions, we get 3! possibilities. However, we count $a_1a_2b_1$ and $a_2a_1b_1$ separately. This should not be the case (again, a_1 and a_2 are indistinguishable). We can just fix this by dividing by the number of permutations of the elements within the category a (it's like if we fix everything else, and we just permute elements of a on their own).
Generalising this idea is not very complicated, and it gives us:

$$\frac{n!}{n_1! \cdots n_r!}$$

Indeed, we permute all those elements, and then we remove the duplicate within each category.

Observation Note that, if $n_i = 1$ for all i, then we get back our formula for order without repetitions where we pick $k = n$ elements.

Example 1 Let's say we have a class of 20 students, where we have to choose a committee of size 4.
If each person of the committee have a different role, then we need $20 \cdot 19 \cdot 18 \cdot 17 = \frac{20!}{(20-4)!}$ possibilities.
If there is one president, one treasurer, and two travel agents, then we have $\frac{20 \cdot 19 \cdot 18 \cdot 17}{1!1!2!}$ since we must not count who is there is no agent 1 and agent 2, the two jobs are exactly the same.
If there roles are indistinguishable, then we just need to pick 4 people out of our 20 students, giving $\binom{20}{4} = \frac{20!}{(20-4)!4!}$.

Example 2 Let's say that we have an integer n, which we want to split into $n = n_1 + \ldots + n_r$ where $n_i \geq 0$ for all i. We want to know the number of ways we can find such numbers.
We recognise the intermediate step in our justification for no order with repetition, meaning that we can use the same reasoning. We can represent our numbers as $|$, giving us strings such as $|||+||++| = 3+2+0+1 = 6$ for instance. We need a total of n bars (since we want them to add to n), and we have $r-1$ plus symbols, giving us $\binom{n+r-1}{r-1}$ possibilities.

Example 3 Let's consider another example, which is very close. This time, we want $n_i > 0$ for all i.
A simple way to do so, is to consider the change of variable $m_i = n_i - 1$. Since $n_i > 0 \iff n_i \geq 1$, we get that we want $m_i \geq 0$ for all i, and:

$$m_1 + \ldots + m_r = (n_1 - 1) + \ldots + (n_r - 1) = n_1 + \ldots + n_r - r = n - r$$

Then, we can do the exact same reasoning as before, giving $\binom{n-1}{r-1}$.

Chapter 3

Probability

3.1 Probability space

Goal

The goal of this part is to make the definition of a probability space. This will first require the definition of three other objects.

Definition: Sample space

The **sample space** of an experiment, noted Ω, is the set of all possible events that can take place in it.
It can be finite, countable or uncountable.

Definition: Event space

An **event space** (or sigma-algebra, or *tribu* in French) $\mathcal{F}$ is a set of subsets of Ω such that:

- $\mathcal{F}$ is nonempty.
- If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$ (where $A^C = \Omega \setminus A$ is the complement of A).
- If $\{A_i\}_{i=1}^{\infty}$ are all elements of $\mathcal{F}$, then:

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

Each element of $A \in \mathcal{F}$, $A \subseteq \Omega$, is named an **event**.

> *Remark* This basically represents the set of interesting events from Ω.
> Typically, if Ω is countable, we can always take $\mathcal{F}$ to be its power set (the set of subsets of Ω). This will always be the biggest sigma-algebra we can define, and another will be a subset of it. In other words, any question we can ask ourselves about the experiment can always be translated into an element of the power set.
> However, if Ω is not countable, then, in general, we cannot just take the power set of Ω. For instance, the power set of $[0, 1]$ is not a sigma algebra.

Properties

An event space $\mathcal{F}$ has the following properties, thanks to its axioms:

- $\bigcup_{i=1}^{n} A_i \in F$
- $\Omega \in \mathcal{F}, \emptyset \in \mathcal{F}$
- $A \cap B \in \mathcal{F}, A \setminus B \in \mathcal{F}, A \ \Delta \ B \in \mathcal{F}$ where Δ is the "xor" of the sets (the symmetric difference).
- $\bigcap_{i=1}^{n} A_i \in F$

Proof 1 To prove the first property, we can just take $A_{n+1} = A_{n+2} = \dots$ and set them to be equal to A_n. Then, we only need to apply the third axiom.

Proof 2 If $\mathcal{F}$ is non-empty, then it has an element A. By the second axiom, $A^C \in \mathcal{F}$. However, by the third property:

$$A \cup A^C = \Omega \in \mathcal{F}$$

Finally, by the second axiom:

$$\Omega^C = \emptyset \in \mathcal{F}$$

Proof 2 and 3 We can write those properties using unions and complements, and they are thus yielded by the previous axioms and properties.

Example 1 Let us consider the sample space of throwing a coin, $\Omega = \{H, T\}$. Then, we can pick either of the two event space:

$$\mathcal{F}_1 = \{\{H, T\}, \emptyset\}, \quad \mathcal{F}_2 = \{\{H, T\}, \{H\}, \{T\}, \emptyset\}$$

Naturally, only the second one is useful.

Remark In general, the following is always a valid sigma algebra:

$$\{\Omega, \emptyset\}$$

Example 2 Let us consider an event where we throw a red die and a green die. Also let A be the event that the red die shows a 4 and B the event that the sum is odd.
Then, $A \cap B$ is the event that the red die shows a 4 and that the sum is odd.

Definition: Probability distribution A **probability distribution** $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ assigns a probability to each element of the event space $\mathcal{F}$, with the following properties:

1. For all $A \in \mathcal{F}$, then $0 \leq \mathbb{P}(A) \leq 1$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $\{A_i\}_{i=1}^{\infty}$ are pairwise disjoint (meaning, for $i \neq j$, then $A_i \cap A_j$), then:

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

The fact that two events are disjoint means that they cannot happen at the same time. The third axiom thus means that, if some events cannot happen at the same time, the probability that any of them happen is just the sum of probabilities.

Properties Let $A, B, \{A_i\}_{i=1}^{\infty}$ be events of some event space. Then:

1. $\mathbb{P}(\emptyset) = 0$
2. $\mathbb{P}\left(A^C\right) = 1 - \mathbb{P}(A)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
4. If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = P(A) + P(B)$.
5. If $A \subset B$, then $P(A) \leq P(B)$ and $P(B \setminus A) = P(B) - P(A)$.
6. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, known as Boole's inequality or the union bound.
7. If $A_1 \subset A_2 \subset \dots$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right)$
8. If $A_1 \supset A_2 \supset \dots$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right)$

The third property is very important. It means that if add the two probabilities, we will have counted their intersection twice. Also, the union bound is used very often

(using it is typically much easier than using the inclusion-exclusion formula (which comes right after)).

Inclusion-exclusion formula

If $A_1, \dots, A_n$ are events of some event space, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < \dots < i_r \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_r})$$

This formula has $2^n - 1$ terms.

Example For instance, for $n = 2$:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

Or, for $n = 3$:

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) \\ + \mathbb{P}(A_1 \cap A_2 \cap A_3)\end{aligned}$$

Remark This formula can typically be found back by drawing a Venn diagram, and being careful on what parts of the diagram we counted, and how many times.
Also, note that we should almost never use this formula when n is big (say more than 4), since it quickly becomes unmanageable.

Proof This proof can be done by induction.

Example 1 We want to know the probability of getting at least one 6 when rolling three fair dice.
Let A_i be the event there is a 6 on die i; we want to compute $\mathbb{P}(A_1 \cup A_2 \cup A_3)$. However, by symmetry, $\mathbb{P}(A_i) = \frac{1}{6}$, $\mathbb{P}(A_i \cap A_j) = \frac{1}{6^2}$ (the probability that die i and die j yield both a 6 is $\frac{1}{6^2}$), and $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{6^3}$. Thus, by the inclusion-exclusion formula:

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = 3 \cdot \frac{1}{6} - 3 \cdot \frac{1}{6^2} + \frac{1}{6^3} = \frac{91}{216}$$

Remark In this case, it is better to consider the complement: the event that no die rolls a 6. Since the die are independent, we simply get that:

$$\mathbb{P}\big(B^C\big) = \frac{5^3}{6^3} \implies \mathbb{P}(B) = 1 - \mathbb{P}\big(B^C\big) = 1 - \frac{5^3}{6^3} = \frac{91}{216}$$

Example 2 We want to compute the probability that a random number from 1 to 1000 is divisible by 2, 3 or 5.
Let D_i be the property that the number is divisible by i. We thus want to compute $\mathbb{P}(D_2 \cup D_3 \cup D_5)$. Also, we notice that $\mathbb{P}(D_i) = \left\lfloor \frac{1000}{i} \right\rfloor / 1000$ since there are $\left\lfloor \frac{1000}{i} \right\rfloor$ numbers divisible by i from 1 to 1000. We can apply the inclusion-exclusion formula:

$$\begin{aligned}\mathbb{P}(D_2 \cup D_3 \cup D_5) &= \mathbb{P}(D_2) + \mathbb{P}(D_3) + \mathbb{P}(D_5) - \mathbb{P}(D_2 \cap D_3) \\ &\quad - \mathbb{P}(D_2 \cap D_5) - \mathbb{P}(D_3 \cap D_5) + \mathbb{P}(D_2 \cap D_3 \cap D_5) \\ &= \mathbb{P}(D_2) + \mathbb{P}(D_3) + \mathbb{P}(D_5) - \mathbb{P}(D_6) - \mathbb{P}(D_{10}) - \mathbb{P}(D_{15}) + \mathbb{P}(D_{30}) \\ &= \frac{500 + 333 + 200 - 166 - 100 - 66 + 33}{1000} \\ &= \frac{367}{500} \\ &= 0.734\end{aligned}$$

Definition: Probability space A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a mathematical object associated with a random experiment, where Ω is its sample space, $\mathcal{F}$ is its event space, and $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$ is its probability distribution.

Remark Any random experiment is modelled by such a probability space. In exams (and exercises), it is important to always give the sample space and the event space.
In other words, this definition is really important.

3.2 Equiprobable events

Equiprobable events If it is finite, we typically try to take the sample space Ω so that each of its elements $\omega \in \Omega$ is equiprobable, i.e:

$$\mathbb{P}(\omega) = \frac{1}{|\Omega|}, \quad \forall \omega \in \Omega$$

In this case, we simply have:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}, \quad \forall A \subseteq \Omega$$

Example 1 Let us consider that we throw two fair dice, one red and one green. The sample space of this experiment is:

$$\Omega = \{11, 12, \dots, 16, 21, 22, \dots, 66\}$$

where the first letter is the outcome of the red dice and the second one is the outcome of the green one. We can take $\mathcal{F}$ to be the power set of Ω.
The event that at least a die outputs a 1 is given by:

$$A = \{11, 12, 13, 14, 15, 16, 21, 31, 41, 51, 61\} \subseteq \mathcal{F}$$

Then, the probability that A takes place is given by:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{2 \cdot 6 - 1}{6^2} = \frac{11}{36}$$

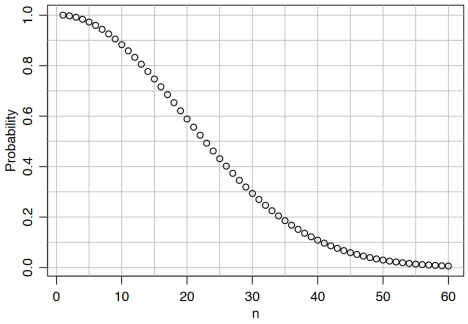
Example 2 We wonder what is the probability that all n people in a room have a different birthday.
The sample space can be written as $\Omega = \{1, \dots, 365\}^n$, where each of those possibilities have probability 365^{-n}. We seek the probability of the following event:

$$A = \{(i_1, \dots, i_n) \mid i_1 \neq i_2 \dots \neq i_n\}$$

We can see that $|A| = \frac{365!}{(365-n)!}$, since there are 365 possibilities for the first person, 364 for the second one, and so on. This gives us that:

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\frac{365!}{(365-n)!}}{365^n} = \frac{365!}{(365-n)!365^n}$$

This curve decays very rapidly. This may seem a bit counterintuitve, hence it is named the birthday paradox.



Example 3 We throw three dice, and we let T_i be the event of the sum being equal to i. We wonder if T_9 or T_{10} is more likely (meaning is it more like to have a sum of 9 or 10). Our sample space is given by:

$$\Omega = \{(r,s,t) \mid r,s,t = 1,\dots,6\} \implies |\Omega| = 6^3$$

We can notice that:

$$9 = 6+2+1 = 5+3+1 = 5+2+2 = 4+4+1 = 4+3+2 = 3+3+3$$

$$10 = 6+3+1 = 6+2+2 = 5+4+1 = 5+3+2 = 4+4+2 = 4+4+3$$

There is the same number of ways to sum to 9 and to sum to 10. However, the die are distinguishable, meaning that there is exactly one way to make $3+3+3$, 3 to make $4+4+1$ and 3! to make $5+3+1$. In other words, we need to be careful in our computations. If we do them correctly, we get:

$$|T_9| = 25, \quad |T_{10}| = 27$$

We thus get that T_{10} is more probable.

— Thursday 2nd March 2023 — **Lecture 4 : Computing the probability people are terrorists**

3.3 Conditional probabilities

Definition: Conditional probability Let A and B be events of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $\mathbb{P}(B) > 0$. Then, the **conditional probability** of A given B (knowing that B took place) is:

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

If $\mathbb{P}(B) = 0$, we are still allowed to consider the following quantity (since both sides are equal to 0, ignoring that $\mathbb{P}(A|B)$ is undefined):

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

Observation We note that any set A can be split into:

$$A = (A \cap B) \cup \left(A \cap B^C\right)$$

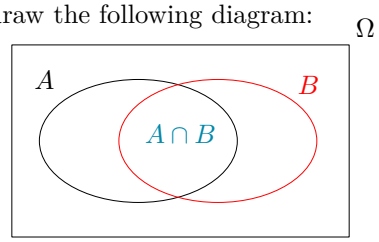
where $A \cap B$ and $A \cap B^C$ are disjoint.
This yields that:

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}\left(A \cap B^C\right) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}\left(A|B^C\right)\mathbb{P}\left(B^C\right)$$

Note that this is correct even if $\mathbb{P}(B) = 0$ or $\mathbb{P}\left(B^C\right) = 0$. Reading this formula orally should make it very intuitive.

Intuition

The idea is to draw the following diagram:



Since we know that B happened, this is like if our new sample space is B and that we are looking for the event $A \cap B$.

Example

We roll two fair dice, one red and one green. Let A and B be the events that the total exceeds 8, and that we get 6 on the red die, respectively. We want to compute the probability of A knowing that B has occurred.
This can easily be computed thanks to our definition:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{4}{6^2}}{\frac{6}{6^2}} = \frac{2}{3}$$

This is very intuitive: if we know that we rolled a 6 on the red die, then the green die must roll any number which is not 1 or 2.
Note that we can show $\mathbb{P}(A) = \frac{5}{18}$. We indeed see that conditioning the problem may have a big result on the probabilities.

Theorem: Conditional probability space

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Also, let $\hat{\mathbb{P}}(A) = \mathbb{P}(A|B)$.
Then $\left(\Omega, \mathcal{F}, \hat{\mathbb{P}}\right)$ is a probability space. In particular:

- If $A \in \mathcal{F}$, then $0 \leq \hat{\mathbb{P}}(A) \leq 1$
- $\hat{\mathbb{P}}(\Omega) = 1$
- If $\{A_i\}_{i=1}^{\infty}$ are pairwise disjoint, then:

$$\hat{\mathbb{P}}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{j=1}^{\infty} \hat{\mathbb{P}}(A_i)$$

Theorem: Law of total probability

Let $\{B_i\}_{i=1}^{\infty}$ be pairwise disjoint events (meaning that $B_i \cap B_j = \emptyset$ for $i \neq j$) of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Also, let A be an event satisfying $A \subset \bigcup_{i=1}^{\infty} B_i$.
Then:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Intuition

Basically, if the B_i cover entirely A, then we can express A as a union of $B_i \cap A$, which are pairwise disjoint (since the B_i originally were pairwise disjoint).

Remark

We have seen a special case of this theorem above, in the definition of conditional probabilities, with B and B^C:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}\left(A|B^C\right)\mathbb{P}\left(B^C\right)$$

Bayes' theorem

Let $\{B_i\}_{i=1}^{\infty}$ be pairwise disjoint events (meaning that $B_i \cap B_j = \emptyset$ for $i \neq j$) of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Also, let A be an event satisfying $A \subset \bigcup_{i=1}^{\infty} B_i$ and $\mathbb{P}(A) > 0$.
Then, for any $j \in \mathbb{N}$:

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Remark 1

Note that this result is also true if the number of B_i is finite, and if the B_i partition Ω (which is a typical application of this theorem).

Remark 2 This theorem is very important. There always are questions on this theorem at exams.

Proof This proof is very straighforward, using our definitions and our law of total probability:

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A \cap B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{\infty}\mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Example We suspect that the man in front of us at the security check at the airport is a terrorist. We know that one person out of 10^6 is a terrorist, and that a terrorist is detected by security check with a probability of 0.9999, but that the alarm may still be triggered when a normal person goes through with a probability 10^{-5}. We wonder what is the probability that he is a terrorist, given that the alarm is triggered when he passes through security.
Let A be the event that the alarm is triggered and T be the one that he is a terrorist. We can apply Bayes' rule:

$$\mathbb{P}(T|A) = \frac{\mathbb{P}(A|T)\mathbb{P}(T)}{\mathbb{P}(A|T)\mathbb{P}(T) + \mathbb{P}(A|T^C)\mathbb{P}(T^C)} = \frac{0.9999 \cdot 10^{-6}}{0.9999 \cdot 10^{-6} + 10^{-5} \cdot (1 - 10^{-6})}$$

Which is approximately:

$$\mathbb{P}(T|A) \approx 0.0909 = 9.09\,\%$$

This shows that $\mathbb{P}(T|A) \neq \mathbb{P}(A|T)$, rather unintuitively if we are not careful (it is a typical bias in statistics to assume they are equal).

Tuesday 7th March 2023 — **Lecture 5 : Success rate at Trum school**

Theorem: Prediction decomposition Let $A_1, \dots, A_n$ be events in probability space. Then:

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \left[\prod_{i=2}^{n} \mathbb{P}(A_i|A_1 \cap \cdots \cap A_{i-1})\right]\mathbb{P}(A_1)$$

Example For instance, we have already seen:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1)$$

But also, for three events:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_3|A_1 \cap A_2)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1)$$

3.4 Independence

Definition: Independence Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two event $A, B \in \mathcal{F}$ are defined to be **independent**, written $A \perp\!\!\!\perp B$, if:

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

Remark When two events are independent, we get that:

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

This is in fact an equivalence:

$$\begin{aligned} & \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \\ \iff & \mathbb{P}(A|B) = \mathbb{P}(A) \\ \iff & \mathbb{P}(B|A) = \mathbb{P}(B) \end{aligned}$$

Intuition Saying that A and B are independent means that the occurrence of one of the two does not affect the occurrence of the other. We thus indeed intuitively get that we should have $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Example A family has two children, giving the following sample space (the first letter represents the sex of the first child, and similarly for the second):

$$\Omega = \{BB, BG, GB, GG\}$$

If we know that the first child is a boy (represented by the event $B_1 = \{BB, BG\}$), then the probability that the second child is a boy (the event $B_2 = \{GB, BB\}$) is:

$$\mathbb{P}(B_2|B_1) = \frac{\mathbb{P}(B_1 \cap B_2)}{\mathbb{P}(B_1)} = \frac{\mathbb{P}(\{BB\})}{\mathbb{P}(B_1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2} = \mathbb{P}(B_2)$$

showing that those two events are actually independent.
However, if now we know that at least a child is a boy (the event $C = B_1 \cup B_2$) and that we want the probability that both are boys (the event $D = \{BB\}$), we get:

$$\mathbb{P}(D|C) = \frac{\mathbb{P}(C \cap D)}{\mathbb{P}(C)} = \frac{\mathbb{P}(D)}{\mathbb{P}(C)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \neq \mathbb{P}(D)$$

This shows that those are not independent, rather unintuitively (this is known as the Two Child Paradox).

Remark Independent does not mean disjoint.
Disjoint means that two events cannot happen at the same time. For instance, if we roll a die, the event of getting a 3 and the one of getting a 6 are disjoint. In this case, we have:

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0 \iff \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

Independent means that one has no influence on the other. For instance, the event of rolling a 6 and that it is sunny outside are independent. In this case, we have:

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

Example Let us consider a pack of cards which is well shuffled. We pick a card at random, and we consider the three following events: the card is an ace A, the card is a king K and the card is an heart H. We see that:

$$\mathbb{P}(A) = \mathbb{P}(K) = \frac{1}{13}, \quad \mathbb{P}(H) = \frac{1}{4} = \frac{13}{52}$$

We can see that A and H are independent:

$$\mathbb{P}(A \cap H) = \frac{1}{52} = \mathbb{P}(A)\mathbb{P}(H)$$

However, A and K are disjoint but not independent (if we know that one happened, the we definitely know that the other did not happened):

$$\mathbb{P}(A \cap K) = \mathbb{P}(\emptyset) = 0 \neq \mathbb{P}(A)\mathbb{P}(K)$$

Types of independence

The events $A_1, \dots, A_n$ are **mutually independent** if, for all sets of indices $F \subset \{1, \dots, n\}$:

$$\mathbb{P}\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \mathbb{P}(A_i)$$

They are **pairwise independent** if:

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j), \quad 1 \leq i < j \leq n$$

They are **conditionally independent given B** if for all sets of indices $F \subset \{1, \dots, n\}$:

$$\mathbb{P}\left(\bigcap_{i \in F} A_i | B\right) = \prod_{i \in F} \mathbb{P}(A_i | B)$$

Remark If we say that events $A_1, \dots, A_n$ are independent, we generally mean that they are mutually independent.

Implications
- Mutual independence implies pairwise independence.
- Pairwise independence only implies mutual independence when $n = 2$.
- Mutual independence neither implies nor is implied by conditional independence.

Remark This subject is really important and almost always has questions during exams.

Example

Let us consider a family with two children, and the events B_1 "the first born is a boy", B_2 "the second child is a boy" and $1B$ "there is exactly one boy".
We have:

$$\mathbb{P}(B_1) = \frac{1}{2}, \quad \mathbb{P}(B_2) = \frac{1}{2}, \quad \mathbb{P}(1B) = \frac{1}{2}$$

We can see that they are pairwise independent:

$$\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1 \cap 1B) = \mathbb{P}(B_2 \cap 1B) = \frac{1}{4}$$

However, they are not mutually independent:

$$\mathbb{P}(B_1 \cap B_2 \cap 1B) = 0 \neq \frac{1}{8} = \mathbb{P}(B_1)\mathbb{P}(B_2)\mathbb{P}(1B)$$

Series-Parallel system

Let's consider an electric system which has components labelled 1 to n, which fail independently of each other. Also, let F_i be the event that the i^{th} component is faulty, with some probability $\mathbb{P}(F_i) = p_i$. The event that the system fails occurs if the current cannot pass from one end to the other (see the picture in the following example).
If the components are arranged in parallel, then for the system to fail all components need to fail:

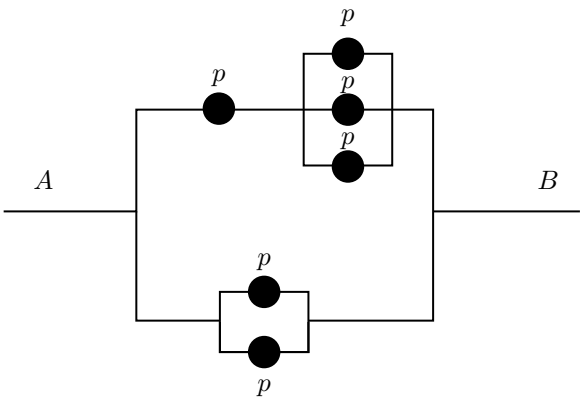
$$\mathbb{P}_P(S) = \mathbb{P}(F_1 \cap \dots \cap F_n) = \prod_{i=1}^{n} p_i$$

If the components are arranged in series, then the system fails as soon as any of the components fail:

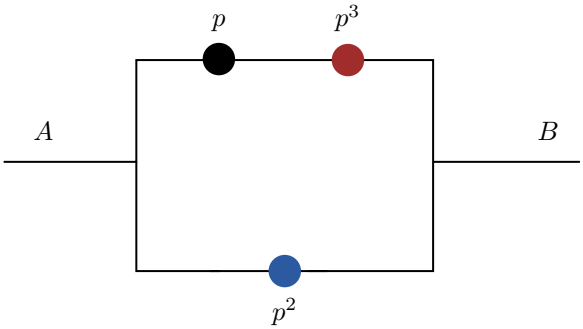
$$P_S(S) = \mathbb{P}(F_1 \cup \dots \cup F_n) = 1 - \mathbb{P}\left(F_1^C \cap \dots \cap F_n^C\right) = 1 - \prod_{i=1}^{n}(1 - p_i)$$

Example

Let us consider the following system:



We want to compute the probability that it fails, knowing that all dots have a fail probability of p.
The idea is, just as for electronics, to replace some circuits by equivalent ones. We can make the following picture, where the red dot has a probability p^3 to fail and the blue one has a probability p^2.



Now, we can merge the two probability from the top line to $1-(1-p)\big(1-p^3\big)$. Finally, we can combine everything to get:

$$p^2\big(1-(1-p)\big(1-p^3\big)\big)$$

Simpson's paradox

Let's say that we read some statistics. *("What's the name of school which does not seem very nice?")* At Trum school, there are 1000 males and 1000 females. The number of success this year was 500 males and 200 females.
However, if really look into the statistics, we could see that there are hard and easy departments, with the following repartition and success rate of people:

Attendance	Hard department	Easy department
Males	1%	99%
Females	99%	1%

Success	Hard department	Easy department
Males	10%	50%
Females	20%	100%

However, in fact, in both cases, the females have always better succeeded than the males; but, we still have $1+990\cdot0.50\approx500$ males who passed and $990\cdot0.2+10\approx200$ females. So, indeed, there were fewer females who passed, but they were in fact better than the males.
This is an important concept because it happens a lot during studies, which leads to many mistakes (including some concerning males and females success at school). We always need to look deeper in the data, to see if there is some categories which can explain the differences.

Chapter 4

Discrete random variables

4.1 Fundamentals

Observation

When analysing some random event, we are often really interested in some values linked to each event (such as the money won at some game, or the number of successes).

Definition: Random variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
A **random variable** $X : \Omega \mapsto \mathbb{R}$ is a function from the sample space Ω to real numbers $\mathbb{R}$.

Definition: Support

The **support** of some random variable X is its codomain (the set of values it can take):

$$D_X = \{x \in \mathbb{R} \mid \exists \omega \in \Omega \text{ such that } X(\omega) = x\}$$

If D_X is countable, then X is named a **discrete random variable**.

Definition: Random variable probability

Let X be some random variable, and $S \subset \mathbb{R}$. Then, we write:

$$\mathbb{P}(X \in S) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in S\})$$

Also, we write:

$$\mathbb{P}(X = x) = \mathbb{P}(X \in A_x)$$

where $A_x = \{\omega \in \Omega \mid X(\omega) = x\} \subset \mathcal{F}$.

Example

We toss a coin multiple times, with probability p to give head, and we note X to be the random variable representing the number of throws until we first get head.
We want to compute $\mathbb{P}(X = 3)$. To do so, we need to first have 2 tails, followed by a head. So:

$$\mathbb{P}(X = 3) = (1-p)p^2$$

We also have that:

$$\mathbb{P}(1.7 \leq X \leq 3.5) = \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = (1-p)p + (1-p)^2 p$$

Definition: Bernoulli random variable

A random variable that only takes the value 0 or 1 is named a **indicator variable** (or **Bernoulli random variable** or **Bernoulli trial**).

Definition: PMF

The **probability mass function** (PMF) of a discrete random variable X is:

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}(A_x), \quad x \in \mathbb{R}$$

where:

$$A_x = \{\omega \in \Omega \mid X(\omega) = x\}$$

Note that this only works since A_x is countable, since the random variable is discrete.

Support We notice that the support of X, also named the support of f_X and noted D_X, is the set of all x such that $f_X(x) > 0$.

Terminology When there is no risk of confusion, we write $f_X = f$ and $D_X = D$.

Properties It has the following properties:

$$f_X(x) \geq 0, \quad \sum_{x_i \in D_X} f_X(x_i) = 1$$

Note that, for now, we always have $f_X(x) \leq 1$ for all x. However, this will no longer be the case for continuous probabilities.

Remark 1 The PMF completely characterises a random variable.

Remark 2 The term "mass" in the PMF will make more sense when we will have seen continuous variable.

Definition: Distribution A **distribution** defines the PMF of some random variable. We use the operator $X \sim D$ to mean that the random variable X follows the distribution D.

Definition: Binomial Let $n \in \mathbb{N}$ and $0 \leq p \leq 1$. A **binomial** random variable X has the following PMF:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n$$

We write $X \sim B(n, p)$ and call n the **denominator** and p the probability of success.

Observation With $n = 1$, this is a Bernoulli variable.

Intuition We repeat n times some experiment with success probability p, and the random variable represents the number of successes.

Definition: Geometric Let $0 \leq p \leq 1$. A **geometric** random variable X has the following PMF:

$$f_X(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

We write $X \sim \text{Geom}(p)$ and we call p the **success probability**.

Intuition We repeat some experiment with success probability p, and the random variable represents the number of times we have to do it until we have the success (counting this last experiment). We indeed first need to do $x - 1$ failures, followed by a success.

Theorem: Lack of memory If $X \sim \text{Geom}(p)$, then:

$$\mathbb{P}(X > n + m | X > m) = \mathbb{P}(X > n)$$

This property is sometimes called **memorylessness**.

Proof We know that $\mathbb{P}(X > n) = (1-p)^n$, since it means we have made at least n failures. Thus:

$$\begin{aligned}\mathbb{P}(X > n+m | X > m) &= \frac{\mathbb{P}(X > n+m \cap X > m)}{\mathbb{P}(X > m)} \\ &= \frac{\mathbb{P}(X > n+m)}{\mathbb{P}(X > m)} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} \\ &= (1-p)^n \\ &= \mathbb{P}(X > n)\end{aligned}$$

□

Example We throw a dice until we get a 6. The probability that we have to make a total of 4 rolls is given by:

$$\mathbb{P}(X = 4) = \left(1 - \frac{1}{6}\right)^3 \frac{1}{6} = \frac{5^3}{6^4}$$

Definition: Negative binomial Let $n \in \mathbb{N}$ and $0 \leq p \leq 1$. A **negative binomial** random variable X has PMF:

$$f_X(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x = n, n+1, \dots$$

We write $X \sim \text{NegBin}(n, p)$.

Observation When $n = 1$, we see that $X \sim \text{Geom}(p)$.

Intuition X models the number of times we need to run an experiment with success probability p to reach n successes.
Indeed, we must have n successes and $n - x$ failures. The x^{th} try must be a success, and we have $\binom{x-1}{n-1}$ ways to pick $n-1$ successes on the $x-1$ other tries.

Example We want to find the probability of seeing 2 heads before 5 tails in a repeated toss of a coin.
Let X be the waiting time for $n = 2$ heads, we want the probability that $\mathbb{P}(X \leq 6)$. We just have:

$$\mathbb{P}(X \leq 6) = \sum_{i=2}^{6} P(X = i) = \sum_{i=2}^{6} \binom{i-1}{1} p^2 (1-p)^{i-2}$$

Definition: Hypergometric Let $w, b, m \in \mathbb{N}$. A random variable X follows an hypergeometric distribution if:

$$\mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}, \quad x = \max(0, m-b), \dots, \min(w, m)$$

We write $X \sim \text{HyperGeom}(w, b; m)$.

Intuition We draw a sample of m balls without replacement from an urn containing w white balls and b black balls. X models the number of white balls drawn.
Indeed, we consider the number of times we can pick x balls from our white balls and $m - x$ from our black balls, and we just divide by the number of ways to pick m balls from $w + b$ balls.

Example We have six tins of food without labelling, of which we know two contain fruits. We wonder what is the distribution of the number of tins of fruit we have when we pick three tins.

This is just an hypergeometric distribution:

$$\mathbb{P}(X = x) = \frac{\binom{2}{x}\binom{4}{3-x}}{\binom{6}{3}}, \quad x = 0, \dots, 2$$

We thus get that:

$$\mathbb{P}(X = 0) = \frac{1}{5}, \quad \mathbb{P}(X = 1) = \frac{3}{5}, \quad \mathbb{P}(X = 2) = \frac{1}{5}$$

Definition: Discrete uniform

Let $a, b \in \mathbb{Z}$ such that $a < b$. A **discrete uniform** random variable X has PMF:

$$f_X(x) = \frac{1}{b - a + 1}, \quad x = a, \dots, b$$

We write $U \sim \mathrm{DU}(a, b)$.

Intuition This definition generalises the outcome of a die throw, which would correspond to the $\mathrm{DU}(1, 6)$ distribution.

Definition: Poisson

Let $\lambda > 0$. A **Poisson** random variable X has the PMF:

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots$$

We write $X \sim \mathrm{Pois}(\lambda)$.

Proof We see that:

$$\sum_{x=0}^{\infty} f_X(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

It is also rather trivial to see that $p_X(x) > 0$, showing that this is indeed a PMF.

□

Intuition This distribution appears in many places in statistics. For instance, we can show that, supposing people arrive at random times, it models the length of a queue.

Tuesday 14th March 2023 — Lecture 7 : High expectations

Definition: CDF

Let X be some random variable with probability distribution $\mathbb{P}$. Its **Cumulative Distribution Function** (CDF) is defined as:

$$F_X(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

If moreover X is discrete, then we can write:

$$F_X(x) = \sum_{\substack{x_i \in D_x \\ x_i \leq x}} \mathbb{P}(X = x_i)$$

When there is no risk of confusion, we write $F = F_X$.

Theorem: Properties of CDF

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \mapsto \mathbb{R}$ be a random variable. Its cumulative distribution function F_X satisfies:

1. $\lim_{x \to -\infty} F_X(x) = 0$.
2. $\lim_{x \to \infty} F_X(x) = 1$
3. F_X is non-decreasing: $F_X(x) \leq F_X(y)$ for $x \leq y$.
4. F_X is continuous to the right: $\lim_{t \to x^+} F_X(x + t) = F_X(x)$
5. $\mathbb{P}(X > x) = 1 - F_X(x)$

6. If $x < y$, then $\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x)$

Observation We can obtain the mass function of a discrete random variable from the CDF by using:

$$f(x) = \lim_{z \to x^+} F(z) - \lim_{y \to x^-} F(y) = F(x) - \lim_{y \to x^-} F(y)$$

This means that describing a problem using a CDF or a PMF are completely equivalent in the discrete case.

Theorem: PMF Transformation If X is a random variable and $Y = g(X)$, then:

$$f_Y(y) = \sum_{x \mid g(x)=y} f_X(x)$$

Proof This is a direct proof:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \mid g(x)=y} \mathbb{P}(X = x) = \sum_{x \mid g(x)=y} f_X(x)$$

□

Example Let $X \sim \text{Pois}(\lambda)$ and $Y = I(X \geq 1)$. We have:

$$f_Y(0) = \mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = e^{-\lambda}$$

$$f_Y(1) = \mathbb{P}(Y = 1) = \sum_{x=1}^{\infty} \mathbb{P}(X = x) = \sum_{x=1}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1 - e^{-\lambda}$$

4.2 Expectation

Definition: Expectation Let X be a discrete random variable for which $\sum_{x \in D_X} |x| f_X(x) < \infty$, where D_X is the support of f_X. The **expectation** (or **expected value** or **mean**) of X is:

$$\mathbb{E}(X) = \sum_{x \in D_X} x\mathbb{P}(X = x) = \sum_{x \in D_x} x f_X(x)$$

Example Let us compute the expected value of some Bernoulli random variable I:

$$\mathbb{E}(I) = 0(1 - p) + 1(p) = p$$

Theorem: Expected value of a function Let X be a random variable with mass function f, and let g be a real-valued function of X. Then:

$$\mathbb{E}[g(X)] = \sum_{x \in D_x} g(x) f(x)$$

when $\sum_{x \in D_X} |g(x)| f(x) < \infty$.

Proof We let $Y = g(X)$. By a theorem we saw previously, we have:

$$f_Y(y) = \sum_{x \in D_x \mid g(x)=y} f_X(x)$$

Therefore:

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y\in D_y} y f_Y(y) \\ &= \sum_{y\in D_y} y \sum_{x\in D_x \mid g(x)=y} f_X(x) \\ &= \sum_{y\in D_y} \sum_{x\in D_x \mid g(x)=y} g(x) f_X(x) \\ &= \sum_{x\in D_x} g(x) f_X(x)\end{aligned}$$

Example

Let $X \sim \text{Pois}(\lambda)$. We want to calculate $\mathbb{E}(X(X-1))$. We can use our theorem:

$$\mathbb{E}(X(X-1)) = \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}e^{-\lambda} = \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}e^{-\lambda} = \lambda^2 \cdot 1 = \lambda^2$$

since we recognised the PMF of the Poisson distribution, which, as all PMF, sums to 1.

Thursday 16th March 2023 — **Lecture 8 : Le poisson du moment**

Theorem: Expectancy properties

Let X be a random variable with a finite expected value $\mathbb{E}(X)$, and let $a, b \in \mathbb{R}$ be constants. Then:

1. $\mathbb{E}$ is a linear operator: $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$
2. If $g(X)$ and $h(X)$ have finite expected values, then:
$$\mathbb{E}[g(X) + h(X)] = \mathbb{E}[g(X)] + \mathbb{E}[h(X)]$$
3. If $\mathbb{P}(X = b) = 1$, then $\mathbb{E}(X) = b$.
4. If $\mathbb{P}(a < X \leq b) = 1$, then $a < \mathbb{E}(X) \leq b$.
5. $\mathbb{E}(X)^2 \leq \mathbb{E}(X^2)$

Remark The three first properties are very useful.

Proof We will only prove the last property, the other are rather easy. Let us consider the following expression, which we know is positive, for any $a \in \mathbb{R}$:

$$0 \leq \mathbb{E}\left[(X-a)^2\right] = \mathbb{E}\left[X^2 - 2aX + a^2\right] = \mathbb{E}(X^2) - 2a\mathbb{E}(X) + a^2$$

Now, we notice that we exactly get our property when $a = \mathbb{E}(X)$:

$$0 \leq \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \iff \mathbb{E}(X)^2 \leq \mathbb{E}(X^2)$$

□

Definition: Moments of a distribution

Let X be a random variable with PMF $f(x)$ such that $\sum_x |x|^r f(x) < \infty$ for some r. Then, we define:

- The r^{th} **moment** of X is $\mathbb{E}(X^r)$.
- The r^{th} **central moment** of X is $\mathbb{E}[(X - \mathbb{E}(X))^r]$.
- The **factorial moment** of X is $\mathbb{E}[X(X-1)\cdots(X-r+1)]$.

Observation The expected value $\mathbb{E}$ is the first moment of X. It represents the average value of X. Note that it has the same units as X.

Definition: Variance and standard deviation

Let X be a random variable with PMF $f(x)$ such that $\sum_x |x|^2 f(x) < \infty$.

Then, we define the **variance** of X to be its second central moment, i.e.:

$$\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}(X))^2\Big]$$

The **standard deviation** of X is defined as:

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

Remark The variance represents the scatter of X around its mean, the average squared distance to the mean. To make its unit match the ones of X, we take the square root, yielding the standard deviation.

Example 1 We throw a dice, and we want to compute its expected value and variance:

$$\mathbb{E}(X) = \frac{1 + \ldots + 6}{6} = \frac{7}{2}$$

$$\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}(X))^2\Big] = \sum_{x=1}^{6} \frac{1}{6}\left(x - \frac{7}{2}\right)^2 = \frac{25 + 9 + 1 + 1 + 9 + 25}{6 \cdot 4} = \frac{35}{12}$$

Example 2 We want to compute the factorial moment of the Poisson distribution:

$$\begin{aligned}\mathbb{E}[X(X-1)\cdots(X-r+1)] &= \sum_{x=0}^{\infty} x(x-1)\cdots(x-r+1)\frac{\lambda^x}{x!}e^{-\lambda} \\ &= \lambda^r \sum_{x-r=0}^{\infty} \frac{\lambda^{x-r}}{(x-r)!}e^{-\lambda} \\ &= \lambda^r\end{aligned}$$

This for instance gives us that $\mathbb{E}(X) = \lambda$.

Theorem: Properties of variance Let X be a random variable which variance exists, and let a, b be constants. Then:

1. $\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2 = \mathbb{E}[X(X-1)] + \mathbb{E}(X) - \mathbb{E}(X)^2$
2. $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$
3. $\mathrm{Var}(X) = 0$ implies that X is constant with probability 1.

Intuition The second property is important. It means that shifting the distribution by a constant b does not have any impact on its average distance from the mean, which should make sense. Also, it means that if we expand it by a factor a, then the average squared distance to the mean is multiplied by a factor a^2.
The third property just means that if the average squared distance to the mean is 0, then the random variable must always give the mean.

Example We have a random variable in $\{1, \ldots, 6\}$. We want to find the PMF which would maximise the variance.
A good first guess would be with uniform probability distribution $f(x) = \frac{1}{6}$ for all x. We get:

$$\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2 = \frac{1}{6}\sum_{i=1}^{6} i^2 - \left(\sum_{i=1}^{6} \frac{i}{6}\right)^2 = \frac{35}{12}$$

Now, let us consider another random variable (which we will still call X for simplicity), for which push all the mass to the extremities, giving $f(1) = f(6) = \frac{1}{2}$ and $f(x) = 0$ otherwise. Its variance is:

$$\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2 = \frac{1^2 + 6^2}{2} - \left(\frac{1+6}{2}\right)^2 = \frac{1}{4}(1-6)^2 = \frac{25}{4}$$

This second PMF has the highest variance and, in fact, we cannot do better. To maximise the average squared distance from the mean, we need to put it further

from the mean. The first PMF instead maximises entropy (which is the amount of randomness a random variable has).

Theorem If X takes its values in $\{0, 1, \ldots\}$ and $\mathbb{E}(X) < \infty$, then:

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)$$

We can generalise this result. Letting $r \geq 2$, we get:

$$\mathbb{E}[X(X-1)\cdots(X-r+1)] = r\sum_{x=r}^{\infty}(x-1)\cdots(x-r+1)\mathbb{P}(X \geq x)$$

Note that this only works with random variables taking their values in non-negative integers.

Proof 1 We see that:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=1}^{\infty} xf(x) \\
&= \sum_{x=1}^{\infty} \mathbb{P}(X = x) \sum_{r=1}^{x} 1 \\
&= \sum_{x=1}^{\infty} \sum_{r=1}^{x} \mathbb{P}(X = x) \\
&= \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)
\end{aligned}$$

Proof 2 Let us consider the following expression:

$$r(x-1)\cdots(x-r+1) = r!\frac{(x-1)!}{(r-1)!(x-r)!} = r!\binom{x-1}{r-1}$$

Thus, we can write:

$$\begin{aligned}
r\sum_{x=r}^{\infty}(x-1)\cdots(x-r+1)\mathbb{P}(X \geq x) &= \sum_{x=r}^{\infty} r!\binom{x-1}{r-1}\sum_{y=x}^{\infty} f_X(y) \\
&= \sum_{y=r}^{\infty} f_X(y) r! \sum_{x=r}^{y}\binom{x-1}{r-1}
\end{aligned}$$

However, we can see that:

$$\begin{aligned}
r!\sum_{x=r}^{y}\binom{x-1}{r-1} &= r\sum_{x=r}^{y}\left[\binom{x}{r} - \binom{x-1}{r}\right] \\
&= r!\binom{y}{r} \\
&= y(y-1)\cdots(y-r+1)
\end{aligned}$$

This gives us that:

$$\begin{aligned}
& r\sum_{x=r}^{\infty}(x-1)\cdots(x-r+1)\mathbb{P}(X \geq x) \\
&= \sum_{y=r}^{\infty} f_X(y)y(y-1)\cdots(y-r+1) \\
&= \mathbb{E}[X(X-1)\cdots(X-r+1)]
\end{aligned}$$

as required.

□

Tuesday 21st March 2023 — **Lecture 9 : Going continuous**

Example

We want to compute the expected value and variance of $X \sim \text{Geom}(p)$.
We know that $\mathbb{P}(X \geq x) = (1-p)^{x-1}$. By our theorem, we get that:

$$\mathbb{E}(X) = \sum_{x=1}^{\infty}(1-p)^{x-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Now, let's compute the variance. Let's first compute the following value:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= 2\sum_{x=2}^{\infty}(x-1)(1-p)^{x-1} \\ &= 2(1-p)\frac{d}{dp}\left[-\sum_{x=1}^{\infty}(1-p)^{x-1}\right] \\ &= 2(1-p)\frac{d}{dp}\left(\frac{-1}{p}\right) \\ &= \frac{2(1-p)}{p^2}\end{aligned}$$

This trick of seeing that a series is the n^{th} derivative of a sum we know how to compute (typically for power series) is very common and must be remembered.
We finally get that the variance is:

$$\text{Var}(X) = \mathbb{E}[X(X-1)] + \mathbb{E}(X) - \mathbb{E}(X)^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

As a quick sanity check, we can notice that it gets smaller as $p \to 1$ (there is less variance when we have more probability to success) and larger as $p \to 0$ (there is a lot of variance when we have a small probability of success). This is what intuition tells us, so this is a good thing.

4.3 Conditional probability distributions

Definition: Conditional probability distribution

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, on which we define a random variable X, and let $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$.
Then, the **conditional probability mass function** of X given B is:

$$f_X(x|B) = \mathbb{P}(X = x|B) = \frac{\mathbb{P}(A_x \cap B)}{\mathbb{P}(B)}$$

Theorem

The function $f_X(x|B)$ is a well defined mass function. In other words:

$$f_X(x|B) \geq 0, \quad \sum_x f_X(x|B) = 1$$

Proposition

Let B be an event of the form $X \in \mathcal{B}$, for some $\mathcal{B} \subset \mathbb{R}$. Then:

$$f_X(x|B) = \frac{I(x \in \mathcal{B})}{\mathbb{P}(X \in \mathcal{B})} f_X(x)$$

Intuition

This means that:

$$f_X(x|B) = \begin{cases} 0, & x \notin \mathcal{B} \\ c f_X(x), & x \in \mathcal{B} \end{cases}$$

where $c = \frac{1}{\mathbb{P}(X \in \mathcal{B})}$ is a normalisation factor.

We thus zero-out every terms which cannot happen, and normalise the rest.

Proof This is a direct proof:

$$\begin{aligned} f_X(x|B) &= \frac{\mathbb{P}(X = x \cap X \in \mathcal{B})}{\mathbb{P}(X \in \mathcal{B})} \\ &= \frac{\mathbb{P}(X \in \mathcal{B}|X = x)}{\mathbb{P}(X \in \mathcal{B})}\mathbb{P}(X = x) \\ &= \frac{I(x \in \mathcal{B})}{\mathbb{P}(X \in \mathcal{B})} f_X(x) \end{aligned}$$

Indeed, $\mathbb{P}(X \in \mathcal{B}|X = x)$ is one if $x \in \mathcal{B}$ and 0 otherwise. It is thus an indicator variable.

□

Example

We want to compute the conditional PMF of $X \sim \text{Geom}(p)$ given that $X > n$. We know that the event $B = \{X > n\}$ has probability $(1-p)^n$ (the first success is after the n^{th} trial is equivalent to saying that the first n experiments failures). Thus:

$$f_X(x|B) = \frac{\mathbb{P}(X = x \cap X > n)}{\mathbb{P}(X > n)} = \frac{I(x > n)}{\mathbb{P}(X > n)} f_X(x) = p(1-p)^{x-n-1}$$

for $x = n+1, n+2, \ldots$. Note that the indicator variable hides in the fact that we only consider $x \geq n+1$.

Definition: Conditional expected value

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where we consider some random variable X and event $B \in \mathcal{F}$. We also suppose that $\sum_x |g(x)| f_X(x|B) < \infty$.
The **conditional expected value** of $g(X)$ given B is defined as:

$$\mathbb{E}[g(X)|B] = \sum_x g(x) f_X(x|B)$$

Theorem

Let X be random variable with expected value $\mathbb{E}(X)$, and $\{B_i\}_{i=1}^{\infty}$ be partition such that $\mathbb{P}(B_i) > 0$ for all i and such that the following sum is absolutely convergent. Then:

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} \mathbb{E}(X|B_i)\mathbb{P}(B_i)$$

This is very similar to the theorem of total probability.

Implication Let B be an event with $\mathbb{P}(B) > 0$ and $\mathbb{P}(B^C) > 0$. Then, our theorem implies that:

$$\mathbb{E}(X) = \mathbb{E}(X|B)\mathbb{P}(B) + \mathbb{E}(X|B^c)\mathbb{P}(B^c)$$

This follows immediately from the theorem, by setting $B_1 = B$, $B_2 = B^C$ and $B_3 = B_4 = \ldots = \emptyset$.

Proof Just as one can think, we will use the theorem of total probability:

$$f(x) = \mathbb{P}(X = x) = \sum_{i=1}^{\infty} \mathbb{P}(X = x|B_i)\mathbb{P}(B_i) = \sum_{i=1}^{\infty} f(x|B_i)\mathbb{P}(B_i)$$

So, this gives that:

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x f(x) \\ &= \sum_x x \sum_{i=1}^{\infty} f(x|B_i)\mathbb{P}(B_i) \\ &= \sum_{i=1}^{\infty} \left[\sum_x x f(x|B_i)\right]\mathbb{P}(B_i) \\ &= \sum_{i=1}^{\infty} \mathbb{E}(X|B_i)\mathbb{P}(B_i)\end{aligned}$$

□

Example 1

We want to compute the expected value of $X \sim \text{Geom}(p)$ given that $X > n$.
We have seen in the previous example that:

$$f_X(x|B) = p(1-p)^{x-n-1}, \quad x = n+1, \dots$$

This gives us that:

$$\mathbb{E}(X|B) = \sum_{x=n+1}^{\infty} x f_X(x|B) = \sum_{x=n+1}^{\infty} xp(1-p)^{x-n-1}$$

Setting $y = x - n$, this gives:

$$\mathbb{E}(X|B) = \sum_{y=1}^{\infty}(n+y)p(1-p)^{y-1} = n\sum_{y=1}^{\infty} p(1-p)^{y-1} + \sum_{y=1}^{\infty} yp(1-p)^{y-1} = n + \frac{1}{p}$$

Indeed, the first sum is the probability distribution of the geometric distributions (and thus sums to 1), and the second one is their expected value. This result is rather intuitive.

Example 2

Let's now say that we want to compute the conditional probability of $X \sim \text{Geom}(p)$ given that $B^C = \{X \leq n\}$.
However, we know $\mathbb{E}(X)$ and $\mathbb{E}(X|B)$: the second one was just computed in the previous example. Thus, we know that:

$$\begin{aligned}&\mathbb{E}(X) = \mathbb{E}(X|B)\mathbb{P}(B) + \mathbb{E}\big(X|B^C\big)\mathbb{P}\big(B^C\big) \\ \iff &\mathbb{E}\big(X|B^C\big) = \frac{\mathbb{E}(X) - \mathbb{E}(X|B)\mathbb{P}(B)}{\mathbb{P}(B^C)} = \frac{\frac{1}{p} - \left(n + \frac{1}{p}\right)(1-p)^n}{1-(1-p)^n}\end{aligned}$$

4.4 Law of small numbers

Definition: Convergence of distributions

Let $\{X_n\}$ and X be random variables with cumulative distribution functions $\{F_n\}$ and F, respectively.
We say that $\{X_n\}$ **converge in distribution** (or converge in laws) to X if, for all $x \in \mathbb{R}$ where F is continuous:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

We write $X_n \xrightarrow{D} X$ (or $X_n \xrightarrow{(d)} X$) as $n \to \infty$.

Remark In the case of discrete random variables, meaning $D_X \subset \mathbb{Z}$, this is equivalent to:

$$\lim_{n\to\infty} f_n(x) = f(x)$$

Theorem: Law of small numbers Let (p_n) be a sequence of probabilities such that $\lim_{n\to\infty} np_n = \lambda > 0$, and let $X_n \sim B(n, p_n)$ be a sequence of random variables following binomial distributions. Finally, let $X \sim \text{Pois}(\lambda)$.
Then:

$$X_n \xrightarrow{\text{D}} X$$

Implication We can use this theorem to approximate binomial probabilities for large n and small p, by using Poisson probabilities.

Proof Let r be fixed. We want to show that the PMF of the binomial distribution tends towards the one of the Poisson distribution:

$$\begin{aligned}\lim_{n\to\infty} \binom{n}{r} p_n^r (1-p_n)^{n-r} &= \lim_{n\to\infty} \underbrace{n^{-r}\binom{n}{r}}_{\to \frac{1}{r!}} \underbrace{(np_n)^r}_{\lambda^r} \underbrace{\left(1 - \frac{np_n}{n}\right)^{n-r}}_{\to e^{-\lambda}} \\ &= \frac{1}{r!}\lambda^r e^{-\lambda}\end{aligned}$$

We use the fact that:

$$\lim_{n\to\infty} n^{-r}\binom{n}{r} = \frac{1}{r!}$$

□

Proposition Let X_n be a hypergeometric variable:

$$\mathbb{P}(X_n = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{x}}$$

Also, let $N, m \to \infty$ such that $\frac{m}{N} \to p$ for some $0 < p < 1$.
Then, the limiting distribution of X_N is $B(n, p)$:

$$\mathbb{P}(X_N = x) \to \binom{n}{x} p^x (1-p)^{n-x}, \quad i = 0, \ldots, n$$

Proof Multiplying by some terms which divide out to 1, it gives:

$$\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} = \frac{m^x}{N^x} \cdot \frac{(N-m)^{n-x}}{N^{n-x}} \cdot \frac{\overbrace{m^{-x}\binom{m}{x}}^{\to \frac{1}{x!}} \overbrace{(N-m)^{-(n-x)}\binom{N-m}{n-x}}^{\frac{1}{(n-x)!}}}{\underbrace{N^{-n}\binom{N}{n}}_{\to \frac{1}{n!}}}$$

which, as $N \to +\infty$, tends towards:

$$p^x(1-p)^{n-x}\frac{n!}{x!(n-x)!} = p^x(1-p)^{n-x}\binom{n}{x}$$

as required.

□

Chapter 5

Continuous random variables

5.1 Fundamentals

Definition: Continuous random variable

Let X be a random variable, and D_x be its support:

$$D_X = \{x \in \mathbb{R} \mid \exists \omega \in \Omega \text{ such that } X(\omega) = x\}$$

This is a **continuous random variable** if D_X is infinite non-countable.

Definition: CDF

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
The **cumulative distribution function** of a continuous random variable X on this probability space is defined just as for discrete random variables:

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\mathcal{B}_x), \quad x \in \mathbb{R}$$

where $\mathcal{B}_x = \{\omega \mid X(\omega) \leq x\} \subset \Omega$.

Observation

The probability that a continuous random variable is an exact value is 0. For instance, the probability that we get exactly 0.5 when picking a random number in $[0, 1]$ is 0.
We need to switch our point of view: we will consider masses. We know that the mass of any particle of some continuous matter weigh 0, but that the whole object still weighs something. When we consider point-wise functions, we must not consider mass, but densities. This thus leads to the following definition.

Definition: PDF

A random variable X is continuous if there exists a function $f(x)$, called the **probability density function** (or density), named PDF, of X, such that:

$$\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du, \quad x \in \mathbb{R}$$

By using the properties of F, we know that this implies that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

Remark

By using the fundamental theorem of calculus, we get that:

$$f(x) = \frac{dF(x)}{dx}$$

This means that $F(x)$ needs to be continuous. This function $F(x)$ can only have jumps for discrete random variables.

Observation

We notice that:

$$\mathbb{P}(X = x) = \lim_{\delta \to 0} \int_{x}^{x+\delta} f(x)dx = 0$$

As expected, a single point has zero probability.

Thursday 23rd March 2023 — **Lecture 10 : Transformations**

Remark

We notice that it is possible to have $f(x) > 1$ for some x, as long as $\int_{-\infty}^{\infty} f(x)dx = 1$. It might just be spiky.
The important thing to consider is that f is not a probability, but a density: it is the area under its curve which represents probabilities.

Definition: Uniform distribution

Let $a < b$. A **uniform random variable** U is a random variable with density:

$$f(u) = \begin{cases} \frac{1}{b-a}, & a \leq u \leq b, \\ 0, & \text{otherwise} \end{cases}$$

We write $U \sim U(a, b)$.

CDF

We can compute the CDF of this function:

$$F(u) = \int_{-\infty}^{u} f(x)du = \begin{cases} 0, & u \leq a \\ \frac{u-a}{b-a}, & a < u \leq b \\ 1, & u > b \end{cases}$$

Definition: Exponential distribution

Let $\lambda > 0$. An **exponential random variable** is a random variable X with density:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

We write $X \sim \exp(\lambda)$.

CDF

We can integrate this function to get its CDF:

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \exp(-\lambda x), & x > 0 \end{cases}$$

Theorem: Lack of memory

An exponential distribution has the lack of memory property:

$$\mathbb{P}(X > x + t | X > t) = \mathbb{P}(X > x), \quad t, x > 0$$

Proof

We can use the CDF of the exponential distribution:

$$\begin{aligned} \mathbb{P}(X > x + t | X > t) &= \frac{\mathbb{P}(X > x + t)}{\mathbb{P}(X > t)} \\ &= \frac{\exp(-\lambda(x + t))}{\exp(-\lambda t)} \\ &= \exp(-\lambda x) \\ &= \mathbb{P}(X > x) \end{aligned}$$

□

Definition: Gamma function

The **Gamma function**, $\Gamma : \mathbb{R}_+ \mapsto \mathbb{R}$ is defined as:

$$\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} e^{-u} du, \quad \alpha > 0$$

Properties This is a generalisation of the factorial to all positive (and in fact, complex) numbers since it follows the following recurrence relation:

$$\Gamma(1) = 1, \quad \Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \quad \alpha > 0$$

This thus yields that:

$$\Gamma(n) = (n-1)!, \quad n \in \mathbb{N}^*$$

We for instance have that:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Remark This function is not important to remember.

Definition: Gamma distribution Let $\alpha, \lambda > 0$. A **gamma random variable** X has density:

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \end{cases}$$

Definition: Laplace distribution Let $\varepsilon \in \mathbb{R}, \lambda > 0$. A **Laplace random variable** (or double exponential) X has density:

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\varepsilon|}, \quad x \in \mathbb{R}$$

CDF We have the following CDF:

$$F(x) = \begin{cases} \dfrac{1}{2} e^{-\lambda|x-\varepsilon|}, & x \leq \varepsilon \\ 1 - \dfrac{1}{2} e^{-\lambda|x-\varepsilon|}, & x > \varepsilon \end{cases}$$

Definition: Pareto distribution Let $\alpha, \beta > 0$. A **Pareto random variable** has the following distribution:

$$f(x) = \begin{cases} 0, & x < \beta \\ \dfrac{\alpha\beta^\alpha}{x^{\alpha+1}}, & x \geq \beta \end{cases}$$

CDF It has the following CDF:

$$F(x) = \begin{cases} 0, & x < \beta \\ 1 - \left(\dfrac{\beta}{x}\right)^\alpha, & x \geq \beta \end{cases}$$

Definition: Expectation Let $g(x)$ be a real-valued function, and X a continuous random variable of density $f(x)$. Then, if $\mathbb{E}(|g(x)|) < \infty$, we define the **expectation** of $g(X)$ to be:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Intuition This is almost the same as for the discrete case, except that we replace sums by integrals.

Expectation and variance In particular, the **expectation** of X is:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Definition: Variance Let X be some random variable such that the following integral exists. Then, its **variance** is:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 f(x)dx = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2$$

Example We want to compute the expectation and the variance of $X \sim U(a, b)$.
Let us first compute:

$$\mathbb{E}(X^r) = \int_{-\infty}^{\infty} x^r \frac{1}{b-a} dx = \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)}$$

This gives us the expectation:

$$\mathbb{E}(X) = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

which is the middle point, very intuitively.
We can then compute the variance:

$$\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2 = \frac{(b-a)^2}{12}$$

Definition: Conditional densities Let $\mathcal{A} \subset \mathbb{R}$ be a reasonable subset, and:

$$\mathcal{A}_x = \{y \mid y \leq x \text{ and } y \in \mathcal{A}\}$$

Then, the conditional PDF is defined as:

$$f_X(x|X \in \mathcal{A}) = \begin{cases} \dfrac{f_X(x)}{\mathbb{P}(X \in \mathcal{A})}, & x \in \mathcal{A} \\ 0, & \text{otherwise} \end{cases}$$

The CDF is given by:

$$F_X(x|X \in \mathcal{A}) = \mathbb{P}(X \leq x|X \in \mathcal{A}) = \frac{\mathbb{P}(X \leq x \cap X \in \mathcal{A})}{\mathbb{P}(X \in \mathcal{A})} = \frac{1}{\mathbb{P}(X \in A)} \int_{A_x} f(y)dy$$

Finally, we define the expectancy:

$$\mathbb{E}[g(X)|X \in \mathcal{A}] = \frac{\mathbb{E}[g(X)I(X \in \mathcal{A})]}{\mathbb{P}(X \in \mathcal{A})}$$

where I is an indicator variable.

Definition: Quantile Let $0 < p < 1$. We define the p quantile of the cumulative distribution function $F(x)$ to be:

$$x_p = \inf\{x \mid F(x) \geq p\}$$

The infimum is required when the function is discontinuous, or when it is flat for some values.

> *Intuition* For most continuous random variables, x_p is unique and is given by $x_p = F^{-1}(p)$.

Definition: Median The **median** of some cumulative distribution function is the $\frac{1}{2}$ quantile.

> *Intuition* It represents the point where we have 50 % chance to be above, and 50 % chance to be below.

Example Let $X \sim \exp(\lambda)$. We want to find the p quantile of X.

We have to solve $F(x_p) = p$:

$$1 - \exp(-\lambda x_p) = p \iff x_p = -\lambda^{-1}\log(1-p)$$

5.2 Transformations

Example

Let $W = -\log(U)$, where $U \sim U(0,1)$. We want to find $F_W(w)$.
We notice that, for some $w > 0$:

$$\begin{aligned}\mathbb{P}(W \leq w) &= \mathbb{P}[-\log(U) \leq w] \\ &= \mathbb{P}(U \geq \exp(-w)) \\ &= 1 - \mathbb{P}\big(U < e^{-w}\big) \\ &= 1 - \mathbb{P}\big(U \leq e^{-w}\big) \\ &= 1 - e^{-w}\end{aligned}$$

which is our result.
Note that we have $\mathbb{P}(U < e^{-w}) = \mathbb{P}(U \leq e^{-w})$ since the probability of any single point is 0.

Definition: Inverse function

Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a function, and $\mathcal{B} \subset \mathbb{R}$ any subset of $\mathbb{R}$.
We write $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ to be the set for which $g\big[g^{-1}(\mathcal{B})\big] = \mathcal{B}$

Remark

We have access to some $U \sim U(0,1)$. Given any arbitrary CDF F_*, we want to find a transformation $g : \mathbb{R} \mapsto \mathbb{R}$ such that, letting $Y = g(U)$:

$$F_Y(y) = F_*(y)$$

In other words, we want:

$$\mathbb{P}(g(U) \leq x) = F_*(x)$$

To do so, we can see that $g = F_*^{-1}$ works:

$$\mathbb{P}(g(U) \leq x) = \mathbb{P}\big(F_*^{-1}(U) \leq x\big) = \mathbb{P}(U \leq F_*(x)) = F_*(X)$$

This result is very general. It means that if we have access to some uniform random generator and we want to generate numbers following any arbitrary PDF $f_*(x)$, we only need to integrate the PDF to get the CDF $F_*(x)$, inverse it, and pass any sample u into this $F_*^{-1}(x)$.

> *Personal note: Usage*
>
> Such transformations can be very important. For instance, in computer science, we have pseudo-random number generators which give some uniform distributions. We may then want to map those to another distribution for different use cases.
> For example, in Computer Graphics, we need to uniformly sample points on a hemisphere. A good idea is to sample points using polar coordinates. However, using this technique blindly with a uniform distribution does not yield a uniform distribution on the sphere (there will be more points near the poles). We thus need to map our uniform points to some other distributions.

Tuesday 28th March 2023 — **Lecture 11 : The Professor did not know the error function ☹**

Theorem: General transformation

Let $Y = g(X)$ be a random variable, and $\mathcal{B}_y =]-\infty, y]$. We consider $g^{-1}(\mathcal{B}_y) = \{x \in \mathbb{R} \mid g(x) \leq y\}$.
If X is continuous, then:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B}_y)\big) = \int_{g^{-1}(\mathcal{B}_y)} f_X(x)dx$$

If X is discrete:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B}_y)\big) = \sum_{x \in g^{-1}(\mathcal{B}_y)} f_X(x)$$

If moreover g is monotone and has a differentiable inverse, then:

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X\big(g^{-1}(y)\big)$$

Proof 1

We show the continuous case.
We know that $X \in g^{-1}(\mathcal{B})$ if and only if $g(X) \in g\big(g^{-1}(\mathcal{B})\big) = \mathcal{B}$. Thus:

$$\mathbb{P}(Y \in \mathcal{B}) = \mathbb{P}(g(X) \in \mathcal{B}) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B})\big)$$

Now, to find $F_y(y)$, we only need to take $\mathcal{B}_y =]-\infty, y]$, giving:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \in \mathcal{B}_y) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B}_y)\big)$$

as expected.

Proof 2

We show the part for the PDF.
We consider g to be increasing, the case where it is decreasing is similar (two minuses cancel out). This implies that $g^{-1}(]-\infty, y]) = \big]-\infty, g^{-1}(y)\big]$, and thus, for $y \in \mathbb{R}$

$$F_Y(y) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B}_y)\big) = \mathbb{P}\big(X \leq g^{-1}(y)\big) = F_X\big(g^{-1}(y)\big)$$

If, moreover X is continuous, we can differentiate:

$$f_Y(y) = \frac{dg^{-1}(y)}{dy} f_X\big(g^{-1}(y)\big)$$

Doing the case for the decreasing function, we get that, completely generally:

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X\big(g^{-1}(y)\big)$$

Remark

To compute $\frac{dg^{-1}(y)}{dy}$, we can notice that:

$$\begin{aligned} &\big(g \circ g^{-1}\big)(y) = y \\ \implies &g'\big(g^{-1}(y)\big)\frac{dg^{-1}(y)}{dy} = 1 \\ \implies &\frac{dg^{-1}(y)}{dy} = \frac{1}{g'(g^{-1}(y))} \end{aligned}$$

Observation

We notice that, if $X \sim U(0,1)$ is uniform, and $g : [0,1] \mapsto \mathbb{R}$ is increasing (and thus $0 \leq g^{-1} \leq 1$ is also increasing), then:

$$F_Y(y) = F_X\big(g^{-1}(x)\big) = g^{-1}(x)$$

since the CDF of some uniform random variable $U(0,1)$ is $F(x) = x$, for $0 \leq x \leq 1$.
This gives us exactly the result we derived before, since a CDF is always increasing.

Example 1

Let $X \sim \exp(\lambda)$ and $Y = \exp(X)$. We want to find F_y and f_y.
We note that $g(x) = e^x$ is increasing, and $g^{-1}(y) = \log(y)$. Thus, for $y > 1$:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(Y \in \mathcal{B}) = \mathbb{P}(g(X) \in \mathcal{B}) = \mathbb{P}\big(X \in g^{-1}(\mathcal{B})\big)$$

This is equal to:

$$\mathbb{P}(X \in]-\infty, \log(y)]) = F_X(\log(y)) = 1 - \exp(-\lambda \log(y)) = 1 - y^{-\lambda}$$

This implies that Y has a Pareto distribution, with $\beta = 1$ and $\alpha = \lambda$. Finally, for $y > 1$:

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X\big(g^{-1}(y)\big) = \left|\frac{1}{y}\right| \lambda e^{-\lambda \log(y)} = \lambda y^{-\lambda-1}$$

Moreover, $f_Y(y) = 0$ for $y \leq 1$ because, then, $\log(y) < 0$ and $f_X(x) = 0$ for $x < 0$. We could also naturally just have used the formula of the Pareto distribution to get the PDF.

Example 2

Let $X \sim \exp(1)$ and $Y = \cos(X)$. We want to find the CDF of Y.
First, because the cosine only gives values in $[0, 1]$, we get that $F_Y(y) = 0$ for $y < -1$ and $F_Y(y) = 1$ for $y \geq 1$.
Moreover, we notice that, for $0 < x < 2\pi$:

$$\cos(X) \leq y \iff \arccos(y) \leq X \leq 2\pi - \arccos(y)$$

By periodicity, and since exponential random variables are always positive, we can split our case:

$$\begin{aligned} &\cos(X) \leq y \\ \iff &X \in g^{-1}(B_y) = \bigcup_{j=0}^{\infty} \{x \in \mathbb{R} \mid 2\pi j + \arccos(y) \leq x \leq 2\pi(j+1) - \arccos(y)\} \end{aligned}$$

This yields that:

$$\begin{aligned} \mathbb{P}(Y \leq y) &= \mathbb{P}\big(X \in g^{-1}(\mathcal{B})\big) \\ &= \sum_{j=0}^{\infty} \mathbb{P}(2\pi j + \arccos(y) \leq X \leq 2\pi(j+1) - \arccos(y)) \\ &= \sum_{j=0}^{\infty} [\exp[-\lambda(2\pi j + \arccos(y))] - \exp[-\lambda(2\pi(j+1) - \arccos(y))]] \\ &= \frac{\exp(-\lambda \arccos(y)) - \exp(\lambda \arccos(y) - 2\pi\lambda)}{1 - \exp(-2\pi\lambda)} \end{aligned}$$

since this is a geometric series.

5.3 Normal distribution

Definition: Normal distribution

Let $\mu \in \mathbb{R}$, $\sigma > 0$. A random variable X has **normal distribution** if it has density:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$. We will show that the standard deviation is σ and the mean is μ.
When $\mu = 0$ and $\sigma^2 = 1$, we call Z a **standard normal**, following:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}$$

Observation

We note that:

$$f(x) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}$$

Integral

The fact that the integral of the PDF is equal to 1 is based on the fact that:

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

The classic way to show this is by computing its square and using a change of variable to polar coordinates:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2} e^{-y^2} dx dy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta \\ &= [\theta]_0^{2\pi} \left[-\frac{e^{-r^2}}{2} \right]_0^{\infty} \\ &= \frac{2\pi}{2} \\ &= \pi \end{aligned}$$

which indeed yields that $I = \sqrt{\pi}$ (since $I > 0$).

Interpretation

The standard deviation σ is a measure of the spread of the values around μ. 68 % of the probability lies in the interval $\mu \pm \sigma$, 95 % in $\mu \pm 2\sigma$ and 99.7 % in $\mu \pm 3\sigma$.
This is often referred to as the 68-95-99.7 rule.

Theorem: Properties

Let $\varphi(z)$ be the PDF, $\Phi(z)$ be the CDF and z_p be the quantiles of $Z \sim \mathcal{N}(0,1)$. Then, for all $z \in \mathbb{R}$:

- The density is symmetric with respect to $z = 0$:

$$\varphi(z) = \varphi(-z)$$

- Using this symmetry of the PDF:

$$\mathbb{P}(Z \leq z) = \Phi(z) = 1 - \Phi(-z) = 1 - \mathbb{P}(Z \geq z)$$

- The standard normal quantiles satisfy:

$$z_p = -z_{1-p}, \quad 0 < p < 1$$

- For all $r > 0$:

$$\lim_{z \to \pm\infty} z^r \varphi(z) = 0$$

  In other words, the moment $\mathbb{E}(Z^r)$ exist for all $r \in \mathbb{N}$.
- We have:

$$\varphi'(z) = -z\varphi(z), \quad \varphi''(z) = \big(z^2 - 1\big)\varphi(z), \quad \varphi'''(z) = -\big(z^3 - 3z\big)\varphi(z), \quad \dots$$

  This implies that $\mathbb{E}(Z) = 0$, $\mathrm{Var}(Z) = \mathbb{E}\big(Z^2\big) = 1$, $\mathbb{E}\big(Z^3\big) = 0$, and so on.
- If $X \sim \mathcal{N}\big(\mu, \sigma^2\big)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

  This implies that, if $X \sim \mathcal{N}\big(\mu, \sigma^2\big)$, then, for $Z \sim \mathcal{N}(0,1)$:

$$X = \mu + \sigma Z$$

Values

$\Phi(z)$ is a non-elementary function. Thus, a typical way to find its value is using a table:

z	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169

The rows represent the first digit of x after the coma, and the columns its second digit. For instance:

$$\Phi(1.54) \approx 0.93822$$

Example 1

The duration of a maths lecture is $\mathcal{N}(47, 4)$, and should be 45. We want to find the probability that the lecture finishes early:

$$\mathbb{P}(X \leq 45) = \mathbb{P}\left(\frac{X-47}{\sqrt{4}} \leq \frac{45-47}{\sqrt{4}}\right) = \mathbb{P}(Z \leq -1) = \mathbb{P}(Z \geq 1) = 1 - \mathbb{P}(Z \leq 1) = 1 - \Phi(1)$$

Using a table of values, we get that:

$$\mathbb{P}(X \leq 45) = 1 - \Phi(1) = 1 - 0.84134 = 0.15866$$

Example 2

Let $Z \sim \mathcal{N}(0, 1)$. We want to compute the PDF and CDF of $Y = |Z|$ and $W = Z^2$. For Y, it is rather easy. We see that, for $y > 0$:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(|Z| \leq y) = \mathbb{P}(-y \leq Z \leq y) = \Phi(y) - \Phi(-y) \implies p_Y(y) = 2\varphi(y)$$

For W, we can use a similar argument:

$$\mathbb{P}(W \leq w) = \mathbb{P}\big(-\sqrt{w} \leq Z \leq \sqrt{w}\big) = \Phi\big(\sqrt{w}\big) - \Phi\big(-\sqrt{w}\big)$$

Theorem: De Moivre-Laplace theorem

Let $0 < p < 1$, $X_n \sim B(n, p)$, $Z \sim \mathcal{N}(0, 1)$, and:

$$\mu_n = \mathbb{E}(X_n) = np, \quad \sigma_n^2 = \mathrm{Var}(X_n) = np(1-p)$$

Then:

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{D} Z$$

Remark This means by definition that:

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{X_n - \mu_n}{\sigma_n} \leq z\right) = \Phi(z)$$

Implication

This allows to give us an approximation of the probability that $X_n \leq r$:

$$\mathbb{P}(X_n \leq r) = \mathbb{P}\left(\frac{X_n - \mu_n}{\sigma_n} \leq \frac{r - \sigma_n}{\sigma_n}\right) = \Phi\left(\frac{r - \mu_n}{\sigma_n}\right)$$

In other words, $X_n \sim \mathcal{N}(np, np(1-p))$.
In practice, the approximation is bad when $\min\{np, n(1-p)\} < 5$.
However, this is just a rule of thumb.

Chapter 6

Multi-dimensional random variables

6.1 Fundamentals

Definition: Discrete joint probability functions

Let (X, Y) be a discrete random variable, meaning that the set of points (x, y) such that $\mathbb{P}[(X, Y) = (x, y)] > 0$ is countable.
The **(joint) probability mass function** of (X, Y) is:

$$f_{X,Y}(x, y) = \mathbb{P}[(X, Y) = (x, y)], \quad (x, y) \in \mathbb{R}^2$$

The **(joint) cumulative distribution function** of (X, Y) is:

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y), \quad (x, y) \in \mathbb{R}^2$$

Definition: Continuous joint probability functions

The random variables (X, Y) is said to be **(jointly) continuous** if there exists a function $f_{X,Y}(x, y)$, called the **(joint) density** of (X, Y), such that:

$$\mathbb{P}[(X, Y) \in A] = \iint_{(u,v)\in A} f_{X,Y}(u, v)dudv, \quad A \subset \mathbb{R}^2$$

By letting $A = \{(u, v) \mid u \leq x, v \leq y\}$, we get the **(joint) cumulative distribution function** of (X, Y):

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)dudv, \quad (x, y) \in \mathbb{R}^2$$

Which implies that:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

For the functions we will study in this course, the order of differentiation does not matter.

Thursday 30th March 2023 — **Lecture 12 : Generalisation of our definitions**

Theorem: Marginal PMF and PDF

The **marginal probability mass function** of a discrete random variable is:

$$f_X(x) = \mathbb{P}(X = x) = \sum_y f_{X,Y}(x, y), \quad x \in \mathbb{R}$$

The **marginal probability density function** of a continuous random variable is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Proof We consider the discrete case, the continuous case is similar.
We are looking for $f_X(x) = \mathbb{P}(X = x)$:

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_y \mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \\ &= \sum_y \mathbb{P}(X = x, Y = y) \\ &= \sum_y f_{X,Y}(x, y)\end{aligned}$$

□

Definition: Conditional PMF and PDF The **conditional probability mass/density function** of Y given X is, supposing that $f_X(x) > 0$:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad y \in \mathbb{R}$$

If (X, Y) is discrete, we also have that:

$$f_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x)$$

Example Let's say every day we receive $X \sim P(100)$ emails (following a Poisson distribution with $\lambda = 100$). Each is a spam independently with probability $p = 0.9$. We want to find the distribution of the total number of emails we received, given that we received 15 good ones.
Let Y be the number of good emails we received. This follows:

$$Y|X = x \sim B(x, 1 - p) \implies \mathbb{P}(Y = y|X = x) = \binom{x}{y}(1 - p)^y p^{x-y}$$

Now, we know that:

$$f_Y(y) = \sum_x f_{X,Y}(x, y) = \sum_x f_{Y|X}(y|x) f_X(x) = \sum_x \binom{x}{y}(1 - p)^y p^{x-y} \frac{\lambda^x}{x!} e^{-\lambda}$$

We can show that this is a Poisson mass function with parameter $\lambda(1 - p)$:

$$f_Y(y) = \frac{[\lambda(1 - p)^y]}{y!} e^{-\lambda(1-p)}$$

Definition: Joint CDF Let $X_1, \ldots, X_n$ be random variables defined on the same probability space. Their **joint cumulative distribution function** is:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

Definition: Joint PMF/PDF Let $X_1, \ldots, X_n$ be random variables defined on the same probability space.
If they are continuous, their **joint probability density function** is:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

where the order of derivatives does not mater with the functions we will consider.
If they are discrete, their **joint probability mass function** is:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

Definition: Marginal Let $\mathcal{A} \subset \{1, 2, \ldots, n\}$. We consider a n-tuple of random variables $X_{\{1,\ldots,n\}} = (X_1, \ldots, X_n)$, writing for instance $X_{\{2,3,5\}} = (X_2, X_3, X_5)$.

If our random variables are continuous, the marginal distribution of $X_{\mathcal{A}}$ is:

$$f_{X_{\mathcal{A}}}(x) = \int_{\overline{\mathcal{A}}} f_{X_{\{1,\dots,n\}}}(x_1,\dots,x_n)dx_{\mathcal{A}}$$

If our random variables are discrete, the marginal distribution of $X_{\mathcal{A}}$ is:

$$f_{X_{\mathcal{A}}}(x) = \sum_{x_{\mathcal{A}}\in\overline{\mathcal{A}}} f_{X_{\{1,\dots,n\}}}(x_1,\dots,x_n)$$

Definition: Conditional density

Let $\mathcal{A},\mathcal{B}\subset\{1,2,\dots,n\}$ be such that $\mathcal{A}\cap\mathcal{B}=\emptyset$. The conditional probability mass/density function of some set of random variables $Y_{\mathcal{B}}$ given another set of random variables $X_{\mathcal{A}}$ is, supposing that $f_{X_{\mathcal{A}}}(x)>0$:

$$f_{Y_{\mathcal{B}}|X_{\mathcal{A}}}(y|x) = \frac{f_{X_{\mathcal{A}},Y_{\mathcal{B}}}(x,y)}{f_{X_{\mathcal{A}}}(x)},\quad y\in\mathbb{R}$$

Example

Let's say that we have $F_{X,Y}(x,y)$ and that we want to find $F_X(x)$. To do so, we notice that:

$$F_X(x) = \mathbb{P}(X\leq x) = \mathbb{P}(X\leq x, Y<+\infty) = \lim_{y\to\infty} F_{X,Y}(x,y)$$

Definition: Multinomial distribution

Let $m\in\mathbb{N}$ and $p_1,\dots,p_k\in[0,1]$ such that $p_1+\dots+p_k=1$.
The random variable $(X_1,\dots,X_k)$ has the **multinomial distribution** of denominator m and probabilities $(p_1,\dots,p_k)$ if its mass function is:

$$f(x_1,\dots,x_k) = \frac{m!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}\quad x_1,\dots,x_k\in\{0,\dots,m\} \text{ and } \sum_{j=1}^{k} x_j = m$$

This is written $(X_1,\dots,X_k)\sim B(m,p_1,\dots,p_k)$

Intuition This is the probability that, having m balls, we have x_1 balls of the first colour, x_2 balls of the second, and so on; knowing that the first colour has probability p_1 to appear, and so on.

Example

Let us say that n people vote for three candidates, independently with probabilities $p_1=0.45$, $p_2=0.4$ and $p_3=0.15$ (where p_i represents the probability to vote for the i^{th} candidate).
Let X_1,X_2,X_3 represent the number of votes for each candidate. This follows exactly a multinomial distribution:

$$\mathbb{P}(X_1=x_1,X_2=x_2,X_3=x_3) = \frac{n!}{x_1!x_2!x_3!}0.45^{x_1}0.4^{x_2}0.15^{x_3}$$

Tuesday 4th April 2023 — **Lecture 13 : Correlation and some causation**

Definition: Independence

Two random variables X,Y defined on the same probability space are **independent** if, for any set $\mathcal{A}$ and $\mathcal{B}$:

$$\mathbb{P}(X\in\mathcal{A},Y\in\mathcal{B}) = \mathbb{P}(X\in\mathcal{A})\mathbb{P}(Y\in\mathcal{B})$$

Implication By letting $\mathcal{A}=]-\infty,x]$ and $\mathcal{B}=]-\infty,y[$, this implies that:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y),\quad \forall x,y\in\mathbb{R}$$

This also implies that:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),\quad \forall x,y\in\mathbb{R}$$

Moreover:

$$f_{Y|X}(y|x) = f_Y(y),\quad \forall y\in\mathbb{R}$$

Equivalence

In fact, we can show that X and Y are independent if and only if:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad \forall x,y \in \mathbb{R}$$

This is also equivalent to:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \forall x,y \in \mathbb{R}$$

for both continuous and discrete random variables.

Example

Let us consider the following probability density function:

$$f_{X,Y}(x,y) = \begin{cases} ce^{-3x-2y}, & x,y > 0 \\ 0, & \text{otherwise} \end{cases}$$

where c allows for normalisation.
We notice that we can write:

$$f_{X,Y}(x,y) = ce^{-3x-2y}I(x>0)I(y>0) = c\big(e^{-3x}I(x>0)\big)\big(e^{-2y}I(y>0)\big)$$

Since we got that $f_{X,Y}(x,y) = g(x)h(y)$ for some functions g and h, we can deduce the independence.
Note that, if we had $x \leq y$ instead of $x,y>0$ as the condition for our PDF, then we could not have done the same proof and the two variables would, in fact, not be independent.

Definition: IID

A **random sample** of size n from a distribution F (or density f) is a set of n independent random variables which all have a distribution F (or density f).
Equivalently, we say that $X_1,\dots,X_n$ are **independent and identically distributed** (iid) with distribution F (or density f) and write $X_1,\dots,X_n \overset{\text{iid}}{\sim} F$ (or $X_1,\dots,X_n \overset{\text{iid}}{\sim} f$).

PDF

By independence the joint density of $X_1,\dots,X_n \overset{\text{iid}}{\sim} f$ is:

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{j=1}^{n} f(x_j)$$

Example

Let us consider $X_1,X_2,X_3 \overset{\text{iid}}{\sim} \exp(\lambda)$, for some $\lambda > 0$. Then:

$$f(x_1,x_2,x_3) = f(x_1)f(x_2)f(x_3) = \lambda^3 \exp[-\lambda(x_1+x_2+x_3)], \quad x_1,x_2,x_3 > 0$$

6.2 Dependence

Definition: Expectation

Let X,Y be random variables of density or mass $f_{X,Y}(x,y)$.
If X,Y are discrete and the following sum converges for $|g(X,Y)|$, then the **expectation** of $g(X,Y)$ is:

$$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)f_{X,Y}(x,y)$$

If X,Y are continuous and the following integral converges for $|g(X,Y)|$, then the **expectation** of $g(X,Y)$ is:

$$\mathbb{E}[g(X,Y)] = \iint g(x,y)f_{X,Y}(x,y)dxdy$$

Definition: Joint and central moments

Let X, Y be random variables of density or mass $f_{X,Y}(x, y)$ and $r, s \in \mathbb{N}$ such that $\mathbb{E}(|X^r Y^s|) < \infty$.
The **joint moments** of X and Y are:

$$\mathbb{E}(X^r Y^s)$$

The **joint central moments** of X and Y are:

$$\mathbb{E}[(X - \mathbb{E}(X))^r (Y - \mathbb{E}(Y))^s]$$

Remark Those definitions are not really important, they mostly allow to define the following (very important) concept.

Definition: Covariance

Let X, Y be random variables of density or mass $f_{X,Y}(x, y)$ such that $\mathbb{E}(|g(X, Y)|) < \infty$. Their covariance is:

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Remark It is possible to show that:

$$\operatorname{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Theorem: Covariance properties

Let X, Y, Z be random variables, and $a, b, c, d \in \mathbb{R}$ be constants. The covariance has the following properties:

- $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$
- $\operatorname{Cov}(a, X) = 0$
- Symmetry: $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
- Bilinearity: $\operatorname{Cov}(a + bX + cY, Z) = b \operatorname{Cov}(X, Z) + c \operatorname{Cov}(Y, Z)$
- $\operatorname{Cov}(a + bX, c + dY) = bd \operatorname{Cov}(X, Y)$
- $\operatorname{Var}(a + bX + cY) = b^2 \operatorname{Var}(X) + 2bc \operatorname{Cov}(X, Y) + c^2 \operatorname{Var}(Y)$
- Cauchy-Schwarz inequality: $\operatorname{Cov}(X, Y)^2 \leq \operatorname{Var}(X) \operatorname{Var}(Y)$

Digression: Cauchy-Scharz inequality Cauchy-Schwarz inequality says that, on some Hilbert space (a vector space with a dot product), then:

$$|\langle \vec{u}, \vec{v} \rangle| \leq \|\vec{u}\| \|\vec{v}\| \iff |\langle \vec{u}, \vec{v} \rangle|^2 \leq \|\vec{u}\|^2 \|\vec{v}\|^2$$

This means that we can consider random variables as a vector space, with the dot product being the covariance, and the norm being the standard deviation (meaning the squared norm is the variance).

Proof We want to show the following property, since it is easier to remember its proof than its result. It directly comes from bilinearity:

$$\begin{aligned} & \operatorname{Var}(a + bX + cY) \\ = \ & \operatorname{Cov}(a + bX + cY, a + bX + cY) \\ = \ & b^2 \operatorname{Cov}(X, X) + 2bc \operatorname{Cov}(X, Y) + c^2 \operatorname{Cov}(Y, Y) \\ = \ & b^2 \operatorname{Var}(X) + 2bc \operatorname{Cov}(X, Y) + c^2 \operatorname{Var}(Y) \end{aligned}$$

□

Theorem: Expectation

If X and Y are independent, and $g(X)$ and $h(Y)$ are functions which expectations exist, then we have:

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

Proof Let us consider the continuous case:

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \iint g(x)h(y)f_{XY}(x,y)dxdy \\ &= \iint g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int g(x)f_X(x)dx \int h(y)f_Y(y)dy \\ &= \mathbb{E}[g(X)]\mathbb{E}[h(Y)]\end{aligned}$$

□

Corollary: Covariance If X and Y are independent, then:

$$\mathrm{Cov}(X,Y) = 0$$

Remark The converse is wrong. It is possible to find random variables such that $\mathrm{Cov}(X,Y) = 0$ but X and Y are not independent.

Theorem: Linearity Let $X_1, \ldots, X_n$ be random variables and $a, b_1, \ldots, b_n$ be constants. Then:

$$\mathbb{E}(a + b_1X_1 + \ldots + b_nX_n) = a + \sum_{j=1}^{n} b_j\mathbb{E}(X_j)$$

$$\mathrm{Var}(a + b_1X_1 + \ldots + b_nX_n) = \sum_{j=1}^{n} b_j^2 \mathrm{Var}(X_j) + \sum_{j\neq k} b_jb_k \mathrm{Cov}(X_j, X_k)$$

If moreover $X_1, \ldots, X_n$ are independent, then $\mathrm{Cov}(X_j, X_k) = 0$ for $j \neq k$ and thus:

$$\mathrm{Var}(a + b_1X_1 + \ldots + b_nX_n) = \sum_{j=1}^{n} b_j^2 \mathrm{Var}(X_j)$$

Remark Those properties can easily be found back.
For instance, let us compute $\mathrm{Var}(16 + 5X_1 - 6X_2)$. First, we know that constants do not matter, so we can write this as:

$$\begin{aligned}&\mathrm{Var}(5X_1 - 6X_2) \\ =\ &\mathrm{Cov}(5X_1 - 6X_2, 5X_1 - 6X_2) \\ =\ &25\,\mathrm{Cov}(X_1, X_1) - 60\,\mathrm{Cov}(X_1, X_2) + 36\,\mathrm{Cov}(X_2, X_2) \\ =\ &25\,\mathrm{Var}(X_1) - 60\,\mathrm{Cov}(X_1, X_2) + 36\,\mathrm{Var}(X_2)\end{aligned}$$

Definition: Average The **average** of random variables $X_1, \ldots, X_n$ is:

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

Theorem: Mean Let $X_1, \ldots, X_n$ be *independent* random variables which all have mean μ and variance σ^2. Then:

$$\mathbb{E}\big(\overline{X}\big) = \mu, \quad \mathrm{Var}\big(\overline{X}\big) = \frac{\sigma^2}{n}$$

Definition: Correlation The **correlation** of X, Y is defined as:

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

This measures the linear dependence between X and Y.

Personal note: Intuition

We have seen that covariance was some kind of dot product, and variance was some kind of square norm. Thus, it is of the form:

$$\frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\|\|\vec{y}\|} = \frac{\|\vec{x}\|\|\vec{y}\| \cos(\vec{x}, \vec{y})}{\|\vec{x}\|\|\vec{y}\|} = \cos(\vec{x}, \vec{y})$$

Thus, this indeed allows to know how colinear the two are. Considering two vectors in 2D, they are correlated if the angle between them is small. This is the same for random variables, except that it is more abstract.

Remark

Two independent random variables are necessarily uncorrelated.
However, two uncorellated random variables are not necessarily independent. For instance, let us consider some random variables $\Theta \sim U(0, 2\pi), X = \cos(\Theta), Y = \sin(\Theta)$. X and Y are definitely dependent (if we know Y, then X only has two possibilities), however we can show that they are uncorrelated.
Correlation only sees linear dependence; and those random variable are dependent, but not linearly.

Example

Let us consider $Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and:

$$X = Z_1, \quad Y_1 = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$$

for $0 < \rho < 1$.
We see that:

$$\operatorname{Var}(X) = \operatorname{Var}(Z_1) = 1$$

Also, since Z_1 and Z_2 are independent:

$$\begin{aligned} \operatorname{Var}(Y) &= \operatorname{Var}\Big(\rho Z_1 + \sqrt{1 - \rho^2} Z_2\Big) \\ &= \operatorname{Var}(\rho Z_1) + \operatorname{Var}\Big(\sqrt{1 - \rho^2} Z_2\Big) \\ &= \rho^2 \operatorname{Var}(Z_1) + \big(1 - \rho^2\big) \operatorname{Var}(Z_2) \\ &= \rho^2 + \big(1 - \rho^2\big) \\ &= 1 \end{aligned}$$

Moreover:

$$\begin{aligned} \operatorname{Cov}(X, Y) &= \operatorname{Cov}\Big(Z_1, \rho Z_1 + \sqrt{1 - \rho^2} Z_2\Big) \\ &= \rho \operatorname{Cov}(Z_1, Z_1) + \sqrt{1 - \rho^2} \operatorname{Cov}(Z_1, Z_2) \\ &= \rho \cdot 1 + \sqrt{1 - \rho^2} \cdot 0 \\ &= \rho \end{aligned}$$

We finally get that:

$$\operatorname{corr}(X, Y) = \rho$$

Theorem: Properties of correlation

Let X, Y be random variables with correlation $\rho = \operatorname{corr}(X, Y)$. Then:

- $-1 \leq \rho \leq 1$. This directly comes from Cauchy-Schwarz inequality.
- If $\rho = \pm 1$, then there exist $a, b, c \in \mathbb{R}$ such that $(a, b) \neq (0, 0)$ and:

  $$aX + bY + c = 0$$

  In other words, X and Y are totally linearly dependent.
- If X, Y are independent, then $\operatorname{corr}(X, Y) = 0$.
- Correlation is not affected by scaling. In other words, for $a, b, c, d \in \mathbb{R}$:

  $$\operatorname{corr}(a + bX, c + dY) = \operatorname{sign}(bd) \operatorname{corr}(X, Y)$$

Remark

Note that correlation does not imply causation.
For instance, we may notice that the number of churches in cities is highly correlated with the number of criminals. However, churches do not imply criminals; the correct explanation is that a city with more inhabitants has more churches and more criminals. The only way to deduce something in statistics is to act on a value, and see if it changes the other.
Also, correlation only measures linear dependence. If we have some kind of quadratic dependence, then the correlation might be 0 (even though we could still say the data is correlated).

Definition: conditional Expectation

Let X, Y be random variables, $x \in \mathbb{R}$ be such that $f_X(x) > 0$, and $g(X,Y)$ be a function such that the following sum and integral converge for $|g(X,Y)|$.
If X, Y are discrete then the **conditional expectation** of $g(X,Y)$ given $X = x$ is:

$$\mathbb{E}[g(X,Y)|X=x] = \sum_y g(x,y) f_{Y|X}(y|x)$$

If X, Y are continuous then the **conditional expectation** of $g(X,Y)$ given $X = x$ is:

$$\mathbb{E}[g(X,Y)|X=x] = \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x) dy$$

Remark Note that $\mathbb{E}[g(X,Y)|X=x]$ is a function of x.

Example

Let $Z = XY$ where X, Y are independent, X has a Bernoulli distribution and Y has a Poisson distribution with parameter λ.
We have that:

$$\mathbb{E}(Z|X=x) = \mathbb{E}(XY|X=x) = \mathbb{E}(xY|X=x) = x\mathbb{E}(Y|X=x) = x\mathbb{E}(Y) = x\lambda$$

since x is a constant, and since X and Y are independent. The fact that we can replace X by x follows from the definition.
Thus:

$$\mathbb{E}(Z|X=0) = 0, \quad \mathbb{E}(Z|X=1) = \lambda$$

Thursday 6th April 2023 — **Lecture 14 : Linearity of expectation**

Theorem

Let X, Y be random variables. If the required expectations exist, then:

$$\mathbb{E}[g(X,Y)] = \mathbb{E}_X[\mathbb{E}[g(X,Y)|X=x]]$$

where $\mathbb{E}_X$ represent expectation according to the distribution of X.
This means that we can freeze a random variable (X here), and then average this result over this variable.

Proof For the discrete case, we have that:

$$\begin{aligned}
\mathbb{E}_X[\mathbb{E}[g(X,Y)|X=x]] &= \sum_x \mathbb{E}[g(X,Y)|X=x] f_X(x) \\
&= \sum_x \sum_y g(x,y) f_{Y|X}(y|x) f_X(x) \\
&= \sum_x \sum_y g(x,y) f_{X,Y}(x,y) \\
&= \mathbb{E}[g(X,Y)]
\end{aligned}$$

Properties

Let X and Y be random variables. The following equality *always* (even if they are dependent) holds:

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

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

If moreover X and Y are independent:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Proof 1 We want to show the first equality.
By definition of $\mathbb{E}[g(x,y)]$, we need $f_{X,Y}(x,y)$ to compute $\mathbb{E}(X+Y)$.
Thus, let us pick $Z = g(X,Y) = X+Y$. By definition:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \mathbb{E}(Z) \\
&= \sum_z z\mathbb{P}(g(X,Y)=z) \\
&= \sum_z z\mathbb{P}(X+Y=z) \\
&= \sum_z z \sum_x \mathbb{P}(X+Y=z|X=x) f_X(x) \\
&= \sum_x \left[\sum_z z\mathbb{P}(X+Y=z|X=x)\right] f_X(x)
\end{aligned}$$

We consider the inner sum, by using the change of variable $u = z-x$:

$$\begin{aligned}
\sum_z z\mathbb{P}(X+Y=z|X=x) &= \sum_z z\mathbb{P}(Y=z-x|X=x) \\
&= \sum_u (u+x) f_{Y|X}(u|x) \\
&= \sum_u u f_{Y|X}(u|x) + x\sum_u f_{Y|X}(u|x) \\
&= \mathbb{E}(Y|X=x) + x
\end{aligned}$$

since u is just a dummy variable.
Coming back to our sum, we find

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_x \left[\sum_z z\mathbb{P}(X+Y=z|X=x)\right] f_X(x) \\
&= \sum_x [\mathbb{E}(Y|X=x) + x] f_X(x) \\
&= \mathbb{E}(Y) + \sum_x x f_X(x) \\
&= \mathbb{E}(Y) + \mathbb{E}(X)
\end{aligned}$$

Proof 2 We want to show the first equality in a different way.
We know that:

$$\mathbb{E}(g(X,Y)) = \sum_{x,y} g(x,y) f_{X,Y}(x,y)$$

Thus:

$$\begin{aligned}\mathbb{E}(X+Y) &= \sum_{x,y}(x+y)f_{X,Y}(x,y)\\ &= \sum_x x\sum_y f_{X,Y}(x,y) + \sum_y y\sum_x f_{X,Y}(x,y)\\ &= \sum_x x f_X(x) + \sum_y y f_Y(y)\\ &= \mathbb{E}(X)+\mathbb{E}(Y)\end{aligned}$$

This proof is simpler.

□

Tuesday 18th April 2023 — **Lecture 15 : The calm before the storm**

6.3 Moment generating functions

Definition: Moment-generating function

The **moment-generating function** (MGF) of a random variable X (also known as the **Laplace transform** of $f_X(x)$) is:

$$M_X(t) = \mathbb{E}\big(e^{tX}\big)$$

for all t such that $M_X(t)$ is defined.

Property

By using properties of the expected value, we see that:

$$M_X(t) = \mathbb{E}\big(e^{tX}\big) = \mathbb{E}\left(\sum_{r=0}^{\infty}\frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty}\frac{t^r}{r!}\mathbb{E}(X^r)$$

Thus, we can get all the moments by differentiation:

$$\mathbb{E}(X^r) = M_X^{(r)}(0)$$

In particular, we always have:

$$M_X(0) = 1, \quad M_x'(0) = \mathbb{E}(X)$$

Example 1

Let us consider X to be an indicator variable with probability p. We get that its MGF is:

$$M_X(t) = (1-p)e^{t\cdot 0} + pe^{t\cdot 1} = 1 - p + pe^t$$

We notice that, indeed:

$$M_X(0) = 1$$

$$M'(t) = pe^t \implies \mathbb{E}(X) = M'(0) = p$$

Example 2

Let $X \sim B(n,p)$. We want to find its MGF:

$$M_X(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{x}\big(pe^t\big)^x(1-p)^{n-x} = \big(1-p+pe^t\big)^n$$

for any $t \in \mathbb{R}$.

Example 3

Let $X \sim \text{Pois}(\lambda)$, we want to find its MGF:

$$M_X(t) = \sum_{x=0}^{\infty} e^{xt}\frac{\lambda^x}{x!}e^{-\lambda} = \sum_{x=0}^{\infty}\frac{\big(\lambda e^t\big)^x}{x!}e^{-\lambda} = \exp\big(\lambda e^t\big)e^{-\lambda} = \exp\big(\lambda\big(e^t-1\big)\big), \quad t \in \mathbb{R}$$

Example 4

Let us consider $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z \sim \mathcal{N}(0,1)$. To compute $M_X(t)$, we first need $M_Z(t)$:

$$\mathbb{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz}\varphi(z)dz = \int_{-\infty}^{\infty} e^{tz}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

To compute this integral, we use the fact that normal PDFs integrate to 1:

$$1 = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \frac{e^{\frac{-\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2\sigma^2}+\frac{2x\mu}{2\sigma^2}}dx$$
$$\iff \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}+x\frac{\mu}{\sigma^2}}dx = \sigma e^{\frac{\mu^2}{2\sigma^2}}$$

Leaving $\sigma = 1$ and $\mu = t$, this allows us to get:

$$\mathbb{E}(e^{tZ}) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}+zt}dz = e^{\frac{t^2}{2}}$$

So, we get that:

$$\mathbb{E}(e^{tX}) = \mathbb{E}\left[e^{t(\mu+\sigma Z)}\right] = e^{t\mu}\mathbb{E}\left[e^{t\sigma Z}\right] = e^{t\mu+\frac{t^2\sigma^2}{2}}$$

Theorem: Properties

If $M_X(t)$ is the MGF of a random variable X, then we have:

- $M_X(0) = 1$
- $M_{a+bX}(t) = e^{at}M_X(bt)$
- $\mathbb{E}(X^r) = M_X^{(r)}(0)$
- $\mathbb{E}(X) = M_X'(0)$
- $\mathrm{Var}(X) = M_X''(0) - M_X'(0)^2$

Theorem: Linear combinations

Let $a, b_1, \ldots, b_n \in \mathbb{R}$ and $X_1, \ldots, X_n$ be independent random variables which MGF exist. Then $Y = a + b_1X_1 + \ldots + b_nX_n$ has the following MGF:

$$M_Y(t) = e^{ta}\prod_{j=1}^{n} M_{X_j}(tb_j)$$

If moreover they are identically distributed (meaning that, overall, they are IID), $S = X_1 + \ldots + X_n$ has the following MGF:

$$M_S(t) = M_X(t)^n$$

Remark This is really what makes MGFs appealing: they allow to turn sums into products when we have independent random variables.

Example

We want to compute the expectation and the variance of $X \sim \exp(\lambda)$. It seems easier to use a MGF:

$$M_X(t) = \int_0^{\infty} e^{xt}\lambda e^{-\lambda x}dx = \lambda\int_0^{\infty} e^{-(\lambda-t)x}dx = \frac{\lambda}{\lambda - t}$$

Which we can differentiate to get:

$$M_X'(t) = \frac{\lambda}{(\lambda-t)^2}, \quad M_X''(t) = \frac{2\lambda}{(\lambda-t)^3}$$

This indeed gives us:

$$\mathbb{E}(X) = M_X'(0) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = M_X''(0) - M_X'(0)^2 = \frac{1}{\lambda^2}$$

Theorem

There exists an injection between CDFs and moment-generating functions.

Intuition In other words, if we recognise the moment-generating function, then we can deduce the probability distribution.
In fact, for all the functions we will see in this class, there is a bijection between CDFs and MGFs.

Example Let $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ be independent random variables. We want to find the distribution of $X_1 + X_2$.
We notice that:

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = \exp\big((\lambda_1+\lambda_2)\big(e^t-1\big)\big)$$

We recognise the MGF of a Poisson variable with parameter $\lambda_1+\lambda_2$. By our theorem, this indeed implies that $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Continuity theorem Let $\{X_n\}, X$ be random variables with cumulative distribution functions $\{F_n\}, F$ and which MGFs $M_n(t), M(t)$ exist at all $|t| < b$, for some $b > 0$ (i.e. they exist in a neighbourhood of $t = 0$).
If there exists a $a \in \mathbb{R}$ such that $0 < a < b$ and $\lim_{n\to\infty} M_n(t) = M(t)$ for all $|t| \le a$, then $X_n \xrightarrow{\text{D}} X$.

Intuition In other words, if $\lim_{n\to\infty} M_n(t) = M(t)$ in a neighbourhood of $t = 0$, then $F_n(x) \to F(x)$ at each $x \in \mathbb{R}$ where F is continuous.

Remark This theorem is sometimes called Lévy's continuity theorem.

Example Let $X_n \sim B(n, p_n)$ and $X \sim \text{Pois}(\lambda)$ such that $\lim_{n\to\infty} np_n = \lambda$. We want to prove the law of small numbers, meaning that $X_n \xrightarrow{\text{D}} X$, in another way.
For a binomial random variable, we know that:

$$M_{X_n}(t) = \big(1 - p_n + p_n e^t\big)^n, \quad \forall t \in \mathbb{R}$$

Moreover, for a Poisson random variable, we have seen that:

$$M_X(t) = \exp\big(\lambda\big(e^t - 1\big)\big), \quad \forall t \in \mathbb{R}$$

So, using that $\lim_{n\to\infty}\big(1 + \frac{a}{n}\big)^n = e^a$, we can see that, for all $t \in \mathbb{R}$:

$$\begin{aligned}\lim_{n\to\infty} M_{X_n}(t) &= \lim_{n\to\infty}\big(1 - p_n + p_n e^t\big)^n \\ &= \left(1 + \frac{np_n(e^t-1)}{n}\right)^n \\ &= \exp\big(\lambda\big(e^t-1\big)\big) \\ &= M_X(t)\end{aligned}$$

We have thus shown our theorem, more easily that what we had done.

Tuesday 25th April 2023 — **Lecture 16 : The storm**

Definition: Mean vector Let $X = (X_1, \ldots, Xp)^T \in \mathbb{R}^p$ be a vector of random variables. Its **expectation** (or **mean vector**) is:

$$\mathbb{E}(X)_{p\times 1} = \begin{pmatrix}\mathbb{E}(X_1)\\ \vdots \\ \mathbb{E}(X_p)\end{pmatrix}$$

Definition: PSD matrix Let Ω be a matrix. It is **positive semi-definite** (PSD), denoted $\Omega \succeq 0$, if it is symmetric and, for all vector $x \in \mathbb{R}^n$:

$$x^T\Omega x \ge 0$$

Implications Since it is symmetric, it is orthonormally diagonalisable with real eigenvalues by the spectral theorem. In other words, there exists a diagonal matrix D and matrix U for which $UU^T = I_p$, such that:

$$\Omega = UDU^T$$

Moreover, since $x^T \Omega x \geq 0$ for any vector x, it implies that all its eigenvalues are positive.

Definition: positive definite matrix Let Ω be a matrix. It is **positive definite**, denoted $\Omega \succ 0$, if it is symmetric and, for all vector $x \in \mathbb{R}^n \setminus \{0\}$:

$$x^T \Omega x > 0$$

Implications A positive definite matrix has all the properties of a PSD matrix, except that all its eigenvalues are strictly positive, and thus it is full rank (meaning that it is inversible).

Definition: (co)-variance matrix Let $X = (X_1, \dots, X_p)^T$ be a vector of random variables. Its **(co)-variance matrix** is:

$$\operatorname{Var}(X)_{p\times p} = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \operatorname{Cov}(X_1, X_2) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_1, X_p) & \operatorname{Cov}(X_2, X_p) & \cdots & \operatorname{Var}(X_p) \end{pmatrix}$$

Remark We notice that, for any vector $a = (a_1, \dots, a_p)^T \in \mathbb{R}^p$, we have:

$$\begin{aligned} \operatorname{Var}\left(\sum_{j=1}^p a_j X_j\right) &= \sum_{i=1}^p \sum_{j=1}^p \operatorname{Cov}(a_i X_i, a_j X_j) \\ &= \sum_{i=1}^p \sum_{j=1}^p a_i a_j \operatorname{Cov}(X_i, X_j) \\ &= a^T \operatorname{Var}(X) a \end{aligned}$$

However this matrix is symmetric (since $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$). We can combine this with the following fact to get this matrix is positive semi-definite:

$$a^T \operatorname{Var}(X) a = \operatorname{Var}\left(\sum_{j=1}^p a_j X_j\right) \geq 0$$

Definition: MGF of random vector Let $X_{p\times 1} = (X_1, \dots, X_p) \in \mathbb{R}^p$ be a random vector. Its **moment-generating function** (MGF) is:

$$M_X(t) = \mathbb{E}\left(e^{t \bullet X}\right) = \mathbb{E}\left(e^{t^T X}\right) = \mathbb{E}\left(e^{\sum_{r=1}^p t_r X_r}\right), \quad \forall t \in \mathcal{T}$$

where $\mathcal{T}$ is the set of $t \in \mathbb{R}^p$ where this converges, meaning:

$$\mathcal{T} = \{t \in \mathbb{R}^p \mid M_X(t) < \infty\}$$

Proposition: Properties of MGF Let $X_{p\times 1}$ be a random vector. Its MGF has the following properties:

- $0 \in \mathcal{T} \subset \mathbb{R}^p$ and $M_X(0) = 1$.
- The mean vector of $X_{p\times 1}$ is:

$$\mathbb{E}(X)_{p\times 1} = M'_X(0) = \frac{\partial M_X}{\partial t}(0) = \begin{pmatrix} \frac{\partial M_x}{\partial t_1}(0) \\ \vdots \\ \frac{\partial M_x}{\partial t_p}(0) \end{pmatrix} \in \mathbb{R}^p$$

- The variance matrix of $X_{p\times 1}$ is:

$$\mathrm{Var}(X)_{p\times p} = \frac{\partial^2 M_X}{\partial t \partial t^T}(0) - M_X'(0)M_X'(0)^T$$

where $\frac{\partial^2 M_X}{\partial t \partial t^T}$ is the Hessian matrix of M_X:

$$\frac{\partial^2 M_X}{\partial t \partial t^T}(t) = \begin{pmatrix} \frac{\partial^2 M_X}{\partial t_1^2}(t) & \cdots & \frac{\partial^2 M_X}{\partial t_1 \partial t_p}(t) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 M_X}{\partial t_p \partial t_1}(t) & \cdots & \frac{\partial^2 M_X}{\partial t_p^2}(t) \end{pmatrix}$$

- Let $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$ be such that they are disjoint (meaning $\mathcal{A} \cap \mathcal{B} = \emptyset$). As usual, we denote $X_{\mathcal{A}}$ for the subvector of X containing $\{X_j \mid j \in \mathcal{A}\}$ (for instance $X_{\{1,5\}} = (X_1, X_5)^T$). $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are independent if and only if:

$$M_X(t) = \mathbb{E}\big(e^{t_{\mathcal{A}} \bullet X_{\mathcal{A}} + t_{\mathcal{B}} \bullet X_{\mathcal{B}}}\big) = M_{X_{\mathcal{A}}}(t_{\mathcal{A}}) M_{X_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T}$$

- There is an injective mapping from MGFs to probability distributions. In other words, different probability distributions which MGFs exist lead to different MGFs.

Definition: Characteristic function

The **characteristic function** of some random variable X is:

$$\varphi_X(t) = \mathbb{E}\big(e^{itX}\big), \quad t \in \mathbb{R}$$

This is also known as a **Fourier transform**.

Remark Many distributions do not have a MGF, their Laplace transform does not converge. Thus, using a Fourier transform can be very useful for proofs, even though they are a bit trickier since they require complex analysis.

6.4 Multivariate normal distribution

Definition: Multivariate normal distribution

Let some random vector $X = (X_1, \dots, X_p)^T$. It follows a **multivariate normal distribution** (or it is jointly Gaussian) if there exists a $p \times 1$ vector $\mu = (\mu_1, \dots, \mu_p)^T \in \mathbb{R}^p$ and a $p \times p$ symmetric matrix with elements ω_{jk} such that:

$$u^T X \sim \mathcal{N}\big(u^T \mu, u^T \Omega u\big), \quad u \in \mathbb{R}^p$$

We write $X \sim \mathcal{N}_p(\mu, \Omega)$.

Remark If they exist, then we have:

$$\mu = \mathbb{E}(X), \quad \Omega = \mathrm{Var}(X)$$

Observation Since $0 \leq \mathrm{Var}\big(u^T X\big) = u^T \Omega u$ for any $\mathbb{R}^p$, this means that Ω must be positive semi-definite.

Terminology Saying that a random vector follows a normal distribution is ambiguous. It may mean that each component follows a normal distribution, or that the vector follows a multivariate normal distribution.

Intuition The idea is that it is constructed such that, no matter the line on which we project our variables, it is a 1D Gaussian distribution.

Example: Degenerate case

Let us consider:

$$X_1 \sim \mathcal{N}(0, 1), \quad X_2 = -X_1$$

We notice that:

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_1, -X_1) = -\mathrm{Var}(X_1) = -1$$

This implies that:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\right)$$

However, in this case, if we consider $u = (1,1)^T \in \mathbb{R}^2$:

$$u^T X = X_1 + X_2 \sim \mathcal{N}(0,0)$$

This is a degenerate case, which is slightly problematic since we divide by the standard deviation in the PDF of the normal distribution. However, since this only happens when we have a linear dependence between X_1 and X_2, this can be fixed by not considering X_2 when doing our distribution computations, and then using $X_2 = -X_1$.
In other words, if we have a degenerate case, we can decrease the dimensions of our matrix (to $\mathbb{R}^{1\times 1}$ in this case) in order for it to match its rank, and thus make it positive definite. When Ω is positive definite (meaning that all its eigenvalues are strictly positive and thus that it is full rank), everything works perfectly fine.

Proposition: Mean and covariance of joint normal distribution

Let $X \sim N_p(\mu, \Omega)$, with $\omega_{ij} = \Omega_{ij}$.
Then:

$$\mathbb{E}(X_j) = \mu_j, \quad \mathrm{Var}(X_j) = \omega_{jj}, \quad \mathrm{Cov}(X_j, X_k) = \omega_{jk}$$

Proof

Let $e_j \in \mathbb{R}^p$ be the vector with a 1 at the j^{th} place and zeros everywhere else. We then have that:

$$X_j = e_j^T X \sim \mathcal{N}(\mu_j, \omega_{jj})$$

It follows that:

$$\mathbb{E}(X_j) = \mu_j, \quad \mathrm{Var}(X_j) = \omega_{jj}$$

For the final property, we can see that:

$$X_j + X_k = (e_j + e_k)^T X \sim \mathcal{N}(\mu_j + \mu_k, \omega_{jj} + \omega_{kk} + 2\omega_{jk})$$

This gives us:

$$\begin{aligned} & \mathrm{Var}(X_j + X_k) = \omega_{jj} + \omega_{kk} + 2\omega_{jk} \\ \iff & \mathrm{Var}(X_j) + \mathrm{Var}(X_k) + 2\,\mathrm{Cov}(X_j, X_k) = \omega_{jj} + \omega_{kk} + 2\omega_{jk} \\ \iff & \mathrm{Cov}(X_j, X_k) = \omega_{jk} \end{aligned}$$

□

Theorem: Properties of the joint normal distribution

Let $X \sim \mathcal{N}_p(\mu, \Omega)$. We have the following properties:

1. The moment-generating function of X is:

$$M_X(t) = \exp\left(t^T\mu + \frac{1}{2}t^T\Omega t\right), \quad t \in \mathbb{R}^p$$

2. Let $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$ be such that $\mathcal{A} \cap \mathcal{B} = \emptyset$. $X_\mathcal{A}$ and $X_\mathcal{B}$ are independent ($X_\mathcal{A} \perp\!\!\!\perp X_\mathcal{B}$) if and only if:

$$\Omega_{\mathcal{A}\times\mathcal{B}} = 0$$

In other words, $X_1, \dots, X_n$ are mutually independent if and only if Ω is diagonal.

3. If moreover $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then:

$$X \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$$

where $1_n \in \mathbb{R}^n$ is a vector with only ones, and $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix.

4. Any linear combination of our variables is also normal. In other words, letting $a \in \mathbb{R}^r$ and $B \in \mathbb{R}^{r \times p}$, we get:

$$a + BX \sim \mathcal{N}_r(a + B\mu, B\Omega B^T)$$

Remark The second property is very important. For instance, if (X_1, X_2) is a multivariate Gaussian, we simply have:

$$\text{Cov}(X_1, X_2) = 0 \iff X_1 \perp\!\!\!\perp X_2$$

Proof 1 We know that $u^T X \sim \mathcal{N}(u^T \mu, u^T \Omega u)$ by hypothesis. We know the MGFs of 1D normal distributions, giving us:

$$M_{u^T X}(t) = \mathbb{E}\left(e^{t u^T X}\right) = \exp\left(t u^T \mu + \frac{1}{2} t^2 u^T \Omega u\right)$$

However, we notice that:

$$M_X(u) = \mathbb{E}\left(e^{u^T X}\right) = \mathbb{E}\left(e^{1 \cdot u^T X}\right) = M_{u^T X}(1)$$

This thus gives us:

$$M_X(u) = \exp\left(u^T \mu + \frac{1}{2} u^T \Omega u\right), \quad \forall u \in \mathbb{R}^p$$

as required.

Proof 2 We only prove the 2D case, but the complete proof is very similar. Let us consider the moment generating functions. By the first property, we have:

$$\begin{aligned} M_X(t) &= \exp\left(t^T \mu + \frac{1}{2} t^T \Sigma t\right) \\ &= \exp\left(t_1 \mu_1 + t_2 \mu_2 + \frac{1}{2} t_1^2 \sigma_1^2 + \frac{1}{2} t_2^2 \sigma_2^2 + t_1 t_2 \Sigma_{1,2}\right) \end{aligned}$$

Then, because both our variables $X_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}^T X$ and $X_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}^T X$ are 1D Gaussian, we know that:

$$M_{X_1}(t) = \exp\left(t_1 \mu_1 + \frac{1}{2} t_1^2 \sigma_1^2\right), \quad M_{X_2}(t) = \exp\left(t_2 \mu_2 + \frac{1}{2} t_2^2 \sigma_2^2\right)$$

Now, we know that two polynomials are equal for all $t \in \mathbb{R}$ if and only if their coefficients are equal. Thus, we get that:

$$M_X(t) = M_{X_1}(t) M_{X_2}(t) \iff \Sigma_{1,2} = 0$$

However, we know that $M_X(t) = M_{X_1}(t) M_{X_2}(t)$ if and only if X_1 and X_2 are independent, finishing our proof.

Proof 3 Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We notice that $X \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$ indeed works, since they all have a mean μ, $\text{Var}(X_i) = \sigma^2$ and $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$.

Proof 4 Let us consider the MGF of $a + BX$:

$$\begin{aligned}
\mathbb{E}\big[\exp\big(t^T(a+BX)\big)\big] &= \mathbb{E}\Big[\exp\Big(t^T a + \big(B^T t\big)^T X\Big)\Big] \\
&= e^{t^T a}\mathbb{E}\Big[\exp\Big(\big(B^T t\big)^T X\Big)\Big] \\
&= e^{t^T a} M_X\big(B^T t\big) \\
&= \exp\Big(t^T a + \big(B^T t\big)^T \mu + \frac{1}{2}\big(B^T t\big)^T \Omega \big(B^T t\big)\Big) \\
&= \exp\Big(t^T(a+B\mu) + \frac{1}{2} t^T \big(B\Omega B^T\big) t\Big)
\end{aligned}$$

However, we recognise that this is the MGF of the $\mathcal{N}_r\big(a + B\mu, B\Omega B^T\big)$ distribution. Because of the injectivity of MGFs, we indeed get that:

$$a + BX \sim \mathcal{N}_r\big(a + B\mu, B\Omega B^T\big)$$

□

Remark We notice that any independent Gaussian variables are jointly Gaussian.
However, if our variables are all Gaussian but are dependent, it does not necessarily mean that they are jointly Gaussian.

Example 1 Let $X \sim \mathcal{N}_p(0, \Omega)$. We know that Ω is positive semi-definite, and thus that it is orthonornormally diagonalisable:

$$\Omega = UDU^T$$

If X is non-degenerate (meaning that Ω is also positive definite and thus that its eigenvalues are non-zero), then:

$$D^{-\frac{1}{2}} U^T X \sim \mathcal{N}_p(0, I_p)$$

Remark Since D is diagonal, we can compute its inverse square root by computing the inverse square root of all of its elements. We notice that it indeed has the property:

$$\Big(D^{-\frac{1}{2}}\Big)^2 = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix}^2 = \begin{pmatrix} \frac{1}{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_n} \end{pmatrix} = D^{-1}$$

Intuition Let's try to make a parallel with the 1D case, where if $X \sim \mathcal{N}\big(0, \sigma^2\big)$, then $\frac{1}{\sigma} X \sim \mathcal{N}(0, 1)$.

We thus want to make sense of $\Big(\sqrt{\Omega}\Big)^{-1}$, since it plays the role of $\frac{1}{\sigma}$. To do so, we notice that:

$$\Omega = UDU^T = UD^{\frac{1}{2}} D^{\frac{1}{2}} U^T = \Big(UD^{\frac{1}{2}}\Big)\Big(UD^{\frac{1}{2}}\Big)^T$$

where we used the fact that $D^{\frac{1}{2}}$ is diagonal.
Thus, it makes a lot of sense to define $\sqrt{\Omega} = UD^{\frac{1}{2}}$, since it gives:

$$\Omega = \sqrt{\Omega}\sqrt{\Omega}^T$$

Now, we can finally see that:

$$\Big(\sqrt{\Omega}\Big)^{-1} = \Big(UD^{\frac{1}{2}}\Big)^{-1} = \Big(D^{\frac{1}{2}}\Big)^{-1} U^{-1} = D^{-\frac{1}{2}} U^T$$

since U is orthogonal.
Finally, we indeed get that it would make a lot of sense that we have $\left(\sqrt{\Omega}\right)^{-1} X \sim \mathcal{N}_p(0, I_p)$

Proof Let's make a constructive proof. We want to find an A such that $AX \sim \mathcal{N}_p(0, I_p)$. By our theorem, we know that:

$$AX \sim \mathcal{N}_r\left(0, A\Omega A^T\right)$$

We thus want to solve $A\Omega A^T = I_p$. Since Ω is positive semi-definite, we can orthonormally diagonalise it:

$$\Omega = UDU^T = UD^{\frac{1}{2}}D^{\frac{1}{2}}U^T$$

where D is diagonal, $D_{ii} > 0$ (thanks to the positive definitness of Ω) and $U^TU = I_n$ (thanks to the orthonormality of U).
Now, we can pick $A = D^{-\frac{1}{2}}U^T$ which indeed gives:

$$A\Omega A^T = D^{-\frac{1}{2}}\underbrace{U^TU}_{I}D^{\frac{1}{2}}D^{\frac{1}{2}}\underbrace{U^TU}_{I}D^{-\frac{1}{2}} = D^{-\frac{1}{2}}D^{\frac{1}{2}}D^{\frac{1}{2}}D^{-\frac{1}{2}} = I$$

where we used $\left(D^{-\frac{1}{2}}\right)^T = D^{-\frac{1}{2}}$ since it is diagonal.

Thursday 27th April 2023 — **Lecture 17 : Blackboard lesson**

Example 2 Let $X \sim \mathcal{N}(1, 4)$ and $Y \sim \mathcal{N}(-1, 9)$ such that (X, Y) are jointly Gaussian, and:

$$\operatorname{corr}(X, Y) = -\frac{1}{6}$$

We want to compute the distribution of $W = X + Y$. Let us begin by computing $\operatorname{Cov}(X, Y)$:

$$\operatorname{Cov}(X, Y) = \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}\operatorname{corr}(X, Y) = 2 \cdot 3 \cdot \frac{-1}{6} = -1$$

Now, we compute the distribution of $\begin{pmatrix} X & Y \end{pmatrix}^T$. The expectation vector is the expectancies stacked, and the covariance matrix has $\operatorname{Cov}(X, Y)$ on its non-diagonal element:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix}\right)$$

We finally notice that we can write $W = \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$, so, by our theorem, we have:

$$\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \mathcal{N}(0, 11)$$

We notice that the fact the expectancy is 0 makes sense: X and Y have opposite expectancies, so adding them up should indeed give:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 1 - 1 = 0$$

Example 3 Let us consider $X_1, \ldots, X_4 \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, and $Y = BX$ where:

$$B = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}$$

We notice that B has orthogonal vectors, which are all of squared norm 4. This means $BB^T = 4I_4$.
Since $X = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \end{pmatrix}^T$ are independent and Gaussian, they are jointly Gaussian:

$$X \sim \mathcal{N}_4\big(0, \sigma^2 I_4\big)$$

We then get that:

$$Y = BX \sim \mathcal{N}_4\big(0, \sigma^2 BB^T\big) = \mathcal{N}_4\big(0, 4\sigma^2 I_4\big)$$

Theorem: Marginal

Let $X \sim \mathcal{N}_p(\mu_{p\times 1}, \Omega_{p\times p})$ and $\mathcal{A} \subset \{1, \dots, p\}$.
Then, we have:

$$X_{\mathcal{A}} \sim \mathcal{N}_{|\mathcal{A}|}(\mu_{\mathcal{A}}, \Omega_{\mathcal{A}})$$

Remark The important thing to remember from this theorem is that any marginal distribution of some jointly Gaussian distribution, is also jointly Gaussian.

Theorem: Conditional

Let $X \sim \mathcal{N}_p(\mu_{p\times 1}, \Omega_{p\times p})$, and $\mathcal{A}, \mathcal{B} \subset \{1, \dots, p\}$ such that $\mathcal{A} \cap \mathcal{B} = \emptyset$ and $\mathcal{A} \cup \mathcal{B} = \{1, \dots, p\}$ (in other words, $\mathcal{A}$ and $\mathcal{B}$ partition $\{1, \dots, p\}$).
If Ω is positive-definite, then:

$$X_{\mathcal{A}} | X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_{|\mathcal{A}|}\big(\mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}}\big)$$

where $\Omega_{\mathcal{A}\mathcal{B}}$ is the matrix Ω where we keep the rows $\mathcal{A}$ and the columns $\mathcal{B}$.

Remark This formula must definitely not be learnt by heart, and will be provided in the cheat sheet at the exam. However, the important thing to remember is that conditioning a jointly Gaussian distribution over another jointly Gaussian distribution also gives a jointly Gaussian distribution.

Example

Let's consider the following random variable vector:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 180 \\ 70 \end{pmatrix}, \begin{pmatrix} 225 & 90 \\ 90 & 100 \end{pmatrix}\right)$$

We notice $\mathcal{A} = \{1\}$ and $\mathcal{B} = \{2\}$, so we can then plug everything in the formula to get $X_1 | X_2 = x_2$:

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}) = 180 + 90 \cdot \frac{1}{100}(x_2 - 70) = 117 + 0.9x_2$$

$$\mathrm{Var}(X_1 | X_2 = x_2) = \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}} = 225 - 90 \cdot \frac{1}{100} \cdot 90 = 144$$

This thus means that:

$$X_1 | X_2 = x_2 \sim \mathcal{N}(117 + 0.9x_2, 144)$$

Tuesday 2nd May 2023 — **Lecture 18 : Weird structure**

Proposition

Let $X = (X_1, \dots, X_p)^T$ and $X_i \sim \mathcal{N}\big(0, \sigma_i^2\big)$ are all independent. In other words:

$$X \sim \mathcal{N}_p\left(0, \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p^2 \end{pmatrix}\right) = \mathcal{N}_p(0, \Omega)$$

Then:

$$f_X(x) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \Omega}} \exp\left(-\frac{1}{2} x^T \Omega^{-1} x\right)$$

Observation This allows us to compute the PDF of X when Ω is diagonal, but we are not yet able to compute it for the other cases.

Remark This follows the observation we made earlier: any independent Gaussian variables are jointly Gaussian (which is not necessarily true if they are dependent).

Proof The proof is rather straightforward:

$$\begin{aligned} f_X(x) &= \prod_{i=1}^{p} f_{X_i}(x_i) \\ &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{x_i^2}{2\sigma_i}\right) \\ &= \frac{1}{\sqrt{2\pi}^p \sqrt{\prod_{i=1}^{p} \sigma_i}} \exp\left(-\frac{1}{2} x^T \begin{pmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_p^2} \end{pmatrix} x\right) \\ &= \frac{1}{\sqrt{2\pi}^p \sqrt{\det \Omega}} \exp\left(-\frac{1}{2} x^T \Omega^{-1} x\right) \end{aligned}$$

□

Theorem: PDF of jointly Gaussian variables Let $X \sim \mathcal{N}_p(0, \Omega)$ be non-degenerate (meaning that Ω is positive definite). Then:

$$f_X(x) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \Omega}} \exp\left(-\frac{1}{2} x^T \Omega^{-1} x\right)$$

Proof We want to use a random variables transformation. Transformation of random vectors comes right after *(and I'm not sure why the structure was chosen to be that way in this course)*, but this is very similar to the 1D case.
Let $X \sim \mathcal{N}(0, I)$. We define $Y = AX$ such that $Y \sim \mathcal{N}(0, \Omega)$. As we saw some lectures ago, this requires:

$$\begin{aligned} A = \sqrt{\Omega}^{-1} = \sqrt{UDU^T}^{-1} = \sqrt{UD^{\frac{1}{2}}\left(UD^{\frac{1}{2}}\right)^T}^{-1} \\ \implies A = \left(UD^{\frac{1}{2}}\right)^{-1} = D^{-\frac{1}{2}} U^T \end{aligned}$$

Now, we need to compute the Jacobian. We notice that the element in the i, j position is:

$$(J(X))_{i,j} = \frac{\partial A_i x}{\partial x_j} = \frac{\partial(\sum_{k=1}^{p} A_{ik} x_k)}{\partial x_j} = A_{ij}$$

We thus get that $J(X) = A$. Now, we can apply the transformation equation (again, it will be explained shortly after, though this is completely analogous to the 1D case):

$$f_Y(y) = \frac{1}{|\det J(X)|} f_X(x) = \frac{1}{|\det A|} \frac{1}{\sqrt{2\pi}^p} e^{-\frac{1}{2} y^T \left(A^{-1}\right)^T A^{-1} y}$$

However, we know that:

$$\begin{aligned} \det(A) = \det(U) \det\left(D^{\frac{1}{2}}\right) = 1 \cdot \det\left(\sqrt{\lambda_1} \cdots \sqrt{\lambda_n}\right) \\ \implies |\det(A)| = \left|\sqrt{\lambda_1 \cdots \lambda_n}\right| = \sqrt{\det(\Omega)} \end{aligned}$$

Moreover, by definition of A, we know that:

$$AA^T = \Omega \iff \left(A^{-1}\right)^T A^{-1} = \Omega^{-1}$$

We have thus indeed got that:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det(\Omega)}} e^{-\frac{1}{2} y^T \Omega^{-1} y}$$

□

Example Let us consider:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

By our conditioning of jointly Gaussian random variables theorem, we know that:

$$X_1 | X_2 = x_2 \sim \mathcal{N}\left(\rho x_2, 1 - \rho^2\right)$$

We want to verify this particular case using the PDF of jointly Gaussian random variables.

Verifiation By definition, we know that:

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)}$$

Now, we can apply our formulas:

$$\begin{aligned} f_{X_2|X_1}(x_2|x_1) &= \frac{\frac{1}{(2\pi)^{\frac{2}{2}}} \frac{1}{\sqrt{\det \Omega}} \exp\left(-\frac{1}{2}\begin{pmatrix} x_1 & x_2 \end{pmatrix} \Omega^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\det \Omega}} \exp\left(-\frac{1}{2}\begin{pmatrix} x_1 & x_2 \end{pmatrix} \Omega^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{x_1^2}{2}\right) \end{aligned}$$

To inverse a positive semi-definite matrix, a good idea is to diagonalise it (we could naturally also use the formula for inverting 2D matrices, though this would be less general). We notice that its eigenvalue-eigenvector pairs are:

$$\left(1 + p, \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right), \quad \left(1 - p, \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)$$

This gives us that:

$$\begin{aligned} \Omega^{-1} &= U D^{-1} U^T \\ &= \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} \frac{1}{1+\rho} & 0 \\ 0 & \frac{1}{1-\rho} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \end{aligned}$$

We then get that:

$$\frac{1}{2}\begin{pmatrix} x_1 & x_2 \end{pmatrix} \Omega^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1 - \rho^2}$$

Let us now consider the exponent of $f_{X_2|X_1}(x_2|x_1)$:

$$-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1 - \rho^2} + \frac{x_2^2}{2} = -\frac{(x_1 - \rho x_2)^2}{2(1 - \rho^2)}$$

We thus finally get that:

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi}(1 - \rho^2)} \exp\left(-\frac{1}{2(1 - \rho^2)}(x_1 - \rho x_2)^2\right)$$

since $\det \Omega = (1-\rho)(1+\rho)$ because it is the product of the eigenvalues.
We recognise the formula of the 1D Gaussian distribution, which indeed tells us that:

$$X_1 | X_2 = x_2 \sim \mathcal{N}\left(\rho x_2, 1-\rho^2\right)$$

6.5 Transformations

Definition: Jacobian

Let $X = (X_1, \dots, X_n)$ be a continuous random vector, and let $Y = g(X)$:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} g_1(X_1, \dots, X_n) \\ \vdots \\ g_n(X_1, \dots, X_n) \end{pmatrix}$$

Their **Jacobian** is given by:

$$J_g(x_1, \dots, x_n) = \det \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} \in \mathbb{R}$$

If this is unambiguous, we shorten $J_g(X) = J(X)$.

Theorem: Transformation

Let $X = (X_1, \dots, X_n)$ be a continuous random vector, and let $Y = g(X)$.
If their Jacobian is non-zero (meaning that g is invertible), then:

$$\begin{aligned} f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) &= f_{X_1,\dots,X_n}(x_1, \dots, x_n) \frac{1}{|J_g(x_1, \dots, x_n)|}\Big|_{x=g^{-1}(y)} \\ &= f_{X_1,\dots,X_n}\left(g^{-1}(y_1, \dots, y_n)\right) \left|J_{g^{-1}}(y_1, \dots, y_n)\right| \end{aligned}$$

Remark

This is very similar to the 1D-case:

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Personal remark: Intuition

First, we note that, when we map a square of area dA centered at x through g, its area will be stretched by a factor $|\det J_g(x)|$. This is an important property of the Jacobian.
Let's now say that we have a probability p to land in this square of area dA. When we map it through g, this probability must definitely not change. However, since the area increased by a factor $|\det J_g(x)|$, the probability density must decrease by a factor $|\det J_g(x)|^{-1}$.

Corollary

Let's say that we are given $f_X(x)$ and $Y = AX$ for some full-rank matrix A.
Then, we know that $|J(x)| = |\det(A)|$, and thus:

$$f_Y(y) = f_X(x)|J(x)|^{-1}\Big|_{x=A^{-1}y} = \frac{f_X\left(A^{-1}y\right)}{|\det A|}$$

Thursday 4th May 2023 — **Lecture 19 : Slowly starting to do statistics**

Example 1

Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We want to compute the joint distribution $Y = (X_1 + X_2, X_1 - X_2)$.
We want to use transformation theorem. We notice that:

$$|J(x_1, x_2)| = |\det B| = |-2| = 2$$

Now, we also need to be able to write our x's as functions of y's:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = B^{-1}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

This tells us that:

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)\frac{1}{\left|J\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)\right|} \\ &= \frac{1}{2}\cdot\frac{1}{2\pi}\exp\left(-\frac{1}{4}\left(y_1^2+y_2^2\right)\right) \end{aligned}$$

We thus see that $Y_1, Y_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,2)$. Naturally, we could also have used the properties of jointly Gaussian distributions, since IID Gaussian variables are jointly Gaussian.

Example 2

Let us consider $X_1, X_2 \overset{\text{iid}}{\sim} \exp(\lambda)$. We want to compute the joint density of:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ \frac{X_1}{X_1+X_2} \end{pmatrix} = g(X)$$

Since they are IID, we have that:

$$f(x_1,x_2) = \lambda^2\exp(-\lambda(x_1+x_2))I(x_1 \geq 0)I(x_2 \geq 0)$$

Now, we need to compute the inverse map of g, h. This is not affine, so there is no general formula. From the problem, we notice that:

$$y_1 y_2 = x_1$$

Now, we can use this to get:

$$y_1 = x_1 + x_2 = y_1y_2 + x_2 \iff x_2 = y_1(1-y_2)$$

This finally gives us that:

$$x = h(y_1,y_2) = \begin{pmatrix} y_1y_2 \\ y_1(y_1-y_2) \end{pmatrix}$$

After that, we need to compute the Jacobian of g:

$$|J_g(x_1,x_2)| = \left|\det\begin{pmatrix} 1 & -\frac{1}{x_1} \\ \frac{x_2}{(x_1+x_2)^2} & -\frac{1}{(x_1+x_2)^2} \end{pmatrix}\right| = \left|-\frac{x_1+x_2}{(x_1+x_2)^2}\right| = \frac{1}{x_1+x_2} = \frac{1}{y_1}$$

Another way to see this is that:

$$|J_g(x_1,x_2)| = |J_h(y_1,y_2)|^{-1} = \left|\det\begin{pmatrix} y_2 & y_1 \\ 1-y_2 & -y_1 \end{pmatrix}\right|^{-1} = |y_1|^{-1} = \frac{1}{y_1}$$

By using our transformation theorem, this yields that:

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \left|\frac{1}{J_g(x_1,x_2)}\right| f_{X_1,X_2}(x_1,x_2) \\ &= y_1\lambda^2\exp(-\lambda(x_1+x_2))I(x_1>0)I(x_2>0) \\ &= y_1\lambda^2\exp(-\lambda y_1)I(y_1y_2>0)I(y_1(1-y_2)>0) \\ &= \underbrace{y_1\lambda^2\exp(-\lambda y_1)I(y_1>0)}_{f_{Y_1}(y_1)}\underbrace{I(0<y_2<1)}_{f_{Y_2}(y_2)} \end{aligned}$$

We notice that this is a product of two densities meaning that they are independent. This tells us that $Y_1 \sim \text{Gamma}(2,\lambda)$ and $Y_2 \sim U(0,1)$ independently.

Definition: Discrete convolution Let X, Y be discrete random variables with PMFs f_X, f_Y. The convolution of their PMFs is:

$$f_X * f_Y(s) = \sum_x f_X(x) f_Y(s-x)$$

Definition: Continuous convolution Let X, Y be continuous random variable with PDFs f_X, f_Y. The convolution of their PDFs is:

$$f_X * f_Y(s) = \int_{-\infty}^{\infty} f_X(x) f_Y(s-x) dx$$

Remark The convolution $f_{X_1} * f_{X_2}$ produces a new function, which we can evaluate: $(f_{X_1} * f_{X_2})(x)$. However, in this course, we simplify the notation by writing:

$$(f_{X_1} * f_{X_2})(x) = f_{X_1} * f_{X_2}(x)$$

Theorem Let X, Y be independent random variables with PMF or PDF f_X, f_Y.
The PMF or PDF of their sum $S = X + Y$ is:

$$f_S(s) = f_X * f_Y(s)$$

Proof The idea is to make the following change of variable:

$$\begin{pmatrix} W \\ S \end{pmatrix} = \begin{pmatrix} X \\ X+Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \implies \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} W \\ S-W \end{pmatrix}$$

The Jacobian is:

$$|J(x,y)| = \left| \det \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right| = 1$$

Thus, by applying our transformation theorem:

$$f_{W,S}(w,s) = f_{X,Y}(x,y) \frac{1}{|J(x,y)|} = f_X(x) f_Y(y) = f_X(w) f_Y(s-w)$$

since X and Y are independent.
We can now marginalise this distribution to get the one of S. In the continuous case, we get:

$$f_S(s) = \int_{-\infty}^{\infty} f_X(w) f_Y(s-w) dw$$

as required.

□

Other proof We want to make another proof, which will be more intuitive and natural.
Let's start with the CDF, since it is easier to reason about them. We notice that, using the theorem of total probability:

$$\begin{aligned} F_{X+Y}(y) &= \mathbb{P}(X+Y \leq y) \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X+Y \leq y | X = z) f_X(z) dz \\ &= \int_{-\infty}^{\infty} \mathbb{P}(Y \leq y-z | X = z) f_X(z) dz \end{aligned}$$

Now, we now that X and Y are independent, so $\mathbb{P}(Y \leq y-z | X = z) = \mathbb{P}(Y \leq y-z)$, giving us that:

$$F_{X+Y}(y) = \int_{-\infty}^{\infty} \mathbb{P}(Y \leq y-z) f_X(z) dz = \int_{-\infty}^{\infty} F_Y(y-z) f_X(z) dz$$

After that, we can differentiate to get the PDF. In this course, we are always working with sufficiently well-behaved functions, so we can bring the derivative inside of the integral:

$$f_{X+Y}(y) = \int_{-\infty}^{\infty} \frac{d}{dy} F_Y(y-z) f_X(z) dz = \int_{-\infty}^{\infty} f_Y(y-z) f_X(z) dz$$

□

Personal remark There is a great 3Blue1Brown video where they explain the intuition behind this result:
`https://www.youtube.com/watch?v=KuXjwB4LzSA`

Theorem Let $X_1, \dots, X_n$ be independent random variables with PDF $f_{X_1}, \dots, f_{X_n}$.
The PDF of their sum $S = X_1 + \dots + X_n$ is:

$$f_S(s) = f_{X_1} * \dots * f_{X_n}(s)$$

Remark This theorem is nice, but it is often not usable in practice: convolutions can be really heavy to compute.
Often, if $n > 2$, it is easier to use MGFs to turn sum of independent random variables into products, and handle such expressions without using any convolution.

6.6 Order statistics

Definition: Order statistics Let $X_1, \dots, X_n$ be random variables.
The **order statistics** of those random variables are the ordered values:

$$X_{(1)} \leq \dots \leq X_{(n)}$$

If moreover they are continuous, then we cannot have any equality, so:

$$X_{(1)} < \dots < X_{(n)}$$

Definition: Minimum Let $X_1, \dots, X_n$ be random variables. The **minimum** is $X_{(1)}$.

Definition: Maximum Let $X_1, \dots, X_n$ be random variables. The **maximum** is $X_{(n)}$.

Definition: Median Let $X_1, \dots, X_n$ be random variables. The **median**, the central value, is

$$\begin{cases} X_{(m+1)}, & n = 2m+1 \\ \dfrac{X_{(m)} + X_{(m+1)}}{2}, & n = 2m \end{cases}$$

Remark This definition is not important.

Theorem Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$ be continuous random variables with PDF f and CDF F.
Then:

1. $\mathbb{P}\big(X_{(1)} \leq x\big) = 1 - (1 - F(x))^n$
2. $\mathbb{P}\big(X_{(n)} \leq x\big) = F(x)^n$
3. $f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} f(x) (1 - F(x))^{n-r}$

Remark The last property does not have to be known.

Proof 1 We notice that:

$$\begin{aligned}\mathbb{P}\big(X_{(1)} \leq y\big) &= \mathbb{P}(\min(X_1, \ldots, X_n) \leq y) \\ &= 1 - \mathbb{P}(\min(X_1, \ldots, X_n) \geq y) \\ &= 1 - \mathbb{P}(X_1 \geq y \cap \ldots \cap X_n \geq y) \\ &= 1 - \mathbb{P}(X_1 > y) \cdots \mathbb{P}(X_n > y) \\ &= 1 - \mathbb{P}(X > n)^n \\ &= 1 - (1 - \mathbb{P}(X < n))^n \\ &= 1 - (1 - F(x))^n\end{aligned}$$

Proof 2 This proof is very similar to the one we just did:

$$\mathbb{P}\big(X_{(n)} \leq x\big) = \mathbb{P}(\max(X_1, \ldots, X_n) \leq x) = \mathbb{P}(X \leq x)^n = F(x)^n$$

□

Example Two people arrive late uniformly at random from 0 to 1 hour after the appointed time. We want to compute the expected time at which they will have both arrived. We have $X_1, X_2 \overset{\text{iid}}{\sim} U(0,1)$, and we want to compute the PDF of $V = X_{(2)}$. By our theorem, we know that:

$$F_V(v) = v^2$$

Thus, we get that:

$$f_V(v) = \frac{d\big(v^2\big)}{dv} = 2v$$

We can now compute the expected value:

$$\mathbb{E}(V) = \int_0^1 2v \cdot v dv = \frac{2}{3}$$

This means that, on average, they will both have arrived 40 minutes after the appointed time.
We could do a similar analysis to get that $\mathbb{E}\big(X_{(1)}\big) = \mathbb{E}(U) = \frac{1}{3}$. Now, if we wanted to know the average waiting time of the first arrived $W = V - U$, we can use the linearity of the expectations to get:

$$\mathbb{E}(W) = \mathbb{E}(V - U) = \mathbb{E}(V) - \mathbb{E}(U) = \frac{1}{3}$$

Note that U and V are definitely not independent, but we do not need this property for the expectations to be linear.

Chapter 7

Approximations

7.1 Inequalities

Theorem: Markov's inequality

Let X be a random variable such that $X \geq 0$, and let $a > 0$. Then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Proof

Since $X \geq 0$, we can see the following inequality, using an indicator variable:

$$X \geq XI(X \geq a) \geq aI(X \geq a)$$

Taking an expectation on both sides, we get that:

$$\mathbb{E}(X) \geq a\mathbb{E}(I(X \geq a)) = a\mathbb{P}(X \geq a)$$

This indeed allows us to conclude that:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

□

Corollaries

Let X be an arbitrary random variables. We have the following corollaries.

1. Let g be a function non-negative everywhere. Then:
$$\mathbb{P}(g(X) \geq a) \leq \frac{\mathbb{E}(g(X))}{a}$$
2. We have that:
$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|)}{a}$$
3. (Chebyshov's inequality) We have that:
$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\left(X^2\right)}{a^2}$$
4. Let g be an increasing function non-negative everywhere. Then:
$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(g(X))}{g(a)}$$
5. We have that:
$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

Proof 3 We know that:

$$\mathbb{P}(X \geq a) \leq \mathbb{P}(|X| \geq a) = \mathbb{P}\Big(|X|^2 \geq a^2\Big)$$

However, we can now apply Markov's inequality to get:

$$\mathbb{P}(X \geq a) \leq \mathbb{P}\big(X^2 \geq a^2\big) \leq \frac{\mathbb{E}\big(X^2\big)}{a^2}$$

Proof 4 We know that:

$$\mathbb{P}(X \geq a) = \mathbb{P}(g(X) \geq g(a)) \leq \frac{\mathbb{E}(g(X))}{g(a)}$$

Proof 5 We can use the fact that $\mathbb{P}(|Y| \geq a) \leq \frac{\mathbb{E}\left(Y^2\right)}{a^2}$ by letting $Y = X - \mathbb{E}(X)$:

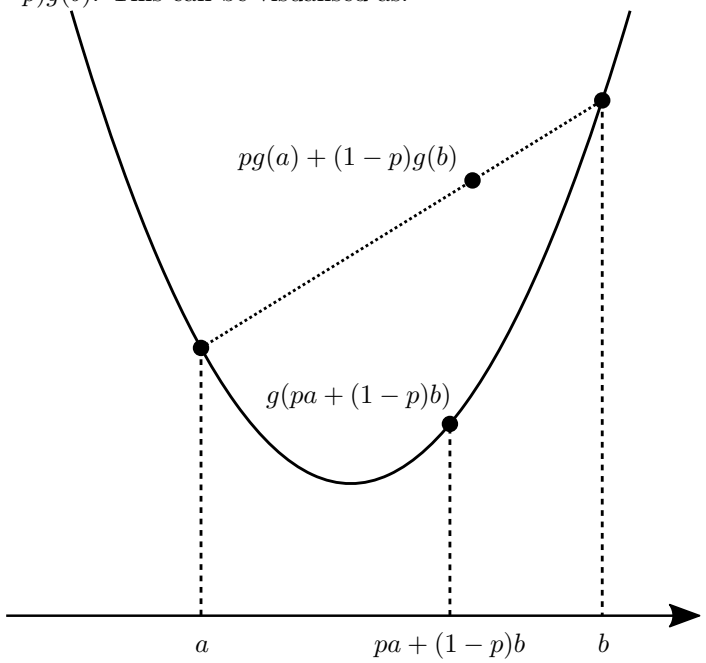
$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{E}\Big(\big(X - \mathbb{E}(X)^2\big)\Big)}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$$

□

Theorem: Jensen's inequality Let g be a convex function, and X be a random variable. Then:

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

Intuition We consider the simpler case where X is a binary random variable: it outputs a with probability p and b with probability $1 - p$.
We know that $g(\mathbb{E}(X)) = g(pa + (1-p)b)$ and $\mathbb{E}(g(X)) = pg(a) + (1-p)g(b)$. This can be visualised as:



However, the fact that g is convex means that any straight line which both endpoints are on the function, is above the function. This indeed means that:

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$$

It is possible to use this argument to make a proof for any discrete random variable. For continuous random variables, we must use other arguments; this one is really only for the intuition.

Remark This result is important.

Exemple

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, and:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We want to bound $\mathbb{P}\big(\big|\overline{X} - p\big| \geq \varepsilon\big)$ for any $\varepsilon > 0$.
The absolute value is not easy to manipulate, so let us begin with squaring both sides:

$$\mathbb{P}\big(\big|\overline{X} - p\big| \geq \varepsilon\big) = \mathbb{P}\Big(\big(X - p\big)^2 \geq \varepsilon^2\Big) \leq \frac{\mathbb{E}\Big(\big(\overline{X} - p\big)^2\Big)}{\varepsilon^2}$$

However, we notice that the numerator is the variance of $\overline{X}$. Indeed, we have that:

$$\mathbb{E}\big(\overline{X}\big) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot n\mathbb{E}(X_i) = p$$

Thus, the numerator is indeed of the form $\mathbb{E}\Big(\big(\overline{X} - \mathbb{E}\big(\overline{X}\big)\big)^2\Big) = \text{Var}\big(\overline{X}\big)$. Let us compute this:

$$\text{Var}\big(\overline{X}\big) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n\,\text{Var}(X_i) = \frac{p(1-p)}{n}$$

since $X_1, \dots, X_n$ are independent.
Putting everything together, we get that:

$$\mathbb{P}\big(\big|\overline{X} - p\big| > \varepsilon\big) \leq \frac{\text{Var}\big(\overline{X}\big)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}$$

This is very interesting since the right hand side tends towards 0 as $n \to \infty$. This means that, the average of Bernoulli random variables tends towards their mean $p = \mathbb{E}(X_i)$.

Theorem: Hoeffding's inequality

Let $Z_1, \dots, Z_n$ be independent random variables such that $\mathbb{E}(Z_i) = 0$ and such that there exists constants $a_i < b_i$ for which $a_i \leq Z_i \leq b_i$ for all i. Also, let $\varepsilon > 0$.
Then, for all $t > 0$:

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq e^{-t\varepsilon} \prod_{i=1}^{n} \exp\left(\frac{t^2(b_i - a_i)^2}{8}\right)$$

Remark Since this is true for any t, we can look for the t which minimises the expression on the right hand side (which only requires to minimise a polynomial since the exponential is an increasing function). This usually yields an inequality which is much better than the theorems we saw before.

Intuition Let $X \geq 0$. We know that:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

However, since e^x is an increasing function which is non-negative everywhere, we get that, for an arbitrary $t > 0$:

$$\mathbb{P}(X \geq a) = \mathbb{P}\big(e^{tX} > e^{ta}\big) \leq \frac{\mathbb{E}\big(e^{tX}\big)}{e^{ta}}$$

This inequality is known as Chernhoff's bound. Hoeffding's inequality's idea is similar, except that it uses the fact that the random variables are bounded instead of non-negative.

Example

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, and:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We again want to bound $\mathbb{P}\big(\big|\overline{X} - p\big| \geq \varepsilon\big)$ for any $\varepsilon > 0$, but using a better inequality. Let $Z_i = \frac{X_i - p}{n}$. We notice that we indeed have the main hypothesis for the Hoeffding's inequality:

$$\mathbb{E}(Z_i) = \frac{\mathbb{E}(X_i) - p}{n} = \frac{p - p}{n} = 0$$

We now see that our expression is equivalent to:

$$\mathbb{P}\big(\big|\overline{X} - p\big| > \varepsilon\big) = \mathbb{P}\big(\overline{X} - p > \varepsilon\big) + \mathbb{P}\big(-\big(\overline{X} - p\big) > \varepsilon\big) = \mathbb{P}\left(\sum_{i=1}^{n} Z_i > \varepsilon\right) + \mathbb{P}\left(-\sum_{i}^{n} Z_i > \varepsilon\right)$$

We know that $0 \leq X_i \leq 1$ and thus $\frac{-p}{n} \leq Z_i \leq \frac{1-p}{n}$. We can thus let $a_i = \frac{-p}{n}$ and $b_i = \frac{1-p}{n}$, giving us that:

$$(b_i - a_i)^2 = \left(\frac{1}{n}\right)^2 = \frac{1}{n^2}$$

We can then plug everything in our formula, noticing it gives the same result for $\sum_{i=1}^{n} Z_i$ and $\sum_{i=1}^{n}(-Z_i)$:

$$\begin{aligned}
\mathbb{P}\big(\big|\overline{X} - p\big| > \varepsilon\big) &= \mathbb{P}\left(\sum_{i=1}^{n} Z_i > \varepsilon\right) + \mathbb{P}\left(-\sum_{i=1}^{n} Z_i > \varepsilon\right) \\
&\leq 2e^{-t\varepsilon} \exp\left(t^2 \frac{1}{n^2}\frac{1}{8}\right)^n \\
&= \exp\left(\frac{t^2}{8n} - t\varepsilon\right)
\end{aligned}$$

We can then finally optimise this by minimising the parabola $\frac{t^2}{8n} - t\varepsilon$, which is minimal at $t = 4\varepsilon n$. This gives us:

$$\mathbb{P}\big(\big|\overline{X} - p\big| > \varepsilon\big) \leq 2e^{-2n\varepsilon^2}$$

This inequality converges much faster to 0: this is an exponential decay whereas Markov's inequality gave something decaying in $\Theta\big(\frac{1}{n}\big)$.

7.2 Convergence

Example

Let us consider the sequence of random variables defined as:

$$\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}, \quad \mathbb{P}\big(X_n = n^2\big) = \frac{1}{n}$$

We notice that:

$$\lim_{n\to\infty} \mathbb{P}(X_n = 0) = 1$$

However, we also get that:

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \lim_{n\to\infty} \left(0\left(1 - \frac{1}{n}\right) + n^2 \frac{1}{n}\right) = +\infty$$

We thus notice that X_n converges in distribution to $X = 0$ constant, even though the moments do not converge to the ones of $X = 0$.

Definition: Convergence

Let $X, X_1, X_2, \dots$ be random variables with cumulative distribution function function $F, F_1, \dots$.
Then, we define:

1. X_n converges to X **almost surely**, written $X_n \overset{\text{a.s.}}{\to} X$, if:

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1$$

2. X_n converges to X **in mean square**, written $X_n \overset{2}{\to} X$, if $\mathbb{E}\big(X_n^2\big) < \infty$ for all n, $\mathbb{E}\big(X^2\big) < \infty$ and:

$$\lim_{n\to\infty} \mathbb{E}\Big((X_n - X)^2\Big) = 0$$

3. X_n converges to X **in probability**, written $X_n \overset{\text{P}}{\to} X$, if, for all $\varepsilon > 0$:

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

4. X_n converges to X **in distribution**, writen $X_n \overset{\text{D}}{\to} X$, if:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

everywhere $F(x)$ is continuous.

Remark The first definition is not important and must not be remembered, but the third one is really important.

Theorem

Let $X, X_1, X_2 \dots$ be random variables. The modes of convergence imply one another as:

$$\begin{matrix} X_n \overset{\text{a.s.}}{\to} X \\ \\ X_n \overset{2}{\to} X \end{matrix} \Rightarrow X_n \overset{\text{P}}{\to} X \implies X_n \overset{\text{D}}{\to} X$$

Note that any other implication is false in general.

Proof We want to show the following proposition:

$$X_n \overset{2}{\to} X \implies X_n \overset{\text{P}}{\to} X$$

Let $\varepsilon > 0$. We want to show that:

$$\mathbb{P}(|X_n - X| > \varepsilon) = 0$$

Markov's inequality yields:

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}\Big((X_n - X)^2 > \varepsilon^2\Big) \leq \frac{\mathbb{E}\Big((X_n - X)^2\Big)}{\varepsilon^2}$$

However, by the definition of mean square convergence, $\mathbb{E}\Big((X_n - X)^2\Big)$ tends towards 0 as $n \to \infty$. Thus, this indeed means that:

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

□

Thursday 11th May 2023 — **Lecture 21 : It's fun we see this theorem that late**

Example 1

Let $X_1, \ldots, X_n$ be independent variables following the same distribution, with mean μ and variance σ^2. Let us consider their mean:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$$

We want to show that $\overline{X}_n \xrightarrow{2} \mu$. In other words, we want to show that the following converges to 0:

$$\mathbb{E}\Big((\overline{X}_n - \mu)^2\Big)$$

However, we note that:

$$\mathbb{E}\big(\overline{X}_n\big) = \frac{1}{n} \cdot n\mathbb{E}(X_i) = \mu$$

We thus see that:

$$\mathbb{E}\Big((\overline{X}_n - \mu)^2\Big) = \operatorname{Var}\big(\overline{X}_n\big) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n \operatorname{Var}(X_i) = \frac{\sigma^2}{n}$$

This indeed goes to 0 as $n \to \infty$.

Example 2

Let $Z \sim \mathcal{N}(0,1)$ and $X_n = (-1)^n Z$. We want to show that $X_n \xrightarrow{\mathrm{D}} Z$ but that they do not converge in any other mode.
Let us begin with showing that $X_n \xrightarrow{\mathrm{D}} Z$. Since the Gaussian is symmetric, $\varphi(x) = \varphi(-x)$, we see that:

$$F_{X_n}(x) = F_Z(x), \quad \forall x, n$$

This indeed yields that:

$$\lim_{n\to\infty} F_{X_n}(x) = F_Z(x)$$

Now, let us now show that $X_n \overset{\mathrm{P}}{\not\to} Z$. We see that, if n is odd:

$$\mathbb{P}(|X_n - Z| > \varepsilon) = \mathbb{P}(-2Z) > \varepsilon = \mathbb{P}\Big(|Z| > \frac{\varepsilon}{2}\Big) = 2\Phi\Big(-\frac{\varepsilon}{2}\Big)$$

However, since this does not go to zero as $n \to \infty$, we have indeed shown that $X_n \overset{\mathrm{P}}{\not\to} Z$.

Recall: Continuity theorem

Let $\{X_n\}, X$ be random variables with cumulative distribution functions $\{F_n\}, F$ and which MGFs $M_n(t), M(t)$ exist at all $|t| < b$, for some $b > 0$ (i.e. they exist in a neighbourhood of $t = 0$).
If there exists a $a \in \mathbb{R}$ such that $0 < a < b$ and $\lim_{n\to\infty} M_n(t) = M(t)$ for all $|t| \leq a$, then $X_n \xrightarrow{\mathrm{D}} X$.

Intuition In other words, if $\lim_{n\to\infty} M_n(t) = M(t)$ in a neighbourhood of $t = 0$, then $F_n(x) \to F(x)$ at each $x \in \mathbb{R}$ where F is continuous.

Example Let X be a random variable with a geometric distribution with a probability of success p. We want to calculate the limit distribution of pX as $p \to 0$.
We have always seen limits of sequences, now we are considering a continuous limit. A way to tackle this is to define $p = \frac{1}{n}$, giving us $X_n = \frac{X}{n}$. This supposes that the limit converges and is thus not perfect, but it works. However, let us do this proof completely formally here, without making any assumption on the convergence.
Let us consider the MGF:

$$\begin{aligned}
\mathbb{E}\big(e^{tpX}\big) &= \sum_{x=1}^{\infty} e^{tpx} p(1-p)^{x-1} \\
&= pe^{tp} \sum_{x=0}^{\infty} \big(e^{tp}(1-p)\big)^x \\
&= \frac{pe^{tp}}{1-(1-p)e^{tp}} \\
&= \frac{p}{e^{-tp}-1+p} \\
&= \frac{1}{1+\frac{e^{-tp}-1}{p}} \to \frac{1}{1-t}
\end{aligned}$$

for $|t| < 1$, using the particular limit $\lim_{x\to\infty} \frac{e^x-1}{x} = 1$.
We notice that this is the MGF of $Y \sim \exp(1)$. By the continuity theorem, we showed that $\frac{1}{p}X \xrightarrow{\mathrm{D}} Y$.

Theorem Let x_0 be a constant, $X, Y, \{X_n\}, \{Y_n\}$ be random variables and h a function continuous at x_0.
We have the following implication:

$$X_n \xrightarrow{\mathrm{D}} x_0 \implies h(X_n) \xrightarrow{\mathrm{P}} h(x_0)$$

Implication In particular, we have:

$$X_n \xrightarrow{\mathrm{D}} x_0 \implies X_n \xrightarrow{\mathrm{P}} x_0$$

Remark We know that we always have:

$$X_n \xrightarrow{\mathrm{P}} X \implies X_n \xrightarrow{\mathrm{D}} X$$

The converse is wrong in general. However, when we have $X = x_0$ constant, then this theorem tells us that we have both directions.

Proof We want to show the particular case where:

$$X_n \xrightarrow{\mathrm{D}} c \implies X_n \xrightarrow{\mathrm{P}} c$$

Since $X_n \xrightarrow{\mathrm{D}} c$, we know by definition that:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

where the CDF of a constant is a step function:

$$F(x) = \mathbb{P}(X \le c) = \begin{cases} 1, & x > c \\ 0, & x < c \end{cases}$$

In particular, this means that, for any $\varepsilon > 0$:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|X_n - c| < \varepsilon) &= \lim_{n\to\infty} \mathbb{P}(c - \varepsilon < X_n < c + \varepsilon) \\ &= \lim_{n\to\infty} (F_n(c+\varepsilon) - F_n(c-\varepsilon)) \\ &= F(c+\varepsilon) - F(c-\varepsilon) \\ &= 1 - 0 \\ &= 1\end{aligned}$$

This indeed yields that $X_n \xrightarrow{\text{P}} c$.

□

Slutsky's lemma Let x_0, y_0 be constants, and $X, Y, \{X_n\}, \{Y_n\}$ be random variables.
If $X_n \xrightarrow{\text{D}} X$ and $Y_n \xrightarrow{\text{P}} y_0$, then we have:

$$X_n + Y_n \xrightarrow{\text{D}} X + y_0$$

Similarly, if $X_n \xrightarrow{\text{D}} X$ and $Y_n \xrightarrow{\text{P}} y_0$, then:

$$X_n Y_n \xrightarrow{\text{D}} X y_0$$

Example Let $X_1, \ldots, X_n$ be IID with mean μ_X and variance σ_X^2, and let $Y_1, \ldots, Y_n$ be IID with mean μ_Y and variance σ_Y^2. We also suppose that $\mu_Y \neq 0$.
We want to show that:

$$\frac{\overline{X}_n}{\overline{Y}_n} \xrightarrow{\text{P}} \frac{\mu_X}{\mu_Y}$$

We have already shown that $\overline{X}_n \xrightarrow{\text{P}} \mu_X$ and $\overline{Y} \xrightarrow{\text{P}} \mu_Y$. In particular, we have $\overline{Y}_n^{-1} \xrightarrow{\text{D}} \mu_Y^{-1}$ which gives us that:

$$\frac{\overline{X}_n}{\overline{Y}_n} \xrightarrow{\text{D}} \frac{\mu_X}{\mu_Y}$$

However, by our theorem, this tells us that:

$$\frac{\overline{X}_n}{\overline{Y}_n} \xrightarrow{\text{P}} \frac{\mu_X}{\mu_Y}$$

Degenerate convergence Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $M_n = \max\{X_1, \ldots, X_n\}$.
We already know that:

$$\mathbb{P}(M_n \leq x) = F(x)^n$$

However, as $n \to \infty$, this tends towards:

$$\mathbb{P}(M_n \leq x) = \begin{cases} 0, & F(x) < 1 \\ 1, & F(x) = 1 \end{cases}$$

since $b^n \to 0$ for any $0 \leq b < 1$.
This is a degenerate limit distribution, so we may want to center and scale:

$$Y_n = \frac{M_n - b_n}{a_n}$$

There sometimes exists sequences a_n, b_n such that the distribution is not degenerate.

Example Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$, and let M_n be their maximum. We want to find a_n, b_n such that:

$$Y_n = \frac{M_n - b_n}{a_n} \xrightarrow{\text{D}} Y$$

where Y has a non-degenerate distribution.
We notice that:

$$\mathbb{P}(Y_n \leq y) = \mathbb{P}(M_n \leq a_n y + b_n) = F(b_n + a_n y)^n = (1 - \exp(-b_n\lambda - a_n\lambda y))^n$$

To simplify everything, we decide to pick $a_n = \frac{1}{\lambda}$ and $b_n = \frac{\log(n)}{\lambda}$, giving:

$$\mathbb{P}(Y_n \leq y) = \left(1 - \frac{\exp(-y)}{n}\right)^n \to \exp(-\exp(y))$$

This is indeed a non-degenerate distribution (which is named the Gumbel distribution).

7.3 Adding many random variables

Theorem: Weak law of large numbers

Let $X_1, X_2, \ldots$ be IID random variables with finite expectation μ. Also, let us consider their average:

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

Then, we have:

$$\overline{X}_n \xrightarrow{\text{P}} \mu$$

Remark We have already shown that, if the $X_1, \ldots$ have a finite variance:

$$\overline{X}_n \xrightarrow{2} \mu$$

However, this theorem does not require this hypothesis.

Remark 2 This theorem can be generalised to the strong law of large numbers which says that, under the same hypothesis:

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu$$

This is not important to know.

Justification We want to make a proof where we don't take the assumption that $\sigma^2 < \infty$. We will use MGFs to do so; we assume convergence, even though we would need to prove it for a formal proof. Let us consider the MGF of $\overline{X}_n$:

$$\begin{aligned} M_{\overline{X}_n}(t) &= \mathbb{E}\left(\exp\left(\frac{t}{n}\sum_{i=1}^{n} X_i\right)\right) \\ &= M_{X_1}\left(\frac{t}{n}\right)^n \\ &= \left(1 + \frac{t}{n}\mathbb{E}(X_1) + \frac{t^2\mathbb{E}\left(X_1^2\right)}{2n^2} + \ldots\right)^n \end{aligned}$$

since the random variables are independent.
Now, we notice that:

$$M_{\overline{X}_n}(t) = \left(1 + \frac{t}{n}\mathbb{E}(X_1) + \frac{t^2\mathbb{E}\left(X_1^2\right)}{2n^2} + \ldots\right)^n \to e^{t\mathbb{E}(X_1)}$$

since $\left(1 + \frac{\lambda}{n} + \frac{c}{n^2}\right)^n \to e^{\lambda}$, everything which decays faster than $\frac{1}{n}$ is just not considered (this is something else we would need to show for a complete proof).

However, we recognise the MGF of the degenerate random variable $Z = \mathbb{E}(X_1)$ constant:

$$M_Z(t) = \mathbb{E}\big(e^{tZ}\big) = \mathbb{E}\Big(e^{t\mathbb{E}(X_1)}\Big) = e^{t\mathbb{E}(X_1)}$$

We have thus shown by the continuity theorem that:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \xrightarrow{\mathrm{D}} \mathbb{E}(X_1)$$

However, since $\mathbb{E}(X_1)$ is constant, we can use our theorem to get:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mathbb{E}(X_1)$$

Observation Let us consider IID random variables $X_1, \dots$ such that $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$. We have seen with the law of large numbers that:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathrm{P}} \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 0$$

Indeed, the variance shrinks to 0. Let us consider another scaling, so that it does not do so:

$$Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$$

This implies that:

$$\mathbb{E}(Y_n) = \mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right) = \frac{1}{\sqrt{n}} \cdot n\mathbb{E}(X_i) = 0$$

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot n\,\mathrm{Var}(X_i) = 1$$

This leads to the following theorem.

Lemma Let $X_1, X_2, \dots$ be independent (but not necessarily identically distributed) random variables which all have the expectation 0 and variance 1. Also, let $Z \sim \mathcal{N}(0,1)$. Then, we have that:

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_n = \sqrt{n} \cdot \overline{X} \xrightarrow{\mathrm{D}} Z$$

Justification Let us consider the MGF of $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$:

$$\begin{aligned} M_{Z_n}(t) &= \mathbb{E}\left(\exp\left(\frac{t}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)\right) \\ &= \left(M_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{t}{\sqrt{n}}\mathbb{E}(X_1) + \frac{t^2\mathbb{E}\big(X_1^2\big)}{\sqrt{n}^2 \cdot 2} + \dots\right)^n \end{aligned}$$

However, we know by hypothesis that $\mathbb{E}(X_1) = \mu = 0$, and:

$$\mathbb{E}\big(X_1^2\big) = \mathbb{E}\big(X_1^2\big) - \underbrace{\mathbb{E}(X_1)^2}_{=0} = \mathrm{Var}(X_1) = 1$$

This yields that:

$$M_{\frac{Y_n}{\sqrt{n}}}(t) = \left(1 + \frac{t^2}{2n} + \dots\right)^n \to e^{\frac{t^2}{2}}$$

by a similar reasoning to the justification of the weak law of large numbers.
However, we recognise the MGF of $\mathcal{N}(0,1)$, which indeed implies that $Z_n \xrightarrow{\text{D}} Z$ by the continuity theorem. Note that we are again hiding the fact that we can really ignore everything $o\left(\frac{1}{n^2}\right)$.

Central limit theorem

Let $X_1, X_2, \dots$ be independent (but not necessarily identically distributed) random variables which all have the same expectation μ and non-zero finite variance σ^2. Also, let $Z \sim \mathcal{N}(0,1)$.
Then, we have that:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_n - \mu}{\sigma} = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{\text{D}} Z$$

Implication

This for instance implies that

$$\mathbb{P}\left(\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \leq z\right) \to \mathbb{P}(Z \leq z) = \Phi(z)$$

Proof

Let us consider the variance and mean of $\frac{X_i - \mu}{\sigma}$:

$$\mathbb{E}\left(\frac{X_i - \mu}{\sigma}\right) = \frac{\mathbb{E}(X_i) - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

$$\operatorname{Var}\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma^2} \operatorname{Var}(X_i - \mu) = \frac{1}{\sigma^2} \operatorname{Var}(X_i) = \frac{\sigma^2}{\sigma^2} = 1$$

This means that we can just use our lemma on those shifted random variables.

□

Chapter 8

Statistical inference

8.1 Introduction

Goal

The goal is, given some observations, we want to make some inference on their probability space. We don't know their distribution, but we know that it is in some family and we want to infer the parameters.
In fact, in the past, we even used the term inverse probability to speak about inveferential statistics.
An **estimator** of the unknown parameters is a function of $y_1, \dots, y_n$. We can also consider a random variable lead by the estimator, $T(Y_1, \dots, Y_n)$.

Definition: Random sample

We will always take $Y_1, \dots, Y_n$ to be IID from the distribution we don't know, named a **random sample**, and $y_1, \dots, y_n$ to be a **realisation** of such random variables. Using those observations, we want to estimate the parameters of the distribution.

Definition: Statistical model

A **statistical model** is a probability distribution $f(y)$ chosen to learn from observed data y or potential data Y.
We note it $f(y) = f(y; \theta)$ to represent the fact that θ is a parameter of the model.

Definition: Statistic

A **statistic** is a function of the random variables, $T = t(Y_1, \dots, Y_n)$.

Example

Let $y_1, \dots, y_n$ be a random sample from a Bernoulli distribution with unknown parameter p. Then, the statistic $t = \sum_{j=1}^{n} y_j$ is considered to be a realisation of the random variable:

$$T = \sum_{j=1}^{n} Y_j \sim B(n, p)$$

8.2 Point estimation

Method of moments

The idea of the **method of moments** is to use the law of large numbers:

$$\frac{1}{n} \sum_{i=1}^{n} y_i^r \xrightarrow{\text{P}} \mathbb{E}(Y_1^r)$$

Thus, we simply set:

$$\mathbb{E}(Y^r) = \frac{1}{n} \sum_{j=1}^{n} y_j^r$$

We take $r = 1, \dots, p$ to have the smallest number of equations which would give a unique solution.

Example 1

Let $Y_1, \dots, Y_n \sim U(0, \theta)$. We want to use the method of moments to estimate θ. We let:

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \mathbb{E}(Y_1) \iff \overline{y} = \frac{\theta}{2}$$

This thus gives us:

$$\hat{\theta}_{MoM} = 2\overline{y} \implies \hat{\Theta}_{MoM} = 2\overline{Y}_n$$

Example 2

Let $Y_1, \dots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown. We can estimate them using the method of moments:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} y_i = \mu \\ \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \mathbb{E}\left(Y_i^2\right) = \mathrm{Var}(Y_i) + \mathbb{E}(Y_i)^2 = \sigma^2 + \mu^2 \end{cases}$$

We notice that this is solved by:

$$\begin{cases} \hat{\mu}_{MoM} = \overline{y} \\ \hat{\sigma}^2_{MoM} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 \end{cases}$$

Definition: Likelihood

Let $y_1, \dots, y_n$ be a random sample from a density $f(y; \theta)$. The **likelihood** for θ is defined as:

$$L(\theta) = f(y_1, \dots, y_n; \theta) = f(y_1; \theta) \cdots f(y_n; \theta)$$

since all samples independent.

Maximum likelihood method

The **maximum likelihood estimate** (MLE) $\hat{\theta}$ of a parameter θ is the value that maximises the likelihood:

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} L(\theta)$$

Remark We often simplify the calculations by maximising $\ell(\theta) = \log(L(\theta))$, which is equivalent to maximising $L(\theta)$ since the logarithm is an increasing function. This is often simpler since we will use derivatives (when the function is differentiable) and differentiating additions is easier than differentiating multiplications.

Example 1

Let us consider a random sample $y_1, \dots, y_n$ from an exponential density with unknown λ.
We want to maximise:

$$L(\theta) = \lambda e^{-\lambda y_1} \cdots \lambda e^{-\lambda y_n} = \lambda^n \exp(-\lambda(y_1 + \ldots + y_n))$$

Applying a natural logarithm on both sides, we get:

$$\ell(\lambda) = \log(L(\lambda)) = n\log(\lambda) - n\lambda\overline{y}$$

We want to maximise this function, so let us compute its derivative:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - n\overline{y}$$

Setting $\frac{d\ell(\lambda)}{d\lambda} = 0$, we get $\lambda = \frac{1}{\overline{y}}$ (since $n > 0$). We notice that we have:

$$\frac{d^2\ell\left(\frac{1}{\overline{y}}\right)}{d\lambda^2} = -\frac{n}{\left(\frac{1}{\overline{y}}\right)^2} < 0$$

This critical point is thus indeed a maximum. This yields that:

$$\hat{\lambda}_{MLE} = \frac{1}{\overline{y}}$$

Example 2

Let us consider a random sample $y_1, \dots, y_n$ from a uniform density with unknown θ. We want to maximise:

$$L(\theta) = \prod_{j=1}^{n} \frac{I(0 < y_j \leq \theta)}{\theta} = \frac{I(0 < y_1, \dots, y_n \leq \theta)}{\theta^n} = \frac{I(\max(y_1, \dots, y_n) \leq \theta)}{\theta^n}$$

We cannot differentiate this function because the indicator random variable is not even continuous. However, we notice that we want θ to be the smallest possible for θ^{-n} to be the greatest. Thus we need to take:

$$\hat{\theta}_{MLE} = \max(y_1, \dots, y_n)$$

We notice that this is different from $\hat{\theta}_{MoM} = 2\overline{y}$, even though both make sense.

Tuesday 23rd May 2023 — **Lecture 23 : Confidence intervals**

Definition: M-estimation

In the ***M*-estimation** method, we maximise a function of the form:

$$\overline{\rho}(\theta; Y) = \sum_{j=1}^{n} \rho(\theta; Y_j)$$

where $\rho(\theta; y)$ is a function of θ, typically concave for all y.
Note that there is an abuse of notation: $\overline{\rho}$ is not the same function as ρ, even though we often write them both the same way (without the bar).

Observation This generalises maximum likelihood estimation: we can take $\rho(\theta; y) = \log(f(y; \theta))$ to get it back.

Least-squares estimator This yields the **least-squares** estimator when we let $\rho(\theta; y) = -(y_j - \theta)^2$.

Example

Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} f$ such that $\mathbb{E}(Y_j) = \theta$ for some θ. We want to find the least-squares estimator of θ.
By definition, we want to maximise:

$$\overline{\rho}(\theta; Y) = -\sum_{j=1}^{m} (\theta - Y_j)^2$$

This function is indeed concave, so it has a single maximum. The derivative gives us:

$$\frac{\overline{\rho}(\theta; y)}{d\theta} = \sum_{j=1}^{n} 2(y_j - \theta)$$

This yields that:

$$\frac{\overline{\rho}(\theta; y)}{d\theta} = 0 \iff \sum_{j=1}^{n} y_j = \sum_{j=1}^{n} \theta = n\theta \iff \hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} y_j = \overline{y}$$

Definition: Bias

The **bias** of the estimator $\hat{\theta}$ of θ is:

$$b(\theta) = \mathbb{E}\left(\hat{\theta}\right) - \theta$$

Interpretations

- If $b(\theta) < 0$ for all θ, we say that $\hat{\theta}$ **underestimates** θ on average.
- If $b(\theta) > 0$ for all θ, we say that $\hat{\theta}$ **overestimates** θ on average.
- If $b(\theta) = 0$ for all θ, we say that $\hat{\theta}$ is **unbiased**.

- If $b(\theta) \approx 0$ for all θ, we say that $\hat{\theta}$ is **in the right place** on average.

Remark This gives us a way to compare estimators: if the bias is much worse in absolute value, then the estimator is worse.

Example 1 Let us consider $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We want to find the bias of $\hat{\mu} = \overline{Y}$:

$$b(\mu) = \mathbb{E}(\hat{\mu}) - \mu = \mathbb{E}(\overline{Y}) - \mu = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}(Y_j) - \mu = \mu - \mu = 0$$

We thus see that $\hat{\mu}$ is unbiased.

Example 2 Let us consider $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We want to find the bias of $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$:

$$b(\hat{\sigma}^2) = \mathbb{E}(\hat{\sigma}^2) - \sigma^2 = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[(Y_j - \overline{Y})^2\right] - \sigma^2$$

Let us consider the expected value term, which we recognise to be the variance since $\mathbb{E}(Y_j - \overline{Y}) = 0$:

$$\mathbb{E}\left[(Y_j - \overline{Y})^2\right] = \operatorname{Var}(Y_j - \overline{Y}) = \operatorname{Var}\left(Y_j - \frac{1}{n}\sum_{k=1}^{n} Y_k\right)$$

We split the sum since we know the variance of a sum of independent random variables is the sum of variance and we would like to use this property:

$$\begin{aligned}
\operatorname{Var}\left(\frac{n-1}{n}Y_j - \frac{1}{n}\sum_{k\neq j} Y_j\right) &= \operatorname{Var}\left(\frac{n-1}{n}Y_j\right) + \sum_{k\neq j}\operatorname{Var}\left(\frac{1}{n}Y_k\right) \\
&= \left(\frac{n-1}{n}\right)^2 \operatorname{Var}(Y_j) + \sum_{k\neq j}\frac{1}{n^2}\operatorname{Var}(Y_j) \\
&= \left[\frac{(n-1)^2}{n^2} + \frac{n-1}{n^2}\right]\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

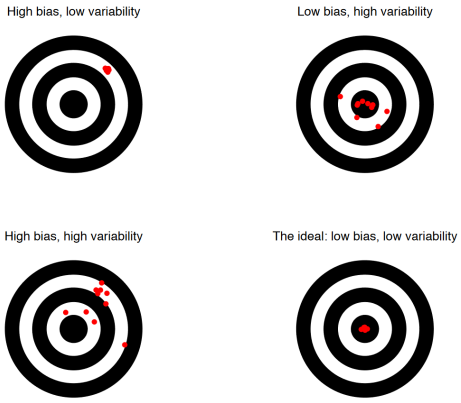
Coming back to our original computation, we get that:

$$b(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

In other words, $\hat{\sigma}^2$ underestimates σ^2, by an amount which should be rather small when n is large.

Remark The bias is important in order to choose the estimator, but it is not the only fact we have to take into account. The variance of the estimator is also really important (recall that the variance represents the mean squared distance to the expected value).

Interpretation We can imagine our point estimator as aiming to the bullseye of a target. If we have a high bias, we are not aiming at the right place, but if we have a lot of variance there our shots are very spread out.



Definition: Mean square error

The **mean square error** (MSE) of the estimator $\hat{\theta}$ of θ is:

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

Proposition

Let $\hat{\theta}$ be an estimator of θ. Then:

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathrm{Var}\left(\hat{\theta}\right) + b(\theta)^2$$

Interpretation The mean square error is thus a good way to compare estimators: we have a measure of both variance and bias.

Proof This is a direct proof:

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\theta}\right) &= \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right) + \mathbb{E}\left(\hat{\theta}\right) - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}(\theta)\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right)\right)\left(\mathbb{E}\left(\hat{\theta}\right) - \theta\right)\right] \\
&\quad + \mathbb{E}\left[\left(\mathbb{E}\left(\hat{\theta}\right) - \theta\right)^2\right] \\
&= \mathrm{Var}\left(\hat{\theta}\right) + 2b\left(\hat{\theta}\right)\mathbb{E}\left[\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right)\right] + b\left(\hat{\theta}\right)^2 \\
&= \mathrm{Var}\left(\hat{\theta}\right) + b\left(\hat{\theta}\right)^2
\end{aligned}$$

Indeed, we have:

$$\mathbb{E}\left[\hat{\theta} - \mathbb{E}\left(\hat{\theta}\right)\right] = \mathbb{E}\left(\hat{\theta}\right) - \mathbb{E}\left[\mathbb{E}\left(\hat{\theta}\right)\right] = \mathbb{E}\left(\hat{\theta}\right) - \mathbb{E}\left(\hat{\theta}\right) = 0$$

□

Observation

Let $\hat{\theta}$ be an unbiased estimator of some parameter θ. We notice that we can compute its variance by using the MSE:

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + b(\theta)^2 = \mathrm{Var}\left(\hat{\theta}\right)$$

Definition: Efficiency

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of the same parameter θ.
We say that $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if:

$$\mathrm{Var}\left(\hat{\theta}_1\right) \leq \mathrm{Var}\left(\hat{\theta}_2\right)$$

Remark We prefer estimators which are more efficient.

Example

Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for a large n. We want to compare the median M (the value such that half of the samples are above and half are below) and the average $\overline{Y}$ as estimators of μ.
We have already computed that:

$$b(\overline{Y}) = 0$$

Moreover:

$$\text{Var}(\overline{Y}) = \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\text{Var}(Y_j) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Now, it is possible to show that, when n is large, the median M is approximately distributed as:

$$M \sim \mathcal{N}\left(\mu, \frac{\pi\sigma^2}{2n}\right)$$

This means that:

$$\mathbb{E}(M) \approx \mu, \quad \text{Var}(M) \approx \frac{\pi\sigma^2}{2n}$$

Thus, it is also unbiased asymptotically. However, the variance of M is slightly worse than $\overline{Y}$, so the latter is slightly more efficient asymptotically. Now, in practice the former might be better if we have many outliers (many points which are far away) or that don't really belong to $\mathcal{N}(\mu, \sigma^2)$ (for instance because of measurement errors).

Proposition

Under certain conditions (for instance that the PDF is "nice" and that the samples are really from the supposed PDF), the maximum likelihood estimator is best for large n: it is unbiased and has a minimal variance.

Remark In practice, we often sacrifice some efficiency for some robustness to outliers.

8.3 Interval estimation

Definition: Pivot

Let $Y = (Y_1, \dots, Y_n)$ be sampled from a distribution F with parameter θ.
A **pivot** is a function $Q = q(Y; \theta)$ which distribution is known and independent of θ. We say that Q is **pivotal**.

Example 1

Let us consider $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$ for an unknown θ and:

$$M = \max(Y_1, \dots, Y_n)$$

We want to show that $Q = \frac{M}{\theta}$ is a pivot, meaning that its distribution is independent of θ. By definition, we have:

$$Q = \frac{1}{\theta}M = \frac{1}{\theta}\max(Y_1, \dots, Y_n) = \max\left(\frac{Y_1}{\theta}, \dots, \frac{Y_n}{\theta}\right)$$

However, $\frac{Y_1}{\theta}, \dots, \frac{Y_n}{\theta} \overset{\text{iid}}{\sim} U(0, 1)$. This indeed means that θ does not matter for Q, and thus that Q is a pivot.

Example 2

Let us consider $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$ for an unknown θ and:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

We want to find an approximate pivot. To do so, let us try to use the central limit theorem:

$$\frac{\overline{Y} - \mathbb{E}(\overline{Y})}{\sqrt{\text{Var}(\overline{Y})}} = \frac{\overline{Y} - \frac{\theta}{2}}{\sqrt{\frac{\theta^2}{12n}}} \xrightarrow{\text{D}} \mathcal{N}(0, 1)$$

Thus, we can take the following random variable, which is asymptotically a pivot (and thus it is an approximate pivot):

$$Q = \frac{\overline{Y} - \frac{\theta}{2}}{\sqrt{\frac{\theta^2}{12n}}}$$

Definition: Confidence interval

Let $Y = (Y_1, \dots, Y_n)$ be data from a parametric statistical model with scalar parameter θ.
A **confidence interval** (CI) for θ, $(L, U) = (\ell(Y), u(Y))$, is a random interval which contains θ with a specified probability, called the **confidence level** of the interval, specified by the following probabilities:

$$\mathbb{P}(\theta < L) = \alpha_L, \quad \mathbb{P}(U < \theta) = \alpha_U$$

Observation The confidence level can be computed by seeing that:

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \mathbb{P}(\theta < L) - \mathbb{P}(U < \theta) = 1 - \alpha_L - \alpha_U$$

Terminology It is typical to take a $\alpha_L = \alpha_U = \frac{\alpha}{2}$ to have a symmetric probability of not containing θ. This is named an **equi-tailed** $(1 - \alpha)$ confidence interval.

Remark The random variables L and U must not depend from θ. Our goal is to find them for some data for which we don't know the distribution, and thus of which we don't know the actual θ.

Construction of a CI

To construct a CI, we first need to find a pivot $Q = q(Y, \theta)$. We then need to choose α_L and α_U, and obtain the quantiles $q_{\alpha_U}, q_{1-\alpha_L}$ of Q. By definition of quantiles, this gives us:

$$\mathbb{P}(q_{\alpha_U} \leq q(Y, \theta) \leq q_{1-a_L}) = (1 - \alpha_L) - \alpha_U$$

We finally need to inverse q as a function of θ, to convert this equation into the form:

$$\mathbb{P}(L \leq \theta \leq U) \leq 1 - \alpha_L - \alpha_U$$

where the bounds L and U only depend on Y, $q_{1-\alpha_L}$ and q_{α_U}, but not on θ.

Remark We require that Q is a pivot to be able to compute the quantiles q_{α_U} and $q_{1-\alpha_L}$, so that they are independent of θ.

Example 1

Let us consider $Y_1, \dots, Y_n \sim U(0, \theta)$. We have already seen that the following random variable is a pivot:

$$Q = \frac{M}{\theta} = \frac{\max(Y_1, \dots, Y_n)}{\theta} = \max\left(\frac{Y_1}{\theta}, \dots, \frac{Y_n}{\theta}\right)$$

Now, let us choose $\alpha_L = \alpha_U = 0.05$. We need to compute the quantiles of Q, which is very easy to compute when we can inverse the CDF. We saw how to compute the CDF of maximums, giving:

$$F_Q(x) = \mathbb{P}(Q \leq x) = \mathbb{P}\left(\max\left(\frac{Y_1}{\theta}, \dots, \frac{Y_n}{\theta}\right) \leq x\right) = F_{\frac{Y_1}{\theta}}(x)^n = x^n$$

for $x \in [0,1]$.
We can now invert our CDF. By definition of the quantiles, they are such that:

$$F_Q(q_\alpha) = \alpha \iff q_\alpha^n = \alpha \iff q_\alpha = \alpha^{\frac{1}{n}}$$

Again by definition of the quantiles, we know that:

$$\mathbb{P}(q_{\alpha_u} \leq Q \leq q_{1-\alpha_L}) = 1 - \alpha_L - \alpha_U \iff \mathbb{P}\Big(\alpha_u^{\frac{1}{n}} \leq Q \leq (1-\alpha_L)^{\frac{1}{n}}\Big) = 1 - \alpha_L - \alpha_U$$

We now only need to consider the inequality inside the probability:

$$\begin{aligned}
&\alpha_U^{\frac{1}{n}} \leq Q \leq (1-\alpha_L)^{\frac{1}{n}} \\
\iff &\alpha_U^{\frac{1}{n}} \leq \frac{M}{\theta} \leq (1-\alpha_L)^{\frac{1}{n}} \\
\iff &\frac{1}{(1-\alpha_L)^{\frac{1}{n}}} \leq \frac{\theta}{M} \leq \frac{1}{\alpha_U^{\frac{1}{n}}} \\
\iff &\frac{M}{(1-\alpha_L)^{\frac{1}{n}}} \leq \theta \leq \frac{M}{\alpha_U^{\frac{1}{n}}}
\end{aligned}$$

Thus, we can let:

$$L = \frac{M}{(1-\alpha_L)^{\frac{1}{n}}}, \quad U = \frac{M}{\alpha_U^{\frac{1}{n}}}$$

We then indeed have:

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha_L - \alpha_U$$

Example 2 We have a sample of $n = 16$ plates with maximum 523308 and average 320869. We suppose that all plates from 0 to θ are given (where θ is the number of cars), uniformly at random. We want to find a confidence interval with $\alpha_U = \alpha_L = 2.5\,\%$. We can apply the bound we have just computed, giving :

$$L = \frac{m}{(1-\alpha_L)^{\frac{1}{n}}} = 524135, \quad U = \frac{m}{\alpha_U^{\frac{1}{n}}} = 659001$$

It is interesting to see that L is slightly above the maximum we have seen, which makes a lot of sense.

Thursday 25th May 2023 — **Lecture 24 : Finally, the null hypotheses**

Example 3 Let us consider again $Y_1, \ldots, Y_n \sim U(0,\theta)$. We have already seen that the following random variable is an approximate pivot:

$$Q = \frac{\overline{Y} - \frac{\theta}{2}}{\sqrt{\frac{\theta^2}{12n}}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$$

We want to find a random interval with symmetric confidence level $\frac{\alpha}{2}$. The quantiles of a normal distribution can be computed numerically and, by symmetry, $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. Since the second is positive for $\frac{\alpha}{2} < 0.5 \iff \alpha < 1$, we prefer it by simplicity for inequalities. We thus have:

$$\mathbb{P}\big(-z_{1-\frac{\alpha}{2}} \leq Q \leq z_{1-\frac{\alpha}{2}}\big) \to 1 - \alpha$$

Now, considering the inequality inside the probability:

$$\begin{aligned} & -z_{1-\frac{\alpha}{2}} \leq \frac{\overline{y}-\frac{\theta}{2}}{\sqrt{\frac{\theta^2}{12n}}} \leq z_{1-\frac{\alpha}{2}} \\ \iff & -z_{1-\frac{\alpha}{2}} \leq \frac{\frac{\overline{y}}{\theta}-\frac{1}{2}}{\frac{1}{\sqrt{12n}}} \leq z_{1-\frac{\alpha}{2}} \\ \iff & \frac{1}{2}-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}} \leq \frac{\overline{y}}{\theta} \leq \frac{1}{2}+\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}} \\ \iff & \frac{\overline{y}}{\frac{1}{2}+\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}}} \leq \theta \leq \frac{\overline{y}}{\frac{1}{2}-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}}} \end{aligned}$$

This yields that we can take:

$$L = \frac{\overline{Y}}{\frac{1}{2}+\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}}}, \quad U = \frac{\overline{Y}}{\frac{1}{2}-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{12n}}}$$

Giving that:

$$\lim_{n\to\infty} \mathbb{P}(\theta \in [L,U]) = 1-\alpha$$

Remark We can always use the CLT to make an approximate confidence interval when we are considering a sum of random variable.

Theorem Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then, we have the following independent estimator:

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sum_{j=1}^{n} \left(Y_j - \overline{Y}\right)^2 \sim \sigma^2 \chi^2_{n-1}$$

where χ^2_ν represents the chi-square distribution with ν degrees of freedom (it will be defined shortly after).

Remark The first result implies the CLT, giving the following pivot:

$$Z = \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$$

This pivot yields the following $(1-\alpha_L-\alpha_U)$ CI for μ:

$$(L,U) = \left(\overline{Y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha_L}, \overline{Y} - \frac{\sigma}{\sqrt{n}} z_{\alpha_U}\right)$$

Interpretation It is important to understand that θ is not sampled randomly in (L,U), it is the inverse. θ is fixed but we don't have access to it. However, we get some random observations, which yield the random interval (L,U).
In other words, if we run the same experiment multiple times (with the same unknown θ), we will get different observations and thus different random intervals.

Definition: Standard error Let $T = t(Y_1, \dots, Y_n)$ be an estimator of θ, let $\tau_n^2 = \text{Var}(T)$ be its variance, and let $V = v(Y_1, \dots, Y_n)$ be an estimator of τ_n^2. Then we call $\sqrt{V}$, or its realisation $\sqrt{v}$, a **standard error** for T.

Observation In most cases, we have:

$$U - L \propto \sqrt{V} \propto \frac{1}{\sqrt{n}}$$

In other words, multiplying the number of samples by 100 increases the precision by a factor 10.

Theorem: Empirical CLT Let T be an estimator of θ based on a sample of size n such that:

$$\frac{T-\theta}{\tau_n} \xrightarrow{\text{D}} Z \sim \mathcal{N}(0,1), \quad \frac{V}{\tau_n^2} \xrightarrow{\text{P}} 1$$

Then, we have:

$$\frac{T-\theta}{\sqrt{V}} \xrightarrow{\text{D}} Z$$

Intuition This means that, if we have an approximation of the variance V, we can use it for the central limit theorem.
This gives an empirical version of the CLT.

Proof This proof comes directly from the hypothesis:

$$\frac{T-\theta}{\sqrt{V}} = \frac{T-\theta}{\tau_n} \cdot \underbrace{\frac{\tau_n}{\sqrt{V}}}_{\to 1} \xrightarrow{\text{D}} Z$$

□

Example We throw a coin 200 times and observe 115 heads. We want to know if the coin is fair.
In other words, we have $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ and:

$$S = X_1 + \ldots + X_n \sim B(n, \theta)$$

Using the central limit theorem, we know that:

$$Q = \frac{S - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$

Using our regular method, we would need to solve the following inequation for θ:

$$-z_{1-\frac{\alpha}{2}} \leq \frac{S-n\theta}{\sqrt{n\theta(1-\theta)}} \leq z_{1-\frac{\alpha}{2}}$$

However, this is very hard to do. We thus use another method, our theorem. Using the law of large numbers, we get that:

$$\frac{S}{n} \xrightarrow{\text{P}} \theta$$

We moreover know that:

$$\tau_n^2 = n\theta(1-\theta)$$

Now, we don't want to get rid of the θ to simplify the inequality. Thus, we can make the following guess:

$$V = n\frac{S}{n}\left(1 - \frac{S}{n}\right)$$

This is indeed such that:

$$\frac{V}{\tau_n^2} = \frac{\frac{S}{n}\left(1-\frac{S}{n}\right)}{\theta(1-\theta)} \xrightarrow{\text{P}} 1$$

We can thus use the empirical version of the CLT, giving a new, easier to use, approximate pivot:

$$Q' = \frac{S-n\theta}{\sqrt{V}} = \frac{S-n\theta}{\sqrt{n\frac{S}{n}\left(1-\frac{S}{n}\right)}} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$

We can then use it to make our confidence integral, yielding:

$$-z_{1-\frac{\alpha}{2}} \leq \frac{S-n\theta}{\sqrt{S(1-\frac{S}{n})}} \leq z_{1-\frac{\alpha}{2}}$$
$$\iff \frac{S}{n} - \frac{z_{1-\frac{\alpha}{2}}}{n}\sqrt{S\left(1-\frac{S}{n}\right)} \leq \theta \leq \frac{S}{n} + \frac{z_{1-\frac{\alpha}{2}}}{n}\sqrt{S\left(1-\frac{S}{n}\right)}$$

This gives that:

$$\alpha = 0.05 \implies 0.506 \leq \theta \leq 0.644, \quad \alpha = 0.01 \implies 0.485 \leq \theta \leq 0.665$$

In the first case, we reject that the coin is fair, whereas in the second case we don't reject. This yields the following section.

8.4 Hypothesis tests

Observation

We notice that we can use confidence intervals to assess the plausibility of a value θ^0 of θ.
If θ^0 lies inside a $(1-\alpha)$ confidence interval, then we cannot reject the hypothesis that $\theta = \theta^0$ at significance level α.
However, if θ^0 lies outside as $(1-\alpha)$ confidence interval, then we reject the hypothesis that $\theta = \theta^0$ at significance level α.

Remark The greater the α that allows us to reject, the surer we are that the hypothesis should be rejected if it lies outside the confidence interval, but also the more values cannot be rejected.

Definition: Hypotheses

The **null hypothesis**, written H_0, is the theory we want to test. The **alternative hypothesis**, written H_1, is the opposite of H_0 (what is true when H_0 is false).

Definition: Types of errors

We call a **false positive** (or type 1 error) if we reject H_0 when it is in fact true.
We call a **false negative** (or type 2 error) if we reject H_1 when it is in fact true.

Remark We are negative when we accept the null hypothesis.

Definition: Simple and composite hypothesis

A **simple hypothesis** entirely fixes the distribution of the data Y. A **composite hypothesis** does not fix the distribution, it leaves some degree of freedom.

Example For instance, let θ_0 be fixed. The hypothesis "the coin is fair", H_0, is simple, but the hypothesis "the coin is not fair", H_1, is composite:

$$H_0 : R \sim B(n, \theta_0)$$
$$H_1 : R \sim B(n, \theta), \quad \theta \in]0,1[\setminus \theta_0$$

Definition: Size and Power

Let H_0, H_1 be simple hypotheses.
The false positive probability is named the **size** and is written α. The true positive probability is named the **power**, written β:

$$\alpha = \mathbb{P}_0(\text{reject } H_0), \quad \beta = \mathbb{P}_1(\text{reject } H_0)$$

Remark We need H_0 and H_1 to be simple hypotheses because we want to be able to speak of $\mathbb{P}_0$ and $\mathbb{P}_1$, which are dependent of a parameter if H_0 or H_1 is composite.

Observation They cannot be optimised both, we can minimise one at the expanse of the other. The trade-off depends on the context.

Example

Let T be a random variable. We consider the following simple null hypothesis and composite alternative hypothesis:

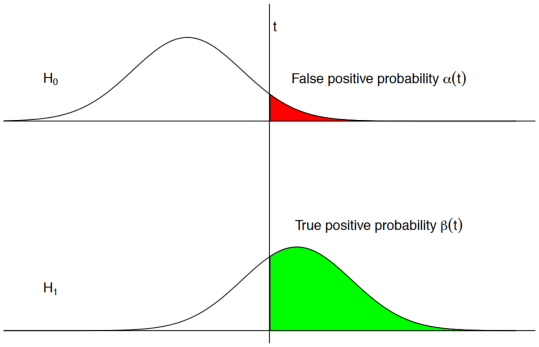
$$H_0 : T \sim \mathcal{N}(0,1), \quad H_1 : T \sim \mathcal{N}(\mu,1)$$

The idea is to reject H_0 if $T > t$, where t is some cut-off. To find this threshold, we compute the probability to have a false positive (reject H_0 incorrectly) and true positive (reject H_0 correctly):

$$\alpha(t) = \mathbb{P}_0(T > t) = 1 - \Phi(t) = \Phi(-t)$$

$$\beta(t) = \mathbb{P}_1(T > t) = \mathbb{P}(T - \mu > t - \mu) = 1 - \Phi(t - \mu) = \Phi(\mu - t)$$

Now, the amount of false positive and true positive (linked to the number of false negative) we accept to have depends on why we are doing this statistical test. For instance, we would not want to diagnose someone with a sickness they don't have. However, when we are looking for sicknesses in the population to know which were eradicated, we don't want to miss any.
We can visualise our choice between size (in red) and power (in green) using the following graphs:

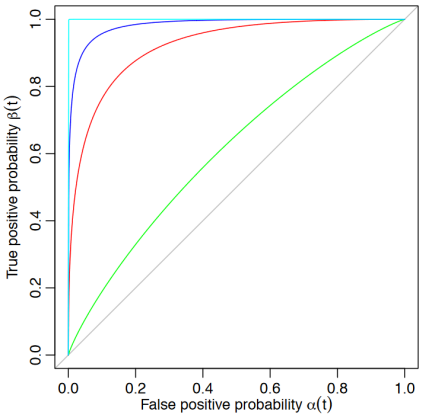


We indeed see that decreasing t increases both the size and the power, but increasing t makes them decrease.

Definition: ROC curve

The **receiver operating characteristic curve** (ROC curve) of a test is the plot of $\beta(t) = \mathbb{P}_1(T > t)$ against $\alpha(t) = \mathbb{P}_0(T > t)$ as the cut-off t varies.
This allows to know what true positive probability we have for any given false positive probability.

Example For instance, in our previous example, we can draw the 4 following ROC curves for $\mu \in \{0, 0.4, 3, 6\}$:



Definition: Pearson statistic Let $O_1, \dots, O_k$ be the number of observations of a random sample of size $n = n_1 + \dots + n_k$ falling into the categories $1, \dots, k$ with expected numbers $E_1, \dots, E_k$. The **Pearson statistic** (also named **chi-square statistic**) is:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Definition: Chi-square distribution Let W be a random variable and $\nu \in \mathbb{N}^*$. W follows the **chi-square distribution** with ν degrees of freedom, written $W \sim \chi^2_\nu$, if it has a density:

$$f_W(w) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} w^{\frac{\nu}{2}-1} e^{-\frac{w}{2}}, \quad w > 0$$

where $\Gamma(a)$ is the gamma function:

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du, \quad a > 0$$

Intuition The PDF of this distribution is not really important, the intuition of where it comes from is much more important.
Let $Z_1, \dots, Z_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then, the following random variable follows a chi-square distribution with ν degrees of freedom:

$$W = Z_1^2 + \dots + Z_\nu^2$$

Expectation We can for instance compute the expectation very easily:

$$\mathbb{E}(W) = \mathbb{E}\left(Z_1^2\right) + \dots + \mathbb{E}\left(Z_\nu^2\right) = \nu$$

Theorem Let $O_1, \dots, O_k$ be multinomial with denominator n and probabilities $(p_1, \dots, p_k)$, where:

$$p_1 = \frac{E_1}{n}, \quad \dots, \quad p_k = \frac{E_k}{n}$$

Then:

$$T \sim \chi^2_{k-1}$$

where T is the Pearson statistic.

Remark 1 The approximation is typically good if:

$$\frac{1}{k}\sum_{i=1}^{k} E_i \geq 5$$

Remark 2 This allows to verify if $O_1, \dots, O_k$ indeed follow a multinomial distribution.

Justification Under the multinomial distribution, we have:

$$\mathbb{E}(O_i) = np_i = E_i$$

$$\text{Var}(O_i) = np_i(1-p_i) = E_i\left(1 - \frac{E_i}{n}\right) \approx E_i$$

This means that, applying the central limit theorem, we (very approximately) have:

$$\frac{O_i - E_i}{\sqrt{E_i}} \approx Z_i \sim \mathcal{N}(0,1)$$

Thus, Pearson's statistics is given by:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \sum_{i=1}^{k} Z_i^2$$

However, we only have $k-1$ degrees of freedom, since we have the following constraint coming from the multinomial distribution:

$$O_1 + \ldots + O_k = n$$

This means that $O_1, \ldots, O_{k-1}$ are free, but O_k is completely determined by the others. Thus, we are summing $k-1$ squared independent standard Gaussians, giving us:

$$T \sim \chi^2_{k-1}$$

To make a formal proof, we would require the multidimensional central limit theorem.

Tuesday 30th May 2023 — **Lecture 25 : Will we manage to finish the course content?**

Definition: *P*-value

Let H_0 be a null hypothesis, and H_1 be its corresponding alternative hypothesis. Moreover, let T be some test statistic we apply on our data, and t_{obs} be its observed value.
The ***P*-value** is defined to be:

$$p_{obs} = \mathbb{P}_0(T \geq t_{obs})$$

Interpretation If p_{obs} is small, then either H_0 is true but we observed something unlikely, or H_0 is false.

Definition: Test significance

Let T be some statistical test, and p_{obs} be its observed P-value.
We say that the test is **significant at level α** if $p_{obs} < \alpha$.

Example

We consider again the case where we throw a coin 200 times, observe 115 heads and we want to know if the coin is fair. This times, instead of a confidence interval, we want to answer this question using the Pearson statistic and a P-value.
Let $S \sim \text{Binom}(200, \theta)$ be the number of heads, where $\theta = \frac{1}{2}$ under the null hypothesis, and $\theta \neq \frac{1}{2}$ otherwise. We see that we can also interpret our problem as a multinomial distribution:

$$(O_1, O_2) \sim \text{Multinomial}(200, (\theta, 1-\theta))$$

where $O_1 = S$ is the number of heads and $O_2 = 200 - S$ is the number of tails.
We can compute the Pearson statistic under the null hypothesis:

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} = 4.5$$

since, under the null hypothesis $E_1 = E_2 = np_i = 100$.
However, we know that $\mathbb{E}(T) = k - 1 = 1$. To know if this is a reasonable fluctuation of the Pearson statistic, we can compute its P-value using quantiles of the chi-square distribution (which can be computed numerically using a computer, or found in a table):

$$p_{obs} = \mathbb{P}_0(T > t_{obs}) = 0.034 \implies 1\,\% < p_{obs} < 5\,\%$$

This means that we can reject the null hypothesis that the coin is fair under significance $5\,\%$. However, under a $1\,\%$ significance, we must accept.

Measure of evidence

The P-value allows us reject the null hypothesis. However, it is often better to consider it as a measure of evidence against H_0: the smaller it is, the more evidence we have that H_0 should be rejected. In fact, knowing p_{obs} is often better than knowing whether H_0 was rejected.
As a rule of thumb, if $p_{obs} = 5\,\%$, we have weak evidence against H_0. If $p_{obs} = 1\,\%$ this is good evidence, and then, as this decreases, this gives more and more evidence.

Thursday 1st June 2023 — **Lecture 26 : The answer is yes**

8.5 Comparison of tests

Observation

We notice that we cannot optimise both the power and the size of a test. Now, if we freeze the size, we would like to have a way to maximise the power.

Definition: Rejection region

A test allows to split our data Y into two regions, the **rejection region**, written $\mathcal{Y}$, and its complement $\overline{\mathcal{Y}}$, such that:

$$Y \in \mathcal{Y} \iff \text{Reject } H_0$$

Equivalently:

$$Y \in \overline{\mathcal{Y}} \iff \text{Reject } H_1$$

Neyman-Pearson lemma

Let H_0 and H_1 be simple hypotheses, and let $f_0(y)$ and $f_1(y)$ be the densities of Y under those hypotheses. Finally, let $\alpha \in [0,1]$ be a size for our test.
We construct the following set:

$$\mathcal{Y}_\alpha = \left\{ y \in \Omega \mid \frac{f_1(y)}{f_0(y)} > t \right\}$$

where the $t \geq 0$ is chosen such that $\mathbb{P}_0(Y \in \mathcal{Y}_\alpha) = \alpha$.
If such a $\mathcal{Y}_\alpha$ exists, then it is the most powerful test of size α. In other words, it maximises $\mathbb{P}_1(Y \in \mathcal{Y}_\alpha)$ amongst all the $\mathcal{Y}'$ for which $\mathbb{P}_0(Y \in \mathcal{Y}') \leq \alpha$.

Proof

By hypothesis, $\mathcal{Y}_\alpha$ exists, meaning that there exists some $t_\alpha \geq 0$ for which $\mathbb{P}_0(Y \in \mathcal{Y}_\alpha) \leq \alpha$.
Let $\mathcal{Y}'$ be an arbitrary critical region of size α or less. Our goal is to show that:

$$\mathbb{P}_1(Y \in \mathcal{Y}_\alpha) \geq \mathbb{P}_1(Y \in \mathcal{Y}') \iff \int_{\mathcal{Y}_\alpha} f_1(y)dy \geq \int_{\mathcal{Y}'} f_1(y)dy$$

We take a step of abstraction, considering an arbitrary PDF $f_j \in \{f_0, f_1\}$ since we will need this later. We aim to show that the following is non-negative, for $j = 1$ (meaning $f_j = f_1$):

$$\begin{aligned} I_j &= \mathbb{P}_j(Y \in \mathcal{Y}_\alpha) - \mathbb{P}_j(Y \in \mathcal{Y}') \\ &= \int_{\mathcal{Y}_\alpha} f_j(y)dy - \int_{\mathcal{Y}'} f_j(y)dy \\ &= \int_{\mathcal{Y}_\alpha \cap \mathcal{Y}'} f_j(y)dy + \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}}'} f_j(y)dy \\ &\quad - \int_{\mathcal{Y}' \cap \mathcal{Y}_\alpha} f_j(y)dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}}_\alpha} f_j(y)dy \end{aligned}$$

where we partitioned $\mathcal{Y}_\alpha = (\mathcal{Y}_\alpha \cap \mathcal{Y}') \cup \left(\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}}'\right)$, and similarly for $\mathcal{Y}'$.
Since the intersection operator is commutative, our sum of integrals simplifies to:

$$I_j = \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}}'} f_j(y)dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}}_\alpha} f_j(y)dy$$

Now, by hypothesis, we see that:

$$\mathbb{P}_0(Y \in \mathcal{Y}') \leq \alpha = \mathbb{P}_0(Y \in \mathcal{Y}_\alpha)$$

However, this implies that:

$$I_0 = \mathbb{P}_0(Y \in \mathcal{Y}_\alpha) - \mathbb{P}_0(Y \in \mathcal{Y}') \geq 0$$

We thus want to make a link from I_1 to I_0. By construction of $\mathcal{Y}_\alpha$, we know that:

$$y \in \mathcal{Y}_\alpha \iff \frac{f_1(y)}{f_0(y)} > t_\alpha \iff f_1(y) > f_0(y)t_\alpha$$

In particular, this allows us to see that:

$$\begin{cases} f_1(y) > t_\alpha f_0(y), & y \in \mathcal{Y}_\alpha \\ f_1(y) \leq t_\alpha f_0(y), & y \in \overline{\mathcal{Y}}_\alpha \end{cases}$$

Since our sets are split according to whether $y \in \mathcal{Y}_\alpha$ or $y \in \overline{\mathcal{Y}}_\alpha$, this allows to simplify our inequality:

$$\begin{aligned} I_1 &= \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}}'} f_1(y)dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}}_\alpha} f_1(y)dy \\ &\geq \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}}'} t_\alpha f_0(y)dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}}_\alpha} t_\alpha f_0(y)dy \\ &= t_\alpha I_0 \\ &\geq 0 \end{aligned}$$

This indeed completes our proof.

□

Example

We throw a coin n times. We want to make an optimal test for the null hypothesis that the coin is fair, for a given size.
Each throw is a Bernoulli variable Y_i so, leaving $R = \sum_{i=1}^n Y_i$, we get:

$$g(r) = \frac{f_1(y)}{f_0(y)} = \frac{\theta^r(1-\theta)^{n-r}}{\left(\frac{1}{2}\right)^r\left(1-\frac{1}{2}\right)^{n-r}} = (2(1-\theta))^n\left(\frac{\theta}{1-\theta}\right)^r$$

We notice that if $\theta > \frac{1}{2}$, then $g(r)$ is increasing; and that it is decreasing when $\theta < \frac{1}{2}$. Let us suppose that $\theta > \frac{1}{2}$ for simplicity, the other case is very similar, except that all inequalities are reversed. In this case, we have:

$$g(r) = \frac{f_1(y)}{f_0(y)} > t \iff r \geq r_1$$

for some $r_1 \in \mathbb{R}$, because g increases.
We could try to isolate it, but this quickly becomes really unmanageable. Instead, since R is a sum of Bernoulli random variables, we can estimate r_1 using the central limit theorem:

$$\alpha = \mathbb{P}_0(Y \in \mathcal{Y}_1) = \mathbb{P}_0(R \geq r_1) = \mathbb{P}_0\left(\frac{R - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \geq \frac{r_1 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right) = 1 - \Phi\left(\frac{r_1 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right)$$

Now, α is given, so we want to invert this. We know that $\Phi(z_\beta) = \beta$, by definition of quantiles and since Φ is continuous, so our equation implies that:

$$\Phi\left(\frac{r_1 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right) = 1 - \alpha \implies z_{1-\alpha} = \frac{r_1 - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \iff r_1 = \frac{n + \sqrt{n}z_{1-\alpha}}{2}$$

Let's say that we made $n = 200$ throws and that we want a size $\alpha = 0.05$. This then yields:

$$r_1 \approx 111.6$$

In other words, if we have more than 111.6 heads, we can reject the null hypothesis that the coin is fair.

Optimal hypothesis test under average error probability

Let H_0 and H_1 be hypotheses with PMFs or PDFs f_0 and f_1, and let $H \in \{0,1\}$ be a random variable stating which hypothesis is correct (meaning that, if $H = 0$, then H_0 should be accepted). Moreover, let Y be some observations.
We want to make a test minimising the average probability of error (the bar over the $\mathbb{P}$ represents the fact that it is an average, not a complement):

$$\begin{aligned}\overline{\mathbb{P}}_e &= \mathbb{P}(\text{accept } H_1|H=0)\mathbb{P}(H=0) + \mathbb{P}(\text{accept } H_0|H=1)\mathbb{P}(H=1)\\ &= \mathbb{P}_0(\text{accept } H_1)\mathbb{P}(H=0) + \mathbb{P}_1(\text{accept } H_0)\mathbb{P}(H=1)\end{aligned}$$

We will moreover consider the prior that, if we don't make any observation, then $\mathbb{P}(\text{truth} = H_1) = \mathbb{P}(\text{truth} = H_0) = \frac{1}{2}$. In other words, we have maximal entropy since we have no information without any observation. This means that we are trying to minimise:

$$\overline{\mathbb{P}}_e = \frac{\mathbb{P}_0(\text{accept } H_1) + \mathbb{P}_1(\text{accept } H_0)}{2}$$

Estimator

Now, the idea of our estimator $\hat{H}$ of H, which purpose is decide which hypothesis is correct while minimising the average error, is to always pick the one which has highest probability to take place:

$$\begin{aligned}\hat{H}(y) &= \operatorname*{argmax}_{h\in\{0,1\}} \mathbb{P}(H=h|Y=y)\\ &= \operatorname*{argmax}_{h\in\{0,1\}} \frac{\mathbb{P}(Y=y|H=h)\mathbb{P}(H=h)}{\mathbb{P}(Y=y)}\\ &= \operatorname*{argmax}_{h\in\{0,1\}} \frac{1}{2\mathbb{P}(Y=y)}\mathbb{P}(Y=y|H=h)\end{aligned}$$

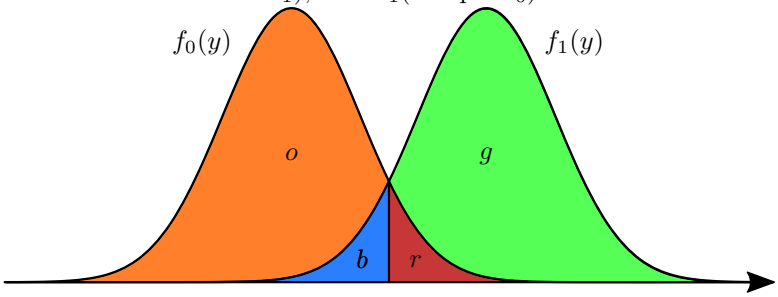
We notice that $\frac{1}{2\mathbb{P}(Y=y)}$ is just a constant with respect to h, so it has no impact on the maximisation of this function. This means that we can just remove it:

$$\hat{H}(y) = \operatorname*{argmax}_{h\in\{0,1\}} \mathbb{P}(Y=y|H=h) = \operatorname*{argmax}_{h\in\{0,1\}} f_h(y)$$

where we recognised the PMFs of our hypotheses. The case for continuous random variables is similar, and gives the PDFs.

Analysis of error

Let us consider the following graph to understand what happens. We notice that the probability of accepting H_1 when the truth is H_0, $\mathbb{P}_0(\text{accept } H_1)$, is the red area (places where the PMF of H_0 is lower than the one of H_1), and $\mathbb{P}_1(\text{accept } H_0)$ is the blue area.



By definition of $\overline{\mathbb{P}}_e$, we want to average those areas, meaning to compute half of their sum. However, we notice that their sum is given by the integral of the minimum between those two functions:

$$\overline{\mathbb{P}}_e = \frac{b+r}{2} = \frac{1}{2}\int_{-\infty}^{\infty} \min(f_0(y), f_1(y))dy$$

Now, to simplify this expression, we notice that the sum of the orange, blue and red regions gives 1, since we exactly have the PDF

of H_0:

$$o + b + r = 1$$

We can do the exact same reasoning to see that the sum of the blue, red and green regions give 1:

$$b + r + g = 1$$

Summing our two equations, we get that:

$$2 = (o + g) + 2(b + r) = \int_{-\infty}^{\infty} |f_0(y) - f_1(y)|dy + 4\overline{\mathbb{P}}_e$$

where we noticed that $o + g$ can be computed using the integral of the absolute difference of the two functions; since they represent the area bounded by the two PDFs. We moreover recognise that the integral is an L_1 norm for functions, telling us that:

$$\overline{\mathbb{P}}_e = \frac{1}{2} - \frac{1}{4}\|f_0 - f_1\|_1$$

As a general intuition, it makes sense that the more distant f_0 and f_1 are (here, according to the L_1 norm), the less we will make wrong predictions. However, let us consider two extreme cases to understand this better. If f_0 and f_1 are completely separated, meaning that $f_0(y) = 0$ when $f_1(y) \neq 0$ and inversely, then $\|f_0 - f_1\|_1 = 2$ and thus:

$$\overline{\mathbb{P}}_e = \frac{1}{2} - \frac{2}{4} = 0$$

This makes sense since, when the PDFs are completely separated, we should always be able to find the correct answer.
If however $f_0(y) = f_1(y)$, meaning that, for any observation, there is no hypothesis which has a higher probability, then $\|f_0 - f_1\|_1 = 0$ and thus:

$$\overline{\mathbb{P}}_e = \frac{1}{2} - \frac{0}{4} = \frac{1}{2}$$

This also makes sense since we can only make a uniformly random guess in that case.

Personal remark after a discussion with Marco Lourenço

In practice, this kind of ideas is not really used. We usually first choose our size (depending on the context), and then try to minimise the false negative probability. Minimising the average of the two is something which is very rarely used.
As an interesting side note, it seems like this is a special case of the maximum a posteriori probability estimate, where we only have two hypotheses, and do not give any prior advantage to any of them:

`https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation`

However, again, this over-specialisation of this test makes it not really used in practice.